

Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Human-Computer Interaction

LLMs as Narrative Facilitators: Successes and Limitations for Multi-user Prompt Fusion for Creating More-than-Human Narratives

Master's Thesis

Christoph Hein
Born Jun 14, 1996 in Kaiserslautern

Matriculation Number 125356

1. Referee: Jun.-Prof. Dr. Jan Ehlers
2. Referee: Dr. Teresa Paulino

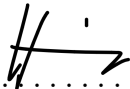
Advisor: Prof. Dr. Nuno Jardim Nunes

Submission date: May 27, 2025

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Kaiserslautern, May 27, 2025



.....

Christoph Hein

Abstract

This study investigates the integration of LLMs into Audience Participation Games to mediate collaborative storytelling on Twitch, addressing the lack of frameworks for equitable multi-user prompting. A mixed-methods approach is used to evaluate the LLM as a game agent and record the user experience. This approach combines qualitative interviews, game experience questionnaire surveys, and exact match analysis. This process is supposed to provide insights that contribute to a better understanding of the user experience when LLMs/GPTs are integrated into gaming contexts. The findings indicate that, while the used LLM (Qwq) effectively merged numerous inputs into a cohesive narrative, it exhibited an "early-input bias," preferring initial contributions and thereby compromising inclusivity in later turns. This resulted in a decrease of more than 50 percent in the EM score by the fifth turn. Participants reported moderate engagement and low perceived competence, suggesting deficiencies in user experience design concerning visibility and feedback. Positively there was next to no tension and frustrations recorded. Mutual influence while prompting the LLM to create the narrative received moderate ratings, indicating a moderate impact. Additionally, the LLM's adherence to more than human perspectives was noted as inconsistent with anthropocentric framing appearing in a few occasions. However, humor increased enjoyment and was partially successful combined with critiques of human ecological impact. Overall, the study is able to contribute to human-AI collaboration in multi-user contexts, but the LLM instructed as a game agent is not able to fulfill the task sufficiently all the time. Due to limitations such as small sample sizes and reliance on prompts to facilitate the agent, the thesis explores more the feasibility of creating narrations with users remotely providing multiple inputs in a prompt.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Motivation	2
1.3	Research Questions	3
1.4	Structure of the Thesis	3
2	Related Work	5
2.1	Large Language Models	5
2.1.1	LLM Overview	5
2.1.2	Background of Natural Language Processing	7
2.1.3	New Developments in the Field of LLMs	8
2.2	LLM Agents	10
2.2.1	LLM-based Agent: empowered & autonomous	10
2.2.2	LLM-Game Agents	14
2.3	To which extent LLMs/GPTs can profit from game environments	17
2.4	How can LLMs/GPTs contribute to Games	18
2.5	Human Computer Interaction studies on Twitch	19
2.5.1	Twitch Platform	20
2.5.2	Audience Participation Games	20
2.6	Non Human Voices with LLMs	25
2.6.1	Story writing with LLMs	26
2.6.2	More than Humans	28
2.7	Ethical Concerns	29
3	Focus Group Study – Evaluation and Findings	31
3.1	Introduction	31
3.2	Methodology	31
3.2.1	Study Design	31
3.2.2	Participant Selection	33
3.2.3	Procedure	33
3.3	Results and Evaluation	35

3.4	Summary and Implications	39
3.4.1	Key Takeaways	39
3.4.2	Next Steps	40
4	Interface for Remote Multiple User Prompts	41
4.1	Model Choice - Hein's Quick Model Choice Approach	41
4.1.1	Game Agent	47
4.2	Development of the LLM Game Agent and Interface	49
5	User Study on Twitch	56
5.1	Introduction	56
5.2	Methodology	57
5.2.1	LoGa - The Game	57
5.2.2	Study Design	58
5.2.3	Data Collection and Analysis	61
5.3	Procedure	63
5.4	Results	65
5.4.1	Qualitative Analysis	66
5.4.2	Quantitative Analysis	68
5.4.3	Questionnaire	74
6	Discussion	80
6.1	LLM adapting Audience Participation Game Requirements	80
6.2	Addressing QMC and similar Methods	81
6.3	Research Questions, Results & open Topics	84
6.3.1	Multi prompt: Prompt Aggregation - Overview	84
6.3.2	User Experience	86
6.3.3	Incorporation of Humor and More than Human Perspective	87
6.3.4	Integration of LLM in Gaming Contexts	88
6.4	Limitations	90
6.5	Conclusion & Future Directions	90
A	Appendix	92
A.1	Game Experience Questionnaire (GEQ)	92
A.2	Extra Questions	94
A.3	Interview Guide	94
A.4	Prompts	95
A.5	Agent Instructions	96
A.6	Tables	97
B	Tools	115

Bibliography

116

Acknowledgement

First, I would like to thank my advisor, Nuno, for opening the door to the Interactive Technologies Institute (ITI) and providing me with the opportunity to grow as an individual on such an interesting project and explore more than human contexts with the people behind LoGa. I would also like to thank greatly Teresa, my supervisor, for her excellent work throughout this process, as well as her willingness to listen and contribute to my progress. The completion of this thesis would not have been possible without the great contributions of the team behind LoGa. Finally, I would like to thank ITI for providing me with a workspace and a scholarship for the last three months of my dissertation in cooperation with Bauhaus University.

Acronyms

ABM - Agent Based Model
AGI - Artificial General Intelligence
AI - Artificial Intelligence
APG - Audience Participation Game
API - Application Programming Interface
CoT - Chain Of Thought
FGD - Focus Group Discussion
G4C - Games For Change
GEQ - Game Experience Questionnaire
GPT - Generative Pre-trained Transformer
HCI - Human Computer Interaction
IRC - Internet Relay Chat
LLM - Large Language Model
LLMGA - Large Language Model Game Agent
LoGa - Story Writing Game
LTSM - Long Short-Term Memory
ML - Machine Learning
MLLM - Multimodal Large Language Model
MtH - More Than Human
NLP - Natural Language Processing
NPC - Non Playable Character
RL - Reinforcement Learning
RNN - Recurrent Neural Network
TPP - Twitch Plays Pokémon
VSP - Video Streaming Platform

List of Figures

3.1	Focus Group during the Method Braindumping	34
3.2	Focus Group Results: LLM Agent Roles identified & Attributes, Perspectives	35
4.1	Quick Model Choice Approach Visualization	42
4.2	Box Plot for Average Rating given	46
4.3	QMC - simplified	48
4.4	Architecture of LLM Agent Interaction	49
4.5	Web based Interface for the LLM Agent	51
4.6	Flask Routes facilitating Web based Interface	52
4.7	Twitch Integration and Chat addition to the Prompt	53
4.8	Visual Character for narrating	55
5.1	The digital LoGa Board Game	57
5.2	Broadcastet View during the User Study	59
5.3	Twitch About Section with Information about the Research Target	60
5.4	Storyboard shows Involvement of the Agent, Streamer and Audience	61
5.5	Experience Level with Twitch/VSP	66
5.6	Word-cloud for top 30 used Words in the Output of the LLMGA	73
5.7	Game Experience as a Function of Twitch Experience	75
5.8	Stream Retaining Rate	77

List of Tables

4.1	Combined Scores for Language Models With and Without Re- searcher Input	45
4.2	Adjusted Scores for Language Models With and Without Re- searcher Input	47
4.3	Final Evaluation Results	47
5.1	Input Integration Analysis across Prompts	69
5.2	Top Ten most repeated Words from the LLMGA	73
5.3	In-game concise - iGEQ	74
5.4	Descriptive Statistics of GEQ social presence	76

Chapter 1

Introduction

This chapter provides a concise summary of the work. It begins by discussing the scope and motivations before outlining the research targets and contributions. The chapter concludes with a description of the work's structure.

1.1 Introduction

In just a few years, large language models (LLMs) have evolved from experimental technology to everyday productivity tools. Noy and Zhang published recently that ChatGPT reduces the time spent on work by 37% and improves the quality of results by an average of 18%. However, we're still only scratching the surface of a fundamental shift in how we work, hire, learn, and create value. It isn't human vs. machine — it's human + artificial intelligence (AI) teams solving problems on an unprecedented scale [Pawat, 2025]. Interacting with LLMs involves prompting. A prompt is an input to a generative AI model that guides its output [Hadi et al., 2023, White et al., 2023]. Prompts may consist of text, images, sounds, or other media. Some examples of prompts include: "Find me relevant research articles to the topic of multi-prompt" or providing an image with the instruction "describe the objects identified on this image", or a recording of a interview accompanied by the text "please summarize the following audio file".

But what if various inputs, opinions, or ideas are combined into one query to prompt a LLM? Examples of possible use cases are: Citizens submit thoughts on a policy, and the system fuses them into a coherent issue brief or recommendation. This could support models like digital town halls. Multiple experts or users brainstorm ideas, which the model then refines and clusters into concepts, pros and cons, or next steps, thus facilitating creative ideation processes. When the LLM is instructed and used not just as a solo tool but as a synthesis engine for group cognition, the model becomes a mediator or facil-

erator. Benefits lie in leveraging AI to augment brainstorming sessions, study groups, and informal debates among people and increasing the interactivity and enjoyment of various forms of digital media and communication, such as debates, role-playing, and AI-driven storytelling. In the context of this thesis, multi-prompt is understood as multi-user prompt fusion. Diverse inputs from multiple users, such as questions, opinions, goals, and constraints, are aggregated, possibly reframed, and contextualized under the usage of LLM agents into a single meta-prompt or structured input for a LLM or specialized agent. LLMs have demonstrated impressive capabilities in diverse domains, showcasing their proficiency in natural language comprehension, extensive knowledge of the world, and sophisticated reasoning abilities. These models have been examined in various contexts, with recent research showcasing their utility in a wide range of applications, including single-agent planners and assistants and particular interesting for this thesis purpose as game agents (GAs) [Hu et al., 2024a, Shao et al., 2023].

This thesis introduces an application space and interface that allows multiple users to simultaneously prompt a LLM. This environment is then used to play a round of LoGA, a participatory, posthumanism-inspired, contextualized storytelling game situated on Madeira Island, live-streamed on Twitch. Viewers are invited to participate in creating the narrative by sharing their creative input with the LLM empowered GA through the Twitch chat.

This study is additionally anchored in the Interactive Technologies Institute’s broader objective of committing to environmental preservation initiatives. These initiatives are based on the Institute’s core motivation: designing innovative interactive systems and services that harmonize the relationships between humans, non-humans, and digital technologies, as demonstrated by the game LoGa. These goals closely align with the United Nations’ Sustainable Development Goals and other initiatives (United Nations, 2023).

1.2 Motivation

Two driving forces motivated this research and study. The experiences and observations made while following a stream on Twitch with a LLM providing a source of engagement and novelty and seeing how the diverse crowd of people in the chat used this interaction to become creative in the framework provided by the streamer. Secondly, there is a need for more scientific work on facilitating multiple user inputs in a single prompt. This could improve interactions with AI for group chat functions and open new opportunities, especially for collaborative, participatory, or consensus-driven systems. By

dealing with multiple prompts and synthesizing diverse inputs, an additional step towards artificial general intelligence (AGI)-one of this century's goals could be made. To contribute in this gap is one of the core motivations of this work, examining prompt aggregation and the user experience with a state of the art LLM applied in the described context. Additional contributing factors are: Using Twitch as a testing environment to simulate user tests with LLMs, applying LLMs to audience participation games, and providing engaging experiences with educational backgrounds and storytelling mechanics. This thesis therefore evaluates such a scenario analyzing user experience and performance with multiple metrics trying to answer the following research questions.

1.3 Research Questions

This thesis aims to answer the next four questions and additionally contribute to the fields of research by offering a representation of non-human entities through LLMs in a variety of scenarios, employed in an unconventional test environment. This unconventional test environment is constituted by a video streaming platform (VSP) for user testing with an LLM.

1. Can a for this specific use case chosen and as agent instructed LLM process multi-prompts in the context of co-writing a story?
2. How are the interaction and user experiences perceived by participants in the context of a LLM facilitating the audience participation game?
3. How well can the LLM incorporate humor and more than human perspectives during narration?
4. How do the overall findings contribute to a better understanding of integrating LLMs/GPTs into gaming contexts and improving gameplay/user experience?

1.4 Structure of the Thesis

This thesis is structured to guide the reader from the foundational concepts of LLMs and LLM-based agents to the specific implementation and evaluation of a user study on Twitch in an interactive game environment.

Chapter 2 (Related Work) provides a thorough overview of LLMs, covering their development, capabilities, and applications in autonomous agents and games. The chapter also explores human-computer interaction (HCI) studies

on Twitch, audience participation games, and the role of LLMs in enabling non-human voices. It concludes with ethical considerations.

Chapter 3 (Focus Group Study: Evaluation and Findings) details a preliminary study that used a focus group to elaborate how a LLM is most effectively used to facilitate the gameplay of LoGa on Twitch. The chapter outlines the study’s methodology, participant selection, and key insights that informed the subsequent design of the interface utilized in the Twitch-based experiment.

Chapter 4 (Interface for Remote Multiple-User Prompts) details the technical development of the LLMGA system, including model selection with a introduced novel method and interface design tailored for real-time interaction in a streaming environment.

Chapter 5 (User Study on Twitch) describes the live experiment conducted on Twitch, including the game description, the study methodology, the data collection process, and the mixed-methods analysis of the qualitative and quantitative results.

The thesis concludes by synthesizing the findings and discussing their implications for HCI and LLM-driven interactions especially for in game contexts as well as the usage of multi-prompts. It also suggests future research directions and explains its limitations.

Chapter 2

Related Work

2.1 Large Language Models

This section provides an overview of LLMs, their capabilities and functionality and new developments in 2024/25 as well as some insight into natural language processing (NLP) and the steps taken to enable these models to solve complex tasks.

2.1.1 LLM Overview

LLMs, such as ChatGPT (OpenAI, 2022)¹, Gemini (DeepMind, 2023)², Claude (Anthropic, 2023)³ LLaMA [Touvron et al., 2023], Alpaca [Taori et al., 2023], and GLM [Zeng et al., 2023], are the latest but not newest paradigm of language models, which evolved from early statistical language models as described by Bellegarda [2004] to neural language models [Melis et al., 2017], then to pre-trained language models [Brown et al., 2020], and finally to LLMs [Zhao et al., 2024]. In essence, the role of a LLM is to respond to inquiries of the following nature. Given a sequence of tokens (i.e., words, parts of words, punctuation marks, emojis, program code etc.), what tokens are most likely to follow, assuming that the sequence is drawn from the same distribution as the vast corpus of public text, code and data on the internet [Shanahan et al., 2023]. In more detail assuming the type of model is a conditional probability distribution $P(w_{n+1} | w_1, \dots, w_n)$, where w_1, \dots, w_n is a sequence of tokens (the context) and w_{n+1} and is the predicted next token. In contemporary implementations, this distribution is realized in a neural network with a transformer architecture, pre-trained on a corpus of textual data to minimize prediction

¹<https://openai.com/index/chatgpt/>

²<https://blog.google/technology/ai/google-gemini-ai/>

³<https://www.anthropic.com/claude>

errors [Shanahan et al., 2023]. The transformers use a self-attention layer, as opposed to the traditional models that use recurrent or convolutional layers, along with an encoder and a decoder. Before transformers, recurrent neural networks (RNNs) with encoder-decoder architecture dominated machine translation and NLP. Self-attention compares different parts of the data. It calculates attention scores for all pairs of input elements, allowing the model to focus on relevant information as it learns [Cordonnier et al., 2019, Fang et al., 2023]. Recurrent layers process one step at a time keeping hidden states for information from previous steps causing complications for long range dependencies [Fang et al., 2023]. Convolutional layers use folding operation on local regions of the input data to indentify spatial data used for image processing [Cordonnier et al., 2019, Fang et al., 2023]. The key difference is self-attention excels at capturing relationships across input, RNNs focus on sequential data processing, and convolutional neural networks are optimized for spatial data analysis [Fang et al., 2023]. OpenAI’s GPT models use only the decoder part of this architecture. These models process natural language inputs (e.g., English or German) or programming code (e.g., Java or Python) through tokenization, as described. To discriminate the difference in parameter scale, the research community has coined the term LLMs for pre-trained LMs having a significant size difference to earlier LMs [Zhao et al., 2024]. The context length or window mentioned in these models’ specifications are the maximum inputs and outputs tokens that the models process at once. For the ChatGPT models, versions are not specified, but the API versions are 16k and 128k tokens for GPT-3.5-turbo and GPT-4-turbo, respectively. The range of tasks which can be solved from a model following this simple tactic is extraordinary. LLMs exhibit strengths in generating discourse and reasoning but might produce factually incorrect outputs, termed hallucinations [Alvarado, 2024]. Text generation, summarization, translation etc., are an easy feed, but also in complex reasoning and planning tasks, such as solving mathematical problems LLMs can record success [Arora et al., 2023]. They can effectively evaluate the validity of their claims and predict their knowledge of answers, showing promising calibration and performance in self-evaluation tasks, particularly when considering multiple samples before making predictions [Kadavath et al., 2022]. Additionally larger models which are scaled up exponentially show emergent abilities. These abilities vary across language models and tasks and are observed at specific computational scales [Wei et al., 2022a]. These abilities enabled tasks such as complex reasoning, nuanced understanding of language, and even deception strategies which have been observed in advanced LLMs like GPT-4 [Hagendorff, 2023]. The technological advancements that have been achieved in these domains are founded on NLP which includes understanding and generation, both in the past and in recent times. In order to understand the utilization of LLMs and

the possibilities that they present, it is beneficial to examine these steps in the next section.

2.1.2 Background of Natural Language Processing

Gururaja et al. determined the most important paradigm shifts in NLP by analyzing published literature in the mid nineties until the early twenties. After early machine translation models from IBM and statistical NLP, one of the early works on neural NLP comes from Collobert and Weston who published a paper in 2008 on single convolutional neural network architecture that, given a sentence, outputs a host of language processing predictions [Collobert and Weston, 2008]. This paper marked a pivotal shift in the community’s perspective on neural models, moving from skepticism to a motivated effort on early neural NLP ultimately leading to a widespread adoption [Gururaja et al., 2023]. Word2Vec, Seq2Seq and Tensorflow were released in 2013, 2014, and 2015 respectively, facilitating a “neural revolution in NLP” [Gururaja et al., 2023]. Mikolov et al. present two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task and the results are compared to the previously best performing techniques based on different types of neural networks. The results show a significant enhancement in accuracy [Mikolov et al., 2013]. Seq2Seq is a general approach to sequence learning that makes minimal assumptions about the sequence structure. The researchers’ described method utilized a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector [Sutskever et al., 2014]. TensorFlow is a machine learning system that works with large amounts of data and in different environments. TensorFlow uses dataflow graphs to represent computation, shared state, and the operations that change that state. It organizes the data flow in a graph, mapping it across multiple machines in a cluster or on a single machine across different parts [Abadi et al., 2016].

The first LLMs (ELMo, GPT and BERT) followed in 2018 [Gururaja et al., 2023]. Generative pre-trained transformer are supposed to perform on many tasks eventually without the need to manually create and label a training dataset for each one [Radford et al., 2019]. GPT-3, which was released in 2020, is an autoregressive language model with 175 billion parameters, which is 10 times more than any previous non-sparse language model. The performance of GPT-3 is tested in the few-shot setting by Brown et al.. For all tasks, GPT-3 is applied without any gradient updates or fine-tuning, with tasks and few-shot demonstrations specified purely via text interaction with the model. GPT-

3 achieves strong performance on many NLP datasets, including translation, question-answering, and close-ended tasks [Brown et al., 2020].

2.1.3 New Developments in the Field of LLMs

⁴ Concurrently, the LLMs of 2024/25 employ sophisticated techniques, such as chain of thought reasoning, to facilitate the breakdown of complex tasks and thereby enhance their problem-solving capabilities. By definition, a chain of thought (CoT) is the decomposition of a problem into a series of intermediate mental steps in natural language that ultimately lead to a rational answer [Wei et al., 2022b]. Leading up to the automatic CoT researches already explored the possibility of CoT-prompting. Wei et al. remarks using a few chain of thought demonstrations provided as exemplars in prompting, leads to improved performances on a range of arithmetic, commonsense, and symbolic reasoning tasks. For instance, prompting a 540B-parameter language model with just eight chain of thought exemplars achieves state of the art accuracy on the GSM8K ⁵ benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier [Wei et al., 2022b]. Open-AI introduced in September the model o1 which makes use of this special feature. Before the language model generates an output, it simulates an internal monologue. It creates an internal CoT corresponding to the complexity of the input. The model works through this CoT step by step before generating a final answer [Zentner, 2025]. Additionally to the increased value, CoT in the responses makes the LLM’s inner monologue visible, the generated responses can be monitored and explained much more easily than before unveiling the so termed black box to a certain degree improving the process of trouble shooting [Zentner, 2025]. OpenAI has already released three different versions of its thinking o1 models, and a further development called o3 is due to be released shortly. Google has also already released an experimental, thinking Gemini Flash version and recently a Chinese startup made headlines with their model Deepseek R1 [Stoppel, 2025]. DeepSeek R1-Zero demonstrates capabilities such as self-verification, reflection, and generating long CoTs, marking a significant milestone for the research community. Furthermore, the researchers showed that distilling more powerful models into smaller ones yields excellent results in reasoning abilities marking an resource conserving approach for training models [DeepSeek-AI et al., 2025]. Already in the end of last year the latest experimental model, o3, achieved a high score on a test that marks progress towards AGI sparking excitement. OpenAI’s o3 scored 87.5% on a relevant benchmark, trouncing the

⁴This Section takes in account articles released until January 2025 to provide an overview of recent events

⁵<https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k>

previous best score for an AI system of 55.5% [Jones, 2025]. Some specialists think that o3 might be producing a series of different chains of thought to help whittle down the best answer from a range of options compared to model o1. Spending more time refining an answer at test time makes a huge difference to the results but also comes at a massive expense. To tackle each task in the ARC-AGI test ⁶, its high-scoring mode took an average of 14 minutes and probably cost thousands of dollars [Jones, 2025]. This sparks new debates about the large energy consumptions and sustainability of LLMs which will be thematized in this work in a later stage.

With these advances, the step to AGI seems to be close, but since there is no technical definition, opinions may diverge. If intelligence is defined as the ability to answer complex questions on a similar competence humans, advanced models such as o1/o3 can already be classified as AGI today, taking into account the test results mentioned above. However, there are many tests developed to measure and quantify possible progress like for example Rein et al. 2023 Google-Proof Q&A, which intends to assess an AI system's performance on PhD-level science problems. Containing a challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry the researchers ensured that the questions are high-quality and extremely difficult. Experts who have or are pursuing PhDs in the corresponding domains reach 65% accuracy (74% when discounting clear mistakes the experts identified in retrospect), while highly skilled non-expert validators only reach 34% accuracy, despite spending on average over 30 minutes with unrestricted access to the web (i.e., the questions are "Google-proof") [Rein et al., 2023]. The questions are also difficult for state-of-the-art AI systems, with the strongest GPT-4 based baseline achieving 39% accuracy [Rein et al., 2023]. As an other example, OpenAI's 2024 MLE-bench pits an AI system against 75 ML engineering-related challenges hosted on Kaggle, an online data-science competition platform. The researchers aim to test real-world ML engineering skills such as training models, preparing datasets, and running experiments. They establish human baselines for each competition using Kaggle's publicly available leaderboard [Chan et al., 2024]. With constantly measuring and improving benchmarks the researches aim to keep up with the rapidly evolving and improving field of LLM and make advancements quantifiable on the way to creating AGI.

The CoT discussed in this section is crucial for evaluating prompt aggregation and understanding user input handling in this thesis. These recently discovered, greatly improved skills enable also a broad spectrum of applications for state-of-the-art LLMs. These include usage within agent-based modeling

⁶<https://github.com/fchollet/ARC-AGI>

(ABM) and acting as an autonomous agent in various fields. These topics will be addressed in the subsequent section.

2.2 LLM Agents

The purpose of this section is to furnish an overview of the domain of LLM-based agents and the integration process of LLMs. The integration process of LLMs encompasses their application areas, such as ABM, which will be subsequently introduced. Furthermore, the section will explore game agents (GAs), a notable application of LLM-based agents, their application, and examples analogous to the initiative of this work, providing a comprehensive overview of the subject. In the field of AI, the term "agent" is frequently used to describe software that processes observations from an external environment and responds to them in a closed loop. LLMs can be used in this instance to simulate human behaviors, given the powerful ability to understand human instructions and provide high-quality generated texts [Shao et al., 2023].

2.2.1 LLM-based Agent: empowered & autonomous

AI agents are artificial entities that sense their environment, make decisions, and take actions. Due to the versatile capabilities LLMs demonstrate, they are regarded as potential sparks for AGI, offering hope for building also general autonomous AI agents. Many researchers have leveraged LLMs as the foundation to build AI agents and have achieved significant progress [Gao et al., 2023a, Xi et al., 2023]. A comparison of LLMs and reinforcement learning models reveals that the former possess a more comprehensive internal world knowledge, thereby facilitating more informed agent actions even in the absence of training on specific domain data. Moreover, LLMs are capable of providing natural language interfaces for interaction with humans, a feature that is both flexible and explainable [Wang et al., 2024]. Thus making it a promising venture to construct LLM-based agents due to their human like intelligence in perceiving the environment and making decisions, which became a reinforced strength due to CoT. A necessary distinction must be made between two categories of agents: empowered and autonomous. This is a topic that is sometimes overlooked in the field, and as a result, attributes are misaligned with the assigned LLM-based agent. This might complicate and provoke misunderstandings when reading and considering research papers for incorporation into one's own research. Empowered agents enhance specific tasks using LLMs, while autonomous agents aim to operate independently across broader objectives Händler [2023], Li et al. [2023a]. The distinction between empowered and

autonomous agents becomes clearer when considering ABM and role-playing contexts.

Agent-Based Models specifically for Social-Ecological Systems

ABMs consist of autonomous, interacting computational objects, called agents, often located in space and time [Macal and North, 2005]. These agents may be identical or each may have unique attributes. Agents can be as few as one or as many as millions. Because these models rely on computation, the behavior of the agents is rule-based. The rules can be simple (e.g., backward induction on an extensive game form) or more complex (e.g., based on heuristics derived from cognitive psychology or neuroscience) [De Marchi and Page, 2014, Macal and North, 2005]. In an ABM, agents receive inputs from their environment and take actions in response to those inputs; as such, many ABMs implement social networks or spatial relationships that play a role in decision making [De Marchi and Page, 2014]. Prediction is the most fundamental ability of agent-based simulation [Gao et al., 2023b]. The environment, whether static or evolving, introduces conditions, instigates competition, defines boundaries, and occasionally supplies resources influencing agent behaviors [Cipi and Cico, 2011]. Outcomes are not directly determined by assumptions but instead emerge from the interactions with both the environment and other agents in the model. The aggregation of these individual behaviors, instantiated in software, produces system-level outcomes such as cooperation and price fluctuations [De Marchi and Page, 2014]. The goal is to mirror the behaviors in reality based on predefined or adaptive rules [Macal and North, 2005]. Examples for a possible application are in economics, in which language models can help forecast market trends, predict consumer behavior, or estimate the impact of policy changes. In sociology, these models can aid in predicting social movements, public opinion shifts, or the adoption of new cultural practices [Gao et al., 2023b].

Since the motivation of this work is to mediate sustainable and ecological approaches using a LLM agent, it is coherent to take a closer look at ABM in an ecological context, the so-called social-ecological systems modeling. SES involves the development and/or application of models to investigate complex problems arising from interactions between humans and natural systems. The field emphasizes the interdependence and interconnectedness of social and ecological components, recognizing that human actions and decisions can have both positive and negative impacts on the environment, and that these impacts can affect human well-being [Reyers et al., 2018]. Nugroho and Uehara list three primary purposes for employing SES: Study how social and ecological outcomes emerge and how the system changes over time; evaluate how new

policies or disturbances affect a complex adaptive SES, including possible unintended consequences; and promote processes that involve participation, which can help people better understand issues and work together to solve problems. It is noteworthy that ABMs can effectively represent human actions in the environmental dimension. This is due to the inherent focus of ABMs on emergent behaviors arising from agent interactions. This capacity aligns seamlessly with the modeling of human efforts to shape their surroundings. Examples of such efforts include land-use alterations, conservation initiatives, and restoration programs [Nugroho and Uehara, 2023].

After this brief excursion into ABM, the role of the LLM-based agent will be now examined in a comparative manner. Overall, empowered agents are characterized by their ability to enrich tasks by supporting specific roles or behaviors. Autonomous agents, on the other hand, are capable of driving complexity and decision-making processes independently. Empowered agents employ LLMs as a central component of their decision-making or interaction capabilities. However, they may necessitate human oversight or external logic systems to function optimally [Xi et al., 2023]. These agents serve as tools for specific tasks where the LLM enhances certain possibilities such as text generation while depending on predefined workflows or humans for critical decisions [Xi et al., 2023]. Autonomous agents are systems built around LLMs that are designed to perform tasks autonomously, without needing continuous human guidance or oversight. These systems integrate advanced reasoning, task planning, and action execution capabilities, in combination to broader abilities such as reinforcement learning, external APIs (Application Programming Interfaces), and self-improvement techniques [Shinn et al., 2023]. They are capable of making independent decisions across multiple tasks, generating, evaluating, and dynamically prioritizing goals [Nakajima, 2023, Wang et al., 2024]. An example for such an autonomous agent is "BabyAGI" an AI-powered task management system with multiple LLM-based agents. One agent creates new tasks based on the previous task's objective and result, another agent prioritizes the task list, and another agent completes tasks. [Nakajima, 2023]. In the context of ABM, a simulation technique where agents interact in an environment according to predefined rules. The emphasis is on the comprehension of emergent behaviors, system dynamics, and the manner in which individual interactions influence outcomes in complex systems [Savin et al., 2023]. Empowered agents, in turn, yield authentic behavioral patterns. For instance, these agents can be programmed to make decisions that mirror those observed in economic or social systems, thereby simulating human-like behavior, enhancing the realism and effectiveness of macroeconomic simulations as well as generate nuanced dialogue and decision reasoning utilizing the LLM

[Li et al., 2023a]. For autonomous agents the operation is characterized by full independence within the context of simulations. This enables the systems to learn from the environment and adapt their behaviors dynamically. LLMs have the potential to function as the central intelligence that guides agents' decisions. This intelligence can incorporate reinforcement learning or evolutionary strategies to facilitate the emergence of behaviors [Wang et al., 2024]. To sum it up empowered agents in ABM are tools to enhance pre-scripted scenarios, while autonomous agents evolve or adapt, contributing to the emergent complexity of the model. Role play with LLMs involves the assumption of specific roles in interactive scenarios, such as a teacher, negotiator, or game character. Role-playing exploits the LLM's capacity to generate contextually appropriate responses and adapt based on user input [Shanahan et al., 2023]. Examining empowered and autonomous agents in this framework discloses that empowered agents support or augment human role-playing, while autonomous agents actively shape and control the narrative or interaction independently. Role-playing empowered agents act as surrogates for human users, providing enriched responses within constrained scenarios. They typically simulate roles under predefined rules and settings, requiring external systems or human intervention to manage the overall interaction flow [Zhang et al., 2024b]. For example, in tabletop role-playing games, a dungeon master's assistant uses a LLM to make suggestions for events without exercising autonomous control over the entire game [Callison-Burch et al., 2022]. Autonomous agents are defined as fully self-directed entities capable of assuming dynamic roles without the need for continuous human guidance. These agents are capable of driving the narrative forward, establishing objectives, and adapting to user or agent interaction, often resolving conflicts autonomously [Xi et al., 2023]. A prominent illustration of such an agent is "Voyager", the first LLM-powered embodied lifelong learning agent in "Minecraft" that continuously explores the world, acquires diverse skills, and makes novel discoveries without human intervention [Wang et al., 2023].

Hauser et al. declares that language is of the most importance in interactions with the environment and provides the basis for high-level abilities. Therefore, creating a language-based interaction with the agent can be seen as a first step toward higher-level task-solving possibilities. Two straightforward steps are all it takes to provide an effective dialogue agent. First, the LLM is embedded in a turn-taking system that interleaves model-generated text with user-provided text/prompts. Second, a dialogue prompt is supplied to the model to initiate a conversation with the user. The dialog prompt typically consists of a preamble that sets the scene for a script- or play-style dialog, followed by a sample dialog between the user and the agent [Shanahan et al., 2023]. A prompt is a

set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities. These kind of prompts by the definition of White et al. are instructions which are given to an LLM to enforce rules, automate processes, and ensure specific qualities (and quantities) of generated output. These prompts are also a form of programming that can customize the outputs and interactions with an LLM. A type of method used in several instances to define and provide a agent with a specific persona, rules, or environment in which the agent is placed. Especially when situated in the setting of a game, this opens up a lot of opportunities which are further introduced in the subsequent section.

2.2.2 LLM-Game Agents

LLM-based game agents (LLMGAs) capable of employing cognitive abilities to gain fundamental insights into gameplay are another subfield of the LLM-based agents explored previously.

The development of GAs could play a pivotal role in the advancement of AGI. The progress of LLMs and their multimodal counterparts (MLLMs) presents an unparalleled opportunity to evolve and empower GAs with human-like decision-making capabilities in complex computer game environments [Hu et al., 2024a]. Digital games are regarded as optimal platforms for the development of AI agents due to their inherent characteristics, which include complexity, diversity, controllability, safety, and reproducibility [Hu et al., 2024a]. Several examples ranging from Chess and Poker to modern video games like Starcraft and Minecraft have been instrumental in AI research. By comparing AI systems with human players in games, researchers are able to evaluate the capabilities of these systems. This was the case in early competitions between computers and a chess master, as well as years later in the game GO and Poker Texas Hold'em through to the very complex real time strategy games Dota 2 or Starcraft 2. In each of these cases, an advanced model eventually defeated the reigning champion [Tsai et al., 2023]. In recent years, LLMs have demonstrated impressive abilities to understand and respond to complex human language queries, so it's not surprising that researchers have begun experimenting with text-based adventures where the game mechanic consists solely of understanding and producing text. A text adventure game provides a text-based environment in which players use text commands to interact with the world, exploring and completing quests. Tsai et al. test in their case study, which is claimed to be one of the first, the ability of GPT-3.5 playing Zork with a human acting as communicating channel between the game state and the LLM. In Zork 1, the character follows natural language commands to search for treasure in the Great Underground Empire: he moves between locations and interacts

with objects; the game program acts as a narrator, describing the player's location and the consequences of his actions. The pipeline used sees, at each step, the human player passing to ChatGPT the current game state described in the natural language of the game world, asking ChatGPT to take a legal action, and then feeding the action taken back to the game, completing a functional cycle that allows the model to experience the game [Tsai et al., 2023]. The overall performance is described by the authors as promising, with some weaknesses such as repetition of previous responses when asked for next actions. Their experiment also showed much higher success when previous actions were fed into following prompts. Overall the game-playing ability of GPT-3.5 is classified on par with state-of-the-art reinforcement learning (RL) approaches [Tsai et al., 2023]. However, most visual novels and text-based games rely on multiple option answers limiting the freedom of choice of the players. Some prompt-based games have recently appeared on Reddit's r/ChatGPTGaming and the prompt marketplace PromptBase, which embraces the idea that simple games can be developed using only natural language prompts [Isaza-Giraldo et al., 2024]. Further developments often involve improved and more complex frameworks, use newer and improved LLMs, or use fine-tuning to achieve higher performance levels. LLMGAs with such extensive work are consequently placed in the more challenging environment of communication games among others. Communication (or conversational) games revolve around turns of communication, negotiation, deduction, and deception among multiple players. The challenge of communication games is to guess the intentions of others behind ambiguous or misleading language, and hiding one's own intentions when necessary [Hu et al., 2024a]. One particular example for a strictly text based use case is the development of an agent with RL for strategic play in Werewolf [Xu et al., 2024]. The researchers goal to develop an agent which generates flexible language actions and possesses strong decision-making abilities is tested in multiple rounds of Werewolf. The game pits two groups against each other, i.e., werewolves and non-werewolves (villagers, other roles with unique abilities: witch, guard and seer), and alternates between night phases, where werewolves secretly attack, and day phases, where survivors discuss and vote to eliminate suspects. The agent performs against other models and even plays with human competitors achieving an human average win rate [Xu et al., 2024]. Xu et al. employ a RL policy to select the optimal action from among the diverse actions generated by the LLM aiming to overcome prior preferences for specific actions resulting in a worse game performance. Furthermore, there is the domain of communication games that overlap with agent simulations, in which projects such as "WarAgent" and "CosmoAgent" are situated. Both of these utilize a multi-agent structure to simulate either diverse ethical paradigms while simulating interactions between entities with

distinct moral principles in the context of extraterrestrial civilizations [Jin et al., 2024] or to simulate multiple war parties, their decisions, and the consequences in historical international conflicts, including the World War I (WWI), the World War II (WWII), and the Warring States Period (WSP) in Ancient China [Hua et al., 2023]. The output is text-based and displays communication protocols and decisions made on the basis of the current situation or predefined mathematical/conceptual models.

It is evident that there exists a multitude of genres of games and advanced LLMGA architectures that are not reliant upon textual inputs and outputs. These agents possess augmented capabilities for both control and perception, in addition to memory storage, which facilitates enhanced cognitive abilities and learning from particular contexts as described by Hu et al.. This development opens up opportunities in the areas of competitive, crafting and exploration games, where a high level of competence of the LLM is required to complete relevant tasks or progress in the game. In order to keep the focus on the project presented in this paper, these additional use cases and rich architectures of LLMGAs will not be explored further, as little of it is relevant to the interface and application explored in this work. Instead, the focus will return to projects with a similar streamlining to this project, focusing on ecological and environmental challenges with the use of LLMGAs.

Games provide a quick and intuitive way for people to immediately see the consequences of their actions, and are a method of raising awareness by artificially incentivizing positive factors and discouraging negative ones [Zhang et al., 2024a].

The "Meaningful Energy Game" by scholars from the Interactive Technologies Institute (ITI) investigates the possibility of implementing LLMs as evaluating agents of open-ended challenges in serious games and its potential to facilitate a meaningful experience for the player. Each level represents a different energy community challenge for the player, and the narrative incrementally increases in scope. This narrative aims to provide perspective on the challenges that energy communities face on different scales and forces the player to pose different types of environmental solutions. The LLM is in charge of mediating the game and evaluating whether or not the player's answers are satisfactory to overcome the proposed challenges [Isaza-Giraldo et al., 2024]. For an other example, EcoEcho immerses players in the role of a conflicted scientist whose misinterpretation of his father's dying wish leads him to impede the advancement of clean energy. Throughout the plot, the scholars leverage GenAI's ability to elicit natural language responses from players and seamlessly translate these verbal inputs into corresponding in-game actions. Through dynamic interactions with various stakeholders, such as journalists, union leaders, and energy ministers, players engage in unsustainable behaviors within the game

world, turning conversational choices into tangible consequences [Zhang et al., 2024a].

Serious Games & Games for Change

The examples showcased draw on the theme of serious games, a specific subset of video games that aim to increase knowledge and awareness while maintaining entertainment value to improve information retention and learning [Gursesli et al., 2023]. As a subset of serious games, Games for Change (G4C) focus on influencing attitudes and promoting positive social and behavioral change, and are attracting interest from developers, researchers, and nonprofit organizations for their potential in climate change education. Despite many promising applications, challenges remain in balancing engagement with educational value. In particular, maintaining the educational purpose without compromising the fun and engaging aspects of the game is challenging. To address this challenge for the targeted use case, the additional step of a focus group is conducted to define the scope and spaces in which novel interaction or gamification elements could be applied.

2.3 To which extent LLMs/GPTs can profit from game environments

After looking extensively on the use cases of LLM-based GA and touching on the genre of serious games, this section focuses on the benefits of employing LLM in a testbed, like games.

Games represent an optimal subject for research involving LLMs/GPTs. Games yield rich multimodal data, which is well-suited for LLM research, and they also provide extensive corpora of text and multimodal data produced by players, viewers, fans, and so forth. Game text data, including transcripts, have already been utilized to train LLM players in multiple cases. The Meta Fundamental AI Research Diplomacy Team (FAIR) developed an agent based on that data corpus that is able to play the full natural language form of the game Diplomacy and demonstrates performance well above the human average in an online league [FAIR]. On the other hand, LLMs inherently grapple with spatial reasoning and planning, two facets that are integral to many games. These games cover a wide spectrum, from strategy board games and digital games where long-term planning is paramount, to first-person shooters that require precise spatial reasoning, to real-time strategy where formulating a reactive plan to achieve the goal of defeating the enemy and reaching the enemy's base is essential [Vinyals et al., 2019]. Consequently, these games can

continue to serve as cutting-edge testbeds for AI playing games and are likely to remain challenging domains for LLM research. Games are also contingent on long-term interactions, particularly in the case of LLM-based game masters. Consequently, games can function as testbeds or benchmarks to explore the limits of recollection under varying context lengths [Gallotta et al., 2024].

2.4 How can LLMs/GPTs contribute to Games

The relationship between games and LLM goes both ways. There are many aspects that can be positively influenced by the use of a LLM. It can even provide the game mechanics, as seen in the example of Infinite Craft ⁷. In this game, developed by Argarwal N., the player starts with a set of core elements (water, fire, wind, and earth). Infinite Craft then asks Llama 2 to imagine the product of combining these elements, resulting in a game with infinite combination possibilities [Gallotta et al., 2024]. Other benefits discussed by de Lima et al. include the integration of GPT into game-based learning environments to provide personalized, subject-specific feedback on players' in-game actions, answer their questions during gameplay, and provide relevant explanations or demonstrations based on the game's tasks and learning content. The scholars discover in a study, that players in GPT-assisted environments exhibit stronger intrinsic motivation and cognitive engagement compared to those in purely game-based settings [de Lima et al., 2023]. In addition, having an in-game character provide feedback is more immersive than relying on an external system. Within games, GPT's interactive narratives and real-time feedback could enable fun, dynamic, and personalized experiences that adapt to individual players' needs, allowing a broader range of users to engage more deeply with the subject [Yang et al., 2024b]. LLMs have been shown to enable more natural and contextually relevant dialogue with non-player characters (NPCs). A recent study of GPT-4-driven NPCs in Minecraft demonstrated the potential for LLMs to enhance NPC interactions and player engagement, as evidenced by collaborative quest completion in a mini-game [Rao et al., 2024]. This early work shows the potential for future games having highly cooperative and adaptive NPCs due to LLM integration. Another aspect could be analyzing player behavior to provide adaptive hints and adjust difficulty levels in real time to improve the user experience. Yang et al. provide insight on lacking research on user experience regarding the integration of LLMs/GPTs into gaming contexts as the over reliance on quantitative metrics can be criticized for obscuring genuine emotional responses and interactive experiences. To address this widely open field, this work aims for evaluating the LLMGA

⁷<https://neal.fun/infinite-craft/> Accessed: 27.03.2025

utilized additionally based on qualitative metrics like follow-up interviews to complete after participating in the game. Additionally, a voluntary questionnaire is intended to bridge the gap between metrics and subjective ratings. In the next sections the proposed test environment is introduced and its relevance for HCI is demonstrated with examples showcasing the potential for user tests.

2.5 Human Computer Interaction studies on Twitch

Where to facilitate and test AI features or other concepts aimed at user participation, or to present some capabilities in video form? In this breach, a video streaming platform (VSP), specifically Twitch, offers some interesting possibilities. Live streaming the performance of a LLM playing a rather complex game like Pokémon Red, or streamers using creative models to play their own story-based version of conquering the world, having multiple LLM characters have a run at a Dungeons and Dragons game, are just a few creative approaches that have happened recently on Twitch. VSPs offer rich audiovisual information and can potentially reach a large community with many users and collect qualitative and quantitative data, making it a testing environment with a lot of potential [Bartolome and Niu, 2023].

HCI research on Twitch has explored various aspects of live streaming, community interactions, and user engagement. Faas et al. studied game development streams using a participant-observer approach, finding that live streams foster learning-centered communities where both streamers and viewers mentor each other. Pellicone and Ahn conducted a qualitative study of an online forum for streamers, revealing how streamers integrate technical, social, and gameplay skills to create engaging performances and highlighting the role of platform design in supporting equitable participation. Seering et al. analyzed 183 million Twitch chat messages to identify factors that influence first-time participation, showing that diverse user presence increases engagement, while information-seeking behavior correlates with lower retention. Sheng and Kairam explored relationship formation on Twitch through interviews and surveys, demonstrating that text-based interactions can foster close connections, though larger communities can hinder relational engagement. These studies illustrate the unique social interactions and affordances of VSPs and provide valuable insights for social media and social computing research [Bartolome and Niu, 2023]. The area of evaluating a LLM and user experience on a VSP is rather unexplored, making it an important target for this thesis.

2.5.1 Twitch Platform

The following subsections introduce the VSP used for user testing, Twitch, and the genre of audience participation games (APG). The second topic has a rather large impact on the project presented, as the goal is to combine elements of the APG field with the use of a GA to collect data about a community collaboratively using the same LLM. This implies that design practices and research done in the field of APG will be a focal point in this section.

Twitch, launched in 2011, is a streaming platform operated by Twitch Interactive, a subsidiary of Amazon, that allows users to broadcast their screen while playing games. It is the most successful video game live streaming platform with 6.9 million monthly active streamers and 2.1 million concurrent viewers in 2020 [Wolff and Shen, 2024]. In 2022 Twitch reached the number of 140 million monthly active users [Bartolome and Niu, 2023] showing its massive growth. People can share their gaming experiences and interact with others in real time. These live streams serve as meeting places for gaming communities. The Twitch streaming medium combines broadcast video with open IRC chat channels. In conjunction with gameplay, it emphasizes viewer participation and community building through the use of communication tools such as chat messages. Twitch streams range in size and nature from intimate communities of fifty viewers to massive broadcasts of tens to hundreds of thousands of viewers [Wolff and Shen, 2024]. Streamers simultaneously play games and communicate via audio and video. Participation in the streams is voluntary. All that is required to chat is a free Twitch account [Deng et al., 2015, Hamilton et al., 2014]. Additional categories apart from gaming encompasses music broadcasts, creative content, real-life content, and very individual topics such as audience participation games, in which people around the world control the player character and his actions just through text messages or influence other aspects of the game.

2.5.2 Audience Participation Games

Among various video game channels on Twitch, the Audience Participation Game (APG) channels enable viewers to participate in video games through channels instead of merely being passive audiences. Lu et al. suggest that streamers and viewers both desire interactions that are more substantial and consequential than commenting and gifting, however, doing so might compromise the user experience, making it a delicate matter to explore for streamers. Choice Chamber was the first game created specifically for the Twitch platform. A 2D platformer, Choice Chamber features an endless stream of enemies, deadly traps and hazards, and a variety of power-ups. What sets it apart from

other genre offerings is the role of the audience. As the streamer plays, polls appear periodically on the screen offering choices between different weapons, different enemy types, changes to the level architecture, etc. While other VSPs only offer support for user comments on the events, Choice Chamber uses Twitch’s built-in chat functionality in a more innovative way. Viewers simply type in their votes from a menu of options, and the winning option affects the game [Stricklin, 2019]. Twitch Plays Pokémon (TPP) is the most famous example of distributed play. With the aid of "VisualBoyAdvance", Python scripts, and an online call to action ("I’m a robot playing Pokémon, I don’t know what I’m doing, tell me which buttons to press"), users of the Twitch chat interface were motivated to play the game through their inputs [Ramirez et al., 2014]. At the height of TPP’s popularity, as many as 100,000 people were simultaneously typing commands, forcing the lead developer to add a "democracy" mode. Keystrokes became votes, and TPP would periodically execute the most popular keystroke instead of tracking 1-1 inputs. In addition, the phenomenon of stream snipers, grievers, and online trolls suggests that an open call to the director’s chair can lead to all kinds of disruptive behavior [Stricklin, 2019]. Based on this example, it becomes clear that an implementation in which users interact simultaneously with a LLM in the context of the game involves similar challenges that are therefore discussed in the focus group.

Little to no research has been conducted on enhancing APG experiences with LLMs. The following scientific work refers to a direct link between an LLM/AI and the Twitch chat, exploring researched use cases. Paliyawan et al. facilitate their interaction with AI in form of a Twitch-based audience participation fighting game, where the strengths of game AI characters are dynamically adjusted based on audience cheering and jeering. In the proposed APG, each user is assigned to either Team P1 or P2 and is tasked with supporting the AI of their respective team by pressing buttons. Users receive scores for their participation, and the names of the top ten users with the highest scores are displayed on a leaderboard. In the competitive mode, users are divided into two distinct groups, while in the collaborative mode, all users are united within a single team (Team P1) [Paliyawan et al., 2022].

A different use case sees an utilization of ChatGPT for the categorization of audience comments in game live streams, treating it as a zero-shot task. Traditional methods for comment analysis in game live streams are lacking in terms of speed and cost-effectiveness. Therefore, the scholars aim was to investigate whether ChatGPT can fulfill these requirements. Through experimental evaluation, their results indicate a majority choice of 54.34% and a human choice of 82.61%, showcasing that ChatGPT, when employed with suitable prompts, can address the aforementioned need [Li et al., 2023b]. However, since neither

example explores the deeper connections and implications of the LLM directly influencing a game retrieving audience prompts directly from the chat, the approach taken in this thesis is a novel concept for research.

Glickman et al. identify a central challenge as information-theoretic in their research: how much of the total information in the game can viewers see while watching, and how much new information can each participant contribute? As the number of players increases and the frequency of meaningful input per player increases, the ability of players to make sense of the game as a system decreases. Also the participation will break down when the number of viewers who input commands in the channel is too large. With many simultaneous input commands, it is difficult for individual players to feel their influence on the games [Striner et al., 2021]. Glickman et al. model suggests design trade-offs between the number of players, meaningful input in chat messages, and the ability to make sense. This becomes even more sensitive when considering adding a LLM to the cycle and giving it some agency over what happens during gameplay. On the other hand, there is also the possibility of using a LLM to moderate and make sense of the input, group it, and filter out unrelated, unwanted, and offensive content during the runtime of the APG. To address these challenges, this project draws on frameworks about audience motivation and participation styles. Incorporating Striner's identified themes for designing an APG into the focus group discussions and ideation phase can serve to better understand the design spaces and how to address the challenges ahead. The thematic map after Striner et al. has three sensitizing concepts - Agency, Pacing, and Community corresponding each to one corner. Themes common to Agency are live-streaming agency, interaction, choice and control. Themes common to the Pacing concept are temporality, progression, attention, distraction, engagement, gametime/downtime, and stream schedule. Themes common to the Community concept are streaming community, cooperation/competition, roles/identity, communication, privacy trade-offs, balance, streamer/viewer relationships, moderation, and power. Attention, choice, control, and engagement consider aspects of both Agency and Pacing. Engagement, streamer/viewer relationships, and balance consider aspects of Pacing and Community. Cooperation/competition, moderation, control, power, and streamer/viewer relationships consider aspects of Community and Agency. No theme considers all three concepts [Striner et al., 2021]. By providing this map during the group discussion the design of the APG and interaction concepts will be more in line with common knowledge and provide a better user experience.

Additionally, Seering et al. defines five categories of audience participants in their study: "helpers," "power seekers," "collaborators," "solipsists," and

"trolls." The motivations of audience participants depend on their goals and their individual and social agency. Having a better understanding of the different user groups that are online on the VSP Twitch is beneficial to making informed design decisions.

Audience Participant Types after [Seering et al., 2017]

- **Helpers:** Focused on assisting the streamer to win but felt they had less power over gameplay. They primarily engaged with the streamer rather than other audience members.
- **Power Seekers:** Wanted to influence the game, often helping the streamer to maximize their impact. They felt more powerful than the streamer but still somewhat limited in influence.
- **Collaborators:** Engaged with both the streamer and other audience members, valuing cooperation over personal impact. They were content regardless of the game's outcome.
- **Solipsists:** Participated for personal benefits like entertainment or learning. They felt powerful but were indifferent to the game's outcome and rarely engaged with others.
- **Trolls:** Focused on disrupting gameplay for entertainment, often switching between helping and hindering. They felt more powerful than the streamer and experimented with different roles to maximize their influence.

Based on this list, while being aware of trolls and their impact on gameplay and the overall experience, the project's interface and interaction should best facilitate productive user groups and reduce troll possibilities. In order to make participation in the game as attractive as possible, game elements and the user interface will be improved on the basis of the results of the focus group.

AI utilized in the context of Twitch and APGs

For this subsection, the Twitch platform was searched for active, running projects involving AI or a LLM. Therefore, the tag search function was used to filter out AI or ChatGPT as a topic. It was possible to identify a few applications of AI where a LLM takes the role of facilitating the content of the stream, or can be directly influenced with chat messages or donation messages. After experiencing a few minutes of the stream, the following examples were

deemed interesting.

The most sophisticated example is Claude the LLM that plays Pokémon ⁸ and involves the team at Anthropic applying their latest model in the context of playing the first edition of Pokémon. The stream shows the LLM playing the game without any special training or input from the users or developers watching. Claude can see what is happening, understand the state of the game, and make decisions using his general reasoning skills, much like a human player would navigate the world. He has a rather slow play speed due to his constant reasoning, and will sometimes use tactics like blacking out by losing every health point of his Pokémon to escape a situation he is unable to navigate.

Two projects that allow users to request a LLM directly with chat messages are first the "TrumpOrBiden2028" ⁹, which simulates and recreates a version of Biden and Trump based on their public personas debating topics suggested by viewers. Similarly, there is a channel called "Ask Jesus" ¹⁰, where viewers can ask questions to an AI trained on Jesus and the teachings of the Bible. In both cases, but especially in the first, it became clear that viewers were using the opportunity to have the AI discuss profane topics and use obscene language for their entertainment.

Finally, a project called "WatchMeForever" ¹¹ generates a sitcom with characters portrayed by AI. The channel claims that it is possible to direct the show with a special channel reward.

There are several other examples of AI integration in Twitch, whether it is used to power a chatbot replying to users messages and contributing to the overall chat experience or as content and topic in the stream itself. Recently, "Doug-Doug" ¹² a streamer had several LLM agents with a character description and abilities solve a Dungeons and Dragons game. They were confronted with numerous encounters with enemies or difficult situations to solve, and used given abilities or information about their environment provided in screenshots to solve them. It was even possible for them to argue with each other. Prior to this project, the streamer had already used a LLM several times to facilitate story-driven experiments. In one of these, viewers would be able to make decisions about conquering European states by playing against the streamer itself or an AI. A sophisticated writing model determined the outcome of tactical prompts and resources used during a player's turn. These examples show how a creative streamer is using LLM capabilities as a source of entertainment while engaging the Twitch chat in a meaningful way which is one source of

⁸<https://www.twitch.tv/claudeplayspokemon> Accessed: 27.03.2025

⁹<https://www.twitch.tv/trumporbiden2028> Accessed: 27.03.2025

¹⁰https://www.twitch.tv/ask_jesus Accessed: 27.03.2025

¹¹<https://www.twitch.tv/watchmealways> Accessed: 27.03.2025

¹²<https://www.twitch.tv/dougDoug>

inspiration of this project.

After an in-depth look at the Twitch platform, APGs, and use cases with AI on a VSP, the focus now shifts to the concept of providing non-human voices utilizing LLMs, which was one of the driving forces behind this thesis, as well as circumstances behind the game used during the user study, LoGa, provided by ITI. The game and its mechanics will be introduced later on in the user study chapter, as it facilitates the study by providing a contextual application space.

2.6 Non Human Voices with LLMs

Giving voice to traditionally voiceless entities-nonhuman species, ecosystems, and other natural systems-requires frameworks that challenge anthropocentric paradigms [Galvão-Ferreira et al., 2024]. While human language and cognition dominate the discourse, emerging research explores the potential of LLMs as mediators between humans and nonhumans. Redesigning human-AI interactions as a new approach to reshaping human-ecology relationships as AI can remind people of different earth beings and forms of existence that are not present at the decision table but may be adversely impacted [Xu and Ge, 2024]. The work of Félix et al. tackles this potential of using LLMs, as artificial assistants to facilitate communication between humans and non-humans. Building on LLMs remarkable abilities in simulating human-like reasoning and communication, emulating different sociodemographic groups and individuals, the scholars propose to extend this potential in inhabiting non-human perspectives. Specifically, they examine whether LLMs can simulate the representation of ecosystems (constituted by nonhuman entities) and influence human decision-making related to environmental issues. Efforts like these are based on a shift which is not merely speculative as studies suggest that LLMs exhibit cognitive processes distinct from human biases, enabling them to deviate from anthropocentric prioritizations [Millière and Rathkopf, 2024, Tjuatja et al., 2024]. Such divergence opens avenues for using LLMs as tools for representing non-human entities, offering perspectives not constrained by human-centric assumptions. Right now there is still a lack of practical approaches that can successfully represent the voices of natural nonhuman entities in a way that influences human behavior and decision-making [Félix et al., 2025].

Despite their reliance on human-generated data and operation based on human language and concepts, LLMs can adopt non-human perspectives through carefully designed prompts and structured interactions. For example, they have been used to interpret animal behavior and translate it into human-understandable narratives [Ye et al., 2023], facilitating interspecies communi-

cation. Beyond individual animals, LLMs can simulate ecosystem-level viewpoints by integrating environmental data (e.g., biodiversity trends, habitat dynamics) to articulate the "interests" of ecosystems [Li et al., 2024b]. For example, interactive, language-based queries to models enable real-time behavioral analysis, helping to identify ecological stressors and inform conservation strategies as described in the chapter about SES modeling. While LLMs show promise in amplifying non-human voices and fostering empathy, challenges remain, such as the potential for misinformation and the need for careful prompt design to ensure thematic depth and accuracy in narratives [Hoffmann et al., 2024].

Building on these foundation, the thesis shifts its focus from exploring LLMs that provide non-human voices in a game-based simulation to providing a tool for storytelling from non-human perspectives. If LLMs can simulate ecosystems or species in decision-making contexts, their application to narrative generation offers a creative extension: creating stories that center non-human experiences. Félix et al. note in their future research directions that AI assistants need to go beyond presenting evidence and facts. For instance, incorporating narrative storytelling, personalized feedback, and other socially reflective design elements could enhance the impact of AI assistants. LLMs have the capacity to produce speculative fiction, educational narratives, and participatory dialogues that amplify non-human voices through the infusion of human creativity and potentially with reference to ecological or behavioral data provided in the prompts. This approach not only mitigates anthropocentric bias in storytelling, but also fosters empathy and deeper engagement with environmental issues [Bardzell and Ghajargar, 2025]. The following sections details the application of LLM in the area of story writing and mediates a better understanding of the term more than human (MtH) in HCI which is frequently used in this thesis.

2.6.1 Story writing with LLMs

The use of AI technologies in creative processes has increased significantly in recent decades, with some researchers viewing AI as a functional tool, while others emphasize its role as a creative collaborator [Guzdial et al., 2019, Rezwana and Maher, 2023].

Guzdial et al. explored how designers conceptualize AI systems in creative contexts and proposed frameworks such as "co-creative" AI systems, in which human and AI agents work harmoniously together. The researchers demonstrated that user interface design can influence how artificial intelligence is perceived in the context of a creative partnership, namely as a friend, collaborator, or manager. Furthermore, as AI technologies have advanced, language

models such as GPT 4 have been applied to the domain of narrative co-creation, allowing AI to function not only as an assistive tool, but also as a storytelling companion or partner [Ghajargar et al., 2022, Rezwana and Maher, 2023]. Ghajargar et al. conducted a comprehensive analysis of generative literature projects that used LLMs in poetic, narrative, and artistic compositions. The conclusion was that these models facilitate the emergence of alternative narrative voices and offer expressive affordances that were previously impractical without computational assistance. The aforementioned technologies have also been applied to interactive or generative literature, allowing users to see, read, and converse with AI via natural language, as Ammanabrolu et al. explores in their AI Dungeon, a system that uses transformers to facilitate real-time, open-ended storytelling, allowing users to directly address the LLM. Their work demonstrates how storytelling can become immersive and personalized when driven by artificial language generation. In addition, Zhang et al. developed a co-creative storytelling tool to encourage interaction between parents and their children. Their research found that AI tools can scaffold imagination and encourage co-constructed narratives while promoting early literacy skills. Linked to this, Čavojský et al.'s work found that LLMs are particularly effective in creating engaging educational narratives that embed ecological data and foster environmental empathy. By simulating conversations with non-human agents or speculative futures, LLMs can help learners develop an emotional connection to environmental systems and challenges [Čavojský et al., 2023]. Invoking this specific connection and awareness through participatory storytelling is already explored outside of using a LLM by the scholars Talgorn and Ullerup in their work Invoking 'Empathy for the Planet' through Participatory Ecological Storytelling: From Human-Centered to Planet-Centered Design. In their initiative, design students and professionals co-created stories using human and nonhuman personas. The resulting narratives revealed that empathy was cultivated through emotional projection and imaginative identification, demonstrating how storytelling acts as a vehicle for self-reflection, collective ecological sense-making, and the integration of more-than-human voices into design practices.

As previously mentioned researchers are tackling this goal with the support of technologies. A significant concern in this kind of ecological storytelling is AI bias, particularly anthropocentric framing in natural language generation. A pioneering study found that LLMs tend to default to human-centered perspectives even when discussing nature or environmental contexts [Hupont et al., 2024]. Additional to this finding other AI biases like speciesist-bias termed by the researcher Hagendorff have to be acknowledged and are addressed in the ethical considerations chapter.

At this point, the use of emotionally resonant storytelling that includes non-

human protagonists could significantly enhance public engagement with ecological issues. If LLMs are trained or guided to operate within this framework, they can become powerful tools for reshaping environmental discourse and inspiring behavioral change. The effort and approach of this thesis, which aims to build on previous referenced work, adapts an LLM in the role of a GA to specifically facilitate participatory ecological story-writing with the creative input of multiple participants simultaneously. With the goal of shifting the perspective to MtH while creating such narratives, the game LoGa, which incorporates these concepts, is used during the study. It exemplifies the efforts of ITI and other researchers to apply MtH in HCI contexts, which will be illuminated next.

2.6.2 More than Humans

Traditionally, HCI has prioritized human needs. However, emerging frameworks advocate for MtH approaches that center ecological entanglements. By designing playful, immersive experiences that disrupt anthropocentric norms, HCI can foster deeper emotional engagement with nature and motivate conservation efforts [Galvão-Ferreira et al., 2024]. These designs shift the focus from human outcomes to reciprocal relationships with non-human entities, aligning with posthumanist theories that dissolve the rigid boundaries between humans, technology, and ecosystems [Forlano, 2016].

Integrating technology into natural environments requires balancing enhancement with ecological sensitivity. For instance, gamified elements like navigation puzzles or scavenger hunts can encourage exploration while preserving the sanctity of outdoor spaces. This approach fosters a sense of accomplishment and connection with nature through meaningful engagement [Galvão-Ferreira et al., 2025]. However, as demonstrated by Nisi et al., technology’s role goes beyond gamification; it can also enable storytelling from non-human perspectives, which is crucial for reimagining human-nature relationships.

The board-based storytelling game LoGa exemplifies this shift. Set in the biodiverse landscapes of Madeira Island, the game invites players to embody non-human entities, such as birds, marine mammals, and volcanic rocks, to explore the entanglements of nature and culture. The Madeira Islands are an exuberant Portuguese volcanic archipelago in the Atlantic, renowned for its rugged landscapes, rich biodiversity, and centuries-old natureculture heritage. Through participatory play, LoGa challenges players to confront the agency of non-human actors and deepen their awareness of ecological heritage [Nisi et al., 2025].

The potential of tools like LoGa is amplified even further when combined with LLMs. While gameplay can immerse humans in non-human perspec-

tives, LLMs can extend this concept by generating narratives that articulate the "voices" of ecosystems or species [Čavojský et al., 2023]. This thesis explores such an approach. Such applications align with the broader aim of MtH-HCI: reconfiguring human-environment relationships by de-centering human agency and enabling sustainable sociotechnical imaginaries [Coskun et al., 2022]. Integrating LLMs into participatory design (e.g., generating dynamic stories for LoGa) allows HCI to bridge speculative storytelling with actionable ecological empathy contributing to giving nonhumans more representativity. After looking into the field of representing non human entities and MtH I want to address some ethical concerns next regarding LLMs usage and in the specific context of this thesis.

2.7 Ethical Concerns

This section focuses on arising ethical concerns regarding the employment of a LLM agent, the representation of a non-human character by it, and the exposure to a community on a social platform like Twitch as well as biases which accompany the employment of a LLM in general.

Employing a LLM Agent on a Platform like Twitch

There are numerous ethical considerations which have to be taken into account while conducting a user study with a LLM involved in such a novel environment such as a VSP. Deploying an LLM agent in an unmoderated, live-streamed setting exposes it to unpredictable human behavior, including attempts to "jailbreak" the system or elicit discriminatory responses. Historical examples, such as Microsoft's Tay, which quickly adopted harmful language when exposed to Twitter users, highlight the dangers of uncontrolled public interaction [Wolf et al., 2017]. While moderated environments can mitigate these risks, scaling to unsupervised deployment requires robust safeguards. As there will be moderation in place as of now, protection against harmful inputs to the LLM Agent and in the chat will be ensured. However, the goal might shift to letting it run without supervision, making it necessary to think about methods to limit the possibilities of it turning into a harmful experience.

- **Input Filtering:** Real-time detection of harmful prompts, such as hate speech and speciesist rhetoric, using hybrid human-AI moderation tools [Hu et al., 2024b].
- **Output constraints:** Preventing anthropomorphized or reductionist responses (e.g., framing ecosystems as "resources" for human use) through

fine-tuning or post-processing [Hagendorff et al., 2023].

Representational Biases in Non-Human Narratives

LLMs inherently reflect the biases of their training data, which prioritizes human language, Western perspectives, and utilitarian framings of nature (e.g., "ecosystem services") [Grasso et al., 2025]. A critical challenge lies in mitigating anthropomorphic bias. The training data and inherent structure of human language may lead to the projection of human traits onto nonhuman entities. This bias could distort the understanding of animal experiences, potentially undermining the validity of the insights generated by speciesist biases. Key concerns include:

- **Anthropomorphic Distortion:** The projection of human emotions or intentions onto non-human entities (e.g., a "lonely" tree or a "vengeful" ocean), which misrepresents their lived experiences [Millière and Rathkopf, 2024].
- **Speciesist bias** is the systematic prioritization of human interests, even in discussions about animals or ecosystems [Hupont et al., 2024]. For example, LLMs may frame biodiversity conservation as a human benefit rather than an intrinsic right of non-human entities.
- **Cultural homogenization:** Overreliance on dominant languages (e.g., English) can erase indigenous or local ecological knowledge [Kamran, 2023].

Using LLMs to "speak for" non-humans raises questions as posed by Forlano "Can an LLM, which is trained using human data, legitimately represent non-human perspectives? Or, does it perpetuate anthropocentric ventriloquism?" The listed ethical concerns do not cover all relevant fields, but they provide an overview of the topic and ensure that these concerns are recognized and addressed when necessary. At this time, the user study is being conducted in a relatively controlled setting, and the presence of harmful content is not anticipated. Nevertheless, if any instance of speciesist bias, or anthropomorphic distortion is identified in the created texts, it will be highlighted.

The next chapter will explain the focus group approach used during the interface creation process to ensure a well-thought-out, participatory LLMGA interface combined with the LoGa game.

Chapter 3

Focus Group Study – Evaluation and Findings

3.1 Introduction

This chapter presents the rationale and approach for conducting a focus group with stakeholders. The primary objective was to refine the mechanics and user interface of the LLM agent to ensure that it enhances audience engagement on Twitch when integrated into the APG version of LoGa. To accomplish this, the focus group explored the following central question: **How can an interactive LLM agent support the LoGa’s gameplay experience on Twitch to make it more fun and engaging?** By gathering insights from a diverse set of stakeholders through discussion, the study aimed to identify improvements in interface design, and AI interactions to enhance engagement and playability on the VSP Twitch. The following sections detail the methodology, key findings from the focus group and implications for the planned user study.

3.2 Methodology

3.2.1 Study Design

The decision to conduct a focus group is made on the basis that it seems beneficial to have multiple parties affected or with influence discussing the possibilities and design choices in a moderated setting. Focus groups are a qualitative research method that involves gathering a small group of people to discuss their perceptions, opinions, and attitudes towards a product, service or topic [Kitzinger, 1995, Morgan, 1996, Pitariu et al., 2009]. This method is particularly useful for understanding user needs and preferences early in the

design process [Krüger et al., 2020, Pathiraja Rathnayaka Hitige et al., 2022, Pitariu et al., 2009]. The method is part of user-centered design approaches, and the interactive and synergistic nature of focus group discussions (FGDs) allows a moderator to promptly clarify questions and probe for further explanation, thereby creating a productive environment for participation [Wong, 2008]. Through an exchange of ideas and opinions, FGDs can reveal a range of participants' perspectives, creating a broader understanding of design spaces [Pathiraja Rathnayaka Hitige et al., 2022].

Traditional focus groups rely on the skills of the moderator who is a relevant factor for the success, and require extensive and time-consuming data analysis afterwards [Krüger et al., 2020]. Another obvious weakness is that an FGD can be dominated by more vocal participants, whose strong and loud voice can misrepresent the sentiment and hide the voices of the silent people [Wong, 2008]. Therefore, creating an open and free discussion environment that allows each participant to express their views and opinions is a top priority Pathiraja Rathnayaka Hitige et al. [2022]. Pitariu et al. guidelines are helpful in conducting a FGD. For excellent results, a diverse group of participants who are stakeholders or represent the target users is necessary. The group size is usually six to ten people to allow for meaningful interactions and enough space to express opinions. The facilitator is able to guide the discussion, encourage participation, and manage group dynamics. A discussion guide with open-ended questions aligned with the research is helpful for a smooth process. To ensure meaningful analysis, the session should be recorded with the consent of the participants.

However, as Krüger et al. mentions, focus groups are not specifically designed to gather insights for experience design. Their suggestion of experience focus groups is an interesting concept to extend the traditional method. In fact, the group principles allow enough freedom to structure and tailor them to individual research. Using this approach better facilitates the incorporation of augmenting materials, such as Striners thematic map, to guide discussions and stimulate idea generation. In addition, the use of open-ended questions and informative material in this setting encourages the inclusion of diverse perspectives, as each individual has a unique level of involvement and expertise in one of the project's domains. With the support of ideation methods such as "braindumping" and "plus five", which are interactive tasks to create visual expressions of user experiences, participants partially document their results, thereby reducing documentation and interpretation time. The defined structure with its specific interactive tasks and the alternation between individual and collaborative work of the participants creates a dynamic and easier to manage scenario for the moderator [Krüger et al., 2020].

The FGD was partially an open discussion with open ended questions but also

incorporated moments of individual work which created visual representations. This process was aided by maps and resources to support participants in the process.

3.2.2 Participant Selection

To incorporate all topics addressed following stakeholders were part of the focus group:

1. LoGa team members, two people responsible for designing multiple versions of the physical game and collecting the underlying data also responsible for writing a scientific report "Rock, Paper, Pigeon: Playing with Posthuman Narratives for Natureculture Storytelling".
2. Coding team consists of one person, tasked with developing the digital adaptation and providing an API for communication with the LLM agent. He is also part of the Twitch experts.
3. Two Twitch experts, users with in-depth knowledge of the platform and streaming dynamics.

Recruitment was done verbally, and participants from LoGa and Coding were selected based on their major roles in each pillar and their importance in providing a completed APG. For Twitch experts, close acquaintances were asked about their knowledge and subsequently asked to participate if the requirement of having used Twitch regularly for watching or streaming themselves was met. The reason for this is that a novice user is aware of the chat interactions and dynamics that occur during a stream, making them an important voice in the discussion of audience engagement. In the following sections, Twitch experts are referred to as TE and LoGa team members as LE.

3.2.3 Procedure

The focus group with four participants commences with handing out consent forms followed by an introduction to the three-pillar structure, which comprises the interface of the LLM Agent, the VSP Twitch, and LoGa the game. Each pillar is introduced within a presentation that illustrates the current state and potential applications. This approach facilitates a comprehensive understanding of the aspects that are not familiar to the participants before engaging in an open discussion. Then the underlying question is posed to the participants and time for discussion is given. Following this, participants engage in ideation using the methods mentioned in study design. The first stage is to write down



Figure 3.1: Focus Group during the Method Braindumping

everything that comes to mind, without filtering or rationalizing, in order to collect possibly all ideas. In this step, participants work together on a poster to collect their ideas, which they write on post-its. In the second step, each participant chooses one idea that they particularly like and adds five ideas that contribute to and expand on it. Each participant is then given five minutes to give a short presentation of his or her idea and explain how it contributes to the overall question posed. The FGD ends with a final moderated discussion of the results, which lasts a few minutes. Afterwards, the participants are thanked for their participation. Data are collected in two forms. The first is a voice recording of the discussions and the second is the visual material created during the FGD. The recording was done with a phone and contains most of the discussions. Unfortunately, much of the open discussion was dominated by clarifications on each topic, as the LoGa team was mostly unaware of the Twitch features, while the other team had a lot of questions about the game, so this part of the recording is mostly ignored during transcription and thematic coding.

3.3 Results and Evaluation

This section presents and interprets all identified themes which were found during thematic coding of the transcripts, as well as an ideation evaluation. The structure of this section reflects categories identified and the associated patterns and perceptions of the focus group participants.

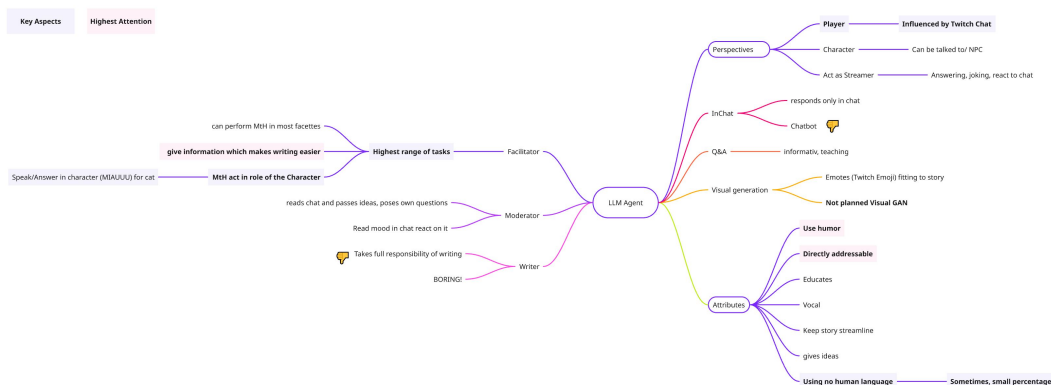


Figure 3.2: Focus Group Results: LLM Agent Roles identified & Attributes, Perspectives

There are multiple roles advocated to the LLM with some attributes reappearing or playing a larger role for the agents purpose while having different range of tasks. Being the **facilitator**, the participants gave it the purpose of supporting creative input or sharing information about the chosen character to deepen knowledge and provide indicators what to write about. "It can comment the action while staying in the character" was suggested by TE1. LE1 also strongly advocated for having it factor in less human centric perspectives. That can come to bear if the story needs correction in terms of ecological facts. If present as a facilitator the LLM supports writing the story, narrate events and fill gaps in the narration. It sums up ideas posed by the audience. TE2 sees additionally the possibility of giving reactions from the audience feeding it to the streamers' attention and clarifying questions about prompts of the story.

Very similar to this role, but with less range of tasks is the **moderator** role, in which the LLM is used to communicate Twitch chat ideas for prompts to the streamer. To do so, it constantly monitors the chat and can pose clarifying questions through the provided interface to bring something to everyone's attention. TE2 came forth with the suggestion: "It could analyze the messages from the chat to determine the current mood and react accordingly on it". LE2 sees in this role a bigger potential to improve fun and collaboration

during game play. Like the facilitator, each story produced is read aloud at the end.

In the role of a **writer**, the previously described tasks closely related to creating the story, such as narrating, filling in gaps, are also served, but as LE1 and LE2 point out, delegating all writing and creating tasks to the LLM will make the experience dull and boring. More interesting are suggestions to let him continue the already started story of the streamer, which could include a turn taking dynamic. A statement that can be made indefinitely is: The creative step must remain with the players, supported or not supported by the LLM. Three smaller roles identified are "**InChat**", "**Q&A**" and the role of **generating other content**. For the first two, both should answer questions and teach in an interesting way, just through a different interface. The InChat appears in the role of the character chosen by the streamer and suggested by LE1 could write in character, meaning that under certain circumstances it will just write gibberish imitating its character in the chat. For the last role, in which new content should be generated, TE1 suggested creating emotes that fit the storyline and events, which could then be used by the audience in the chat to show their excitement. During the discussion, it emerged that the LLM could be present as the chosen character, meaning it could be talked to during gameplay and have a different kind of dialog similar to an NPC. It could instead also take on the role of a player who is heavily influenced by the Twitch chat and therefore plays for everyone watching and creates a story according to their will, or it could take on some responsibilities while acting as a streamer, answering questions, exposing stupid comments, and making jokes to lighten the mood.

Attributes that were discussed for several roles include using non-human language to facilitate the MtH approach, contributing to the writing of the story, and getting these ideas from the Twitch chat. A recurring and central theme was the LLM using humor to answer questions or react and comment. This was discussed because LE1 wants to reduce the effect of appearing as a teacher. However, this attribute was mentioned several times as being very influential in audience perception and interest, making it a possible target for the agent. Additionally the LLM being able to keep the story more streamline and taking a MtH perspective seems most promising.

Category: Story and writing

Writing was defined by LE1 and LE2 as a core element of the LoGa game. In addition, supporting the narrative with visuals was seen as an attention-grabbing concept. There may be gaps to be filled in participants' stories based on prompts provided to the agent. A selective process could influence players regardless of the viewers playing through the LLM or the streamer to increase

the effort to possibly write the best story or part of the story cf. (LE1). Another mechanic suggested by TE1 is to have the streamer and LLM take turns completing each other's story or following up on the last sentence, creating a highly interactive experience.

Category: Twitch and viewer participation

Discussing Twitch and its viewers, the possibility of letting them control the LLM output by prompting in the Twitch chat seems a likely scenario. The Twitch chat can then play against or with the streamer by influencing the story with their ideas. TE1 described an interesting phenomena on Twitch, where viewers can create the entertainment themselves, which is then only adapted by the streamer. This can be very powerful because the community is engaged with itself and everyone strives to contribute. Therefore, the chat should be a hundred percent involved in the creative process of writing the story. This is most likely to work with funny content, which has a higher overall success rate in holding attention cf. (TE2). Other concepts for interaction involve the audience voting on parts of the story to decide how to continue or which finished story is the best. The audience could also take part in the turn taking and continue already started stories. They will respond to the LLM by producing text in the chat or answering questions, depending on whether the answer is purely informative or in MtH style. The LLM answering fully in character could generate more interest and have the audience trying to figure out how to understand or communicate, but could also lose all interest if done incorrectly. Another volatile idea from TE1 is to have the audience collect emotes based on the stories, which speaks to the fascination of receiving individual achievements during the game.

Category: More than Human

LE1 describes this as the LLM taking on the role and perspective of, for example, the Monk Seal, which is a regional animal of Madeira. When its language is adapted or the focus is shifted to its sensations and goals, the streamer and audience must adapt as well, fostering a deeper understanding of the MtH world. Incorporating this strategy can support creative writing with MtH input provided by the LLM cf. (LE2). Adapting a particular style to respond to queries and being aware of the ecological facts woven into the story are two other issues that came up in the discussion. An opportunity to point out the differences between a human biased perspective and a MtH perspective is described by LE2 when comparing the finished story of the streamer with the one written with the support of the LLM. This is based on the assumption that the LLM has a better ability to adapt MtH perspectives than the other players.

Category: Playing experience

The previous playing experience was described by LE1 and LE2 as lackluster of excitement and they would hope for a more fun and collaborative environment. This can be facilitated by having multiple parties play together, however, there is the need of engagement during play time especially the longer periods of writing are dull and will not work well with audience engagement cf. (LE1). There is the possibility of pairing the streamer with other players as well as have the viewers take part by playing through the LLM cf. (TE1). This could unfold in a competitive manner trying to create the most liked story or cooperative, completing each others story with creative inputs. Overall the pace and rate of interaction should be increased by having the LLM comment or read out the story part after each stage. This will result in more exchange along all participants. This goes hand in hand with the Idea of TE2 to swap the stories after one sentence creating a more engaging and dynamic course of play.

Having the viewers playing through prompting the LLM poses some questions involving the playing experience. Mechanics to pick a random user or divide user in their respective audience participation group and categorize prompts are strategies which have the possibility to provide a smooth and productive playing experience. Important would be according to LE1 to swap active players from the audience after each turn. An other mechanic which can increase viewer detention are unexpected events. TE2 describes them as the LLM reacting to certain prompts in a surprising way to an unlikely percentage. This is similar to the idea of the LLM adapting MtH language for a low percentage during interaction. Other ideas involved having a mystery, respectively, charade game on the hand which core idea was to have the participants try to fool each other in believing that they are the AI writing the story which poses an interesting concept and has an uncommon focus to shift the imitation from the LLM imitating humans to humans trying to appear as LLM which is a paradox to recent developments cf. (LE2).

Category: Streamer

The streamer was identified as an important role in maintaining attention, and his responsibilities were described by TE1 as not only playing a game, but reading the chat, answering questions, changing the gameplay, and adapting to what is occupying the chat. For example, if there is a discussion going on, the streamer might pause the game, look up a particular topic, and join the discussion cf. (TE2). He must be addressable at all times and could play against the audience in the case of an APG. Looking at the role of the streamer, it is inconceivable that he can be treated similarly to the original players during the writing process. Writing without involving the other players, the audience

playing with the LLM results in boredom for all parties.

Category: identified risks

Lastly, some risks were touched on, namely the possibility of trolls writing in the chat, which is one of the audience participation groups trying to disrupt the stream. The other is unproductive prompting that does not result in the game progressing. Risks were a really small part of the discussion, which means that the design process will be inspired by the APGs presented in related work and their decisions to minimize the described behaviors.

3.4 Summary and Implications

This section summarizes the key insights and the most prominent points to include in designing the LLM integration with the APG and mentions the next steps taken to realize the planned user study.

3.4.1 Key Takeaways

After the FGD, it is clear that the chosen model will have an impact on the key feature of using humor and creating humorous content, as well as contributing the MtH perspective. Ideally, the LLM agent will provide more knowledge about the played character and take care of adhering to ecological facts in the story. There are several functionalities that need to be added to make the experience enjoyable and interesting. MtH can be present with spontaneous output in the role of the character. Prompts need to be collected from the audience in the Twitch chat and need to be recognized and handled with specific mechanics to provide the best experience for the productive audience participation groups. LoGa could benefit from some substantial changes that provide a more dynamic and optimized sequence in which engagement with each other is promoted. This can be done by swapping after a sentence and reading out story parts for each turn. The Twitch chat needs to be involved in the creative process of writing the story, which can be done by applying the LLM response to the game interface so that the agent writes a story based on Twitch chat prompts while the streamer does so as well. To reduce confusion and inflated prompts, it seems necessary to cut out some of the provided prompts for story writing, banking on the fact that the viewers will provide some of their ideas. This would also require an integration of the LLM and its interface to the game logic.

3.4.2 Next Steps

For development a few things have to be adapted in the code however due to time constraints and it being a rather first time approach, game changes are kept to a minimum meaning the character choice was overhauled with multiple non human choices added. Moreover, an overview of the full story at the end was included thanks to the coding team of LoGa. The model and its desired capabilities of using humor proficient and MtH perspectives were found through a process which will be specified in the following chapter.

Enabling viewers to take part in story creation and adding their creative inputs is realized with the Twitch API integration to the code and appropriate filtering techniques, both of which are explained further in the development process chapter. MtH and humor will also be specifically targeted in the system prompt which gives the chosen LLM the instruction at hand how to function as the GA. A more detailed look follows in the next chapter into the system prompt.

Chapter 4

Interface for Remote Multiple User Prompts

Disclaimer: The code shown in this chapter was written supported by LLM usage and Github Copilot trouble shooting and is based on existing code ¹

This chapter presents the comprehensive methodology used to enable and evaluate the capabilities of LLMs as facilitators of an APG in which the audience uses multi-prompting to create an ecologically based narrative. The goal is to enable multiple users to remotely contribute to a prompt and thereby participate in an activity, such as a game. The approach includes the proposal of a novel method for LLM selection, subsequently tested during this thesis called "Quick Model Choice" (QMC). The second part of the chapter includes all development steps taken and code explanations for providing an interface for user interaction with the LLMGA.

4.1 Model Choice - Hein's Quick Model Choice Approach

The in Fig 4.1 portrayed method is a code-free, moderate-effort seven step process of selecting a LLM from the widely available models with the aim of filtering out the one that excels in certain areas important to the researcher's goal. It makes use of prompt engineering to refine the base prompt, which combines general purpose with attributes for which a chosen model (M) is specifically tested. The base prompt can be refined to achieve a higher satisfaction with the response (R). Following the collection of responses from each model selected for evaluation, the refinement process commences with

¹<https://github.com/DougDougGithub/Babagaboosh> Last accessed on 18.03.2025

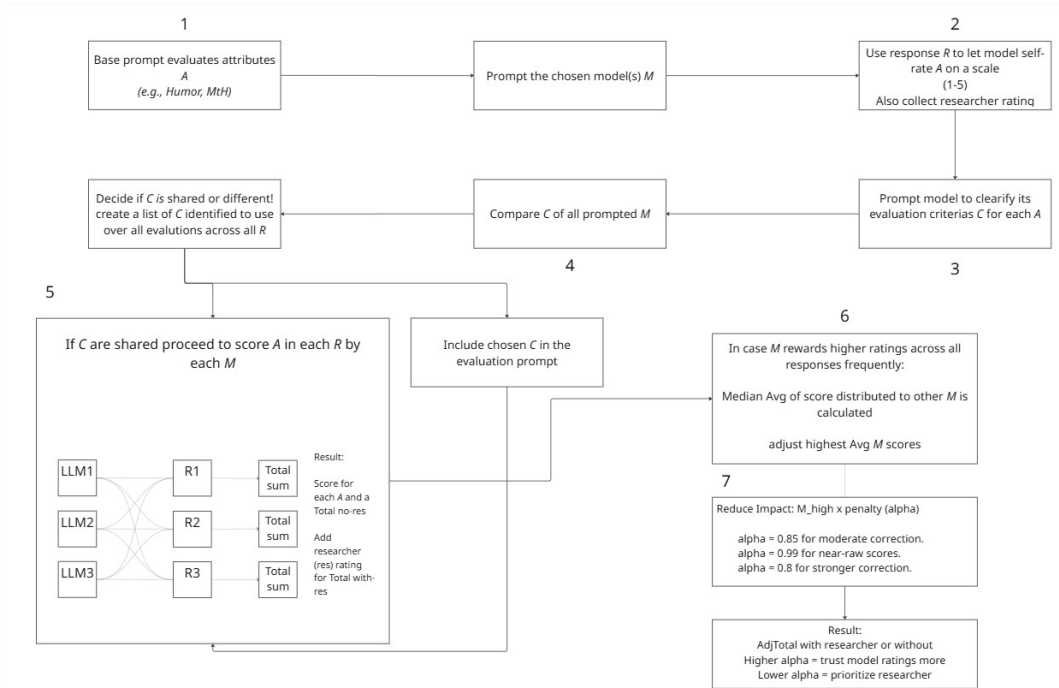


Figure 4.1: Quick Model Choice Approach Visualization

the models rating their own responses based on the required attributes (A) using a Likert scale. In the step three and four, the criteria (C) of the model on which the rating was founded are acquired and comparatively analyzed with the criteria of the other models. In the event of a significant deviation in the criteria, the researcher is tasked with constructing a common list of criteria, as an intermediate step between four and five, from which the evaluation of the responses can commence. This is accomplished by selecting criteria that are common or, in the researcher’s subjective opinion, more significant, to be included in the final rating. The subsequent phase, designated as "assessment," involves the provision of a consistent task for the evaluation of every response for each model. The objective of this evaluation is to assess each output generated based on the base prompt. The specific attributes within the responses will be analyzed using a selected Likert scale. (see Step 5). It is important to note that the preceding steps may necessitate varying degrees of effort to adequately articulate the task demanded by the LLM. Consequently, the implementation of multiple iterations of prompt engineering may be warranted. Moreover every time a prompt is sent it was in a new conversation to prevent any biases and side effects from previous responses. The result of step five is a score for each model provided by all other models for each attribute utilized in the response. If it is deemed advantageous to factor in a human and more

subjectively opinion the researcher rating is added to the score. The ultimate outcome of this process is a synthesis of all ratings, thereby providing a comprehensive perspective on which model demonstrates the greatest proficiency in utilizing the desired attributes. It is possible to complete the calculation and selection process after step five. In the event of LLMs awarding higher ratings with a greater frequency than lower and average ratings, and in the event of significant discrepancies between the researcher's score and that of the LLMs, it may be advantageous to implement a penalty factor (alpha) ranging from 0.8 to 0.99 on LLMs who have allocated average ratings above a certain threshold or have demonstrated a marked tendency towards high ratings. By adjusting alpha the confidence in the evaluation of the LLM can be varied and expressed. In order to achieve the adjustment from step seven, the mean rating is calculated and recorded as the median in step six. Subsequently, the penalty is applied to identified models (M_high), resulting in the correction of their scores and the production of a more robust result. The final score can be considered irrespective of the researchers' opinion, and it provides a statistical and thorough method of selecting a LLM to use in owns research tasks.

Prompting

Prompting makes up a substantial parts of this method. In the case of this thesis, the models were specifically asked to use humor and represent the MtH perspective when responding to a query. For this purpose, the task was described similar to the one used in the final application, with additional inputs simulating the messages of the viewers such as: "It has the size of a house, its yellow, it loves fish, it contemplates to leave or kill itself, how much is the fish, are you single?", in order to test the handling of several more or less independent and serious prompts. This simulates and aligns close with the reality of stream chats which are made up of previous discussed participant types like helpers and possibly trolls after Seering et al.. In the previous described FGD the two attributes namely humor and MtH were made out to be of most importance for providing an engaging, motivating experience while being educated further on ecological facts and experiencing a shift in perspective. Therefore the focus is set on discovering the most capable model in these disciplines thereby including their application in the base prompt. As consequence, the LLM was explicitly asked to use humor to expose and debunk ecologically false facts possibly included in the prompts, and to tell the story from the perspective of a chosen non-human character and reduce human bias while doing so. The base and all other prompts are also included in the appendix for closer inspection (see A.4). In the described process the initial base prompt worked so well there was no iterations with prompt engineering conducted. To make

sure the instructions are fully recognized the task to specify the role of the LLM itself was included at the end. For the evaluation task and especially for refining and aligning the criteria some iteration of the prompts were used and every prompt sent during this method was placed in a new conversation with the responding model.

Evaluation

In the first iteration of this proposed method the evaluation was done before aligning any criteria which resulted in a less transparent result. Every model rated the responses of each other model and itself on the basis of humor incorporated and MtH perspective taken on the common Likert scale between one to five. This was instructed by a zero shot prompt that included the task of rating the attributes separately and providing context on what basis, which values to use, and their meaning. Some prompt engineering was conducted to alter the results and provide more context which was deemed to be helpful. In the prompt described an explanation like following was added: "The text provided is generated by a LLM based on a prompt with multiple suggestions for a description of a monk seal some more serious some less".

Refinement

After consultation, previous inquired criteria from the models were compared and larger differences for the analysis in MtH and some for humor were noticed. Due to this discrepancies, the previous result was declared as unverified and a list of shared and desired criteria was built on which the evaluation took place after. For the attribute humor the criteria of witty, timing and pacing, originality and creativity, surprising and relevance and context were provided. For MtH the criteria were ecological, non human bias, environmental challenges, unique subjectivity, and critique. When clarifying which attributes should be examined and rated on the specified criteria, it was advantageous to include examples for each criteria in the evaluation prompt which were mentioned previously by one of the models. So for the criteria witty following description was included: "Incorporates clever wordplay, puns, and dialogue that makes the reader laugh out loud. The humor should be sharp and memorable". After refining the prompt, for evaluation the rating commenced as shown in Fig. 4.1 step five, resulting in scores for each attribute awarded by the LLMs and the researcher.

Table 4.1: Combined Scores for Language Models With and Without Researcher Input

Model Evaluation Scores			
Model	Model Only Score	Researcher Score	Total With Researcher
Gemma3 27b	43.0	5.0	48.0
Qwq-32b	48.5	7.0	55.5
Deepseek-r1	46.0	6.0	52.0
Mistral 24b	39.5	6.0	45.5
Llama3.3 70b	46.0	6.0	52.0
Qwen2.5 72b	37.5	4.5	42.0

Calculation/Selection

The calculation process comprises the aggregation of all scores for each LLM in the designated attribute column. Subsequently, the mean score awarded during the evaluation process is calculated for each model without and with the researchers awarded scores (see Eqn. 4.1 & 4.2). The i is the instance/table and j refers to the model.

$$\text{Total}(T_i) = \sum_{j \in \text{Models}} (\text{Humor}_{i,j} + \text{MtH}_{i,j}) \quad (4.1)$$

$$\text{Total}_{\text{with-res}}(T_i) = \text{Total}(T_i) + (\text{Humor}_{\text{res}} + \text{MtH}_{\text{res}}) \quad (4.2)$$

Additional to this preliminary result from Tab. 4.1 it is possible to ascertain which model has given the highest rating across its evaluations, both as the mean and the median. In the case of this evaluation, the over all median score of 4.0 was obtained from two models, namely: "Mistral 24b and Llama3.3 70b" (see Fig. 4.2 for full results).

This led to the application of a penalty in a subsequent equation 4.3, as these models were considered to have an overly positive evaluation and assessment, eminent by comparing their sum of values with those of other models and the researcher. The models in question allocated a total of 47 or 48 points to other models during the evaluation process and gave at least one time the rating of a perfect five. This represents an increase of 10 to 11 points compared to Qwen2.5. In this particular instance, the value of 85% was identified as the threshold for confidence in the score derived from these two models. Subsequent to this determination, a penalty of 15% of the initial value is applied.

$$\text{AdjScore}(T_i) = \text{Total}(T_i) - \alpha \times \sum_{j \in \text{M_high}} (\text{Humor}_{i,j} + \text{MtH}_{i,j}) \quad (4.3)$$

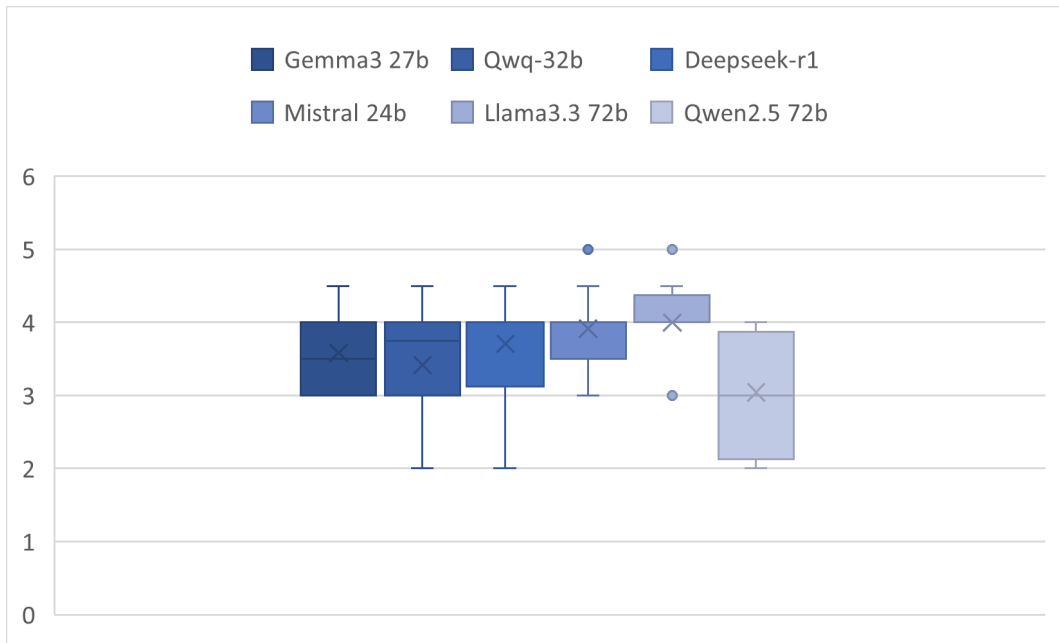


Figure 4.2: Box Plot for Average Rating given

Example Adjustment for the score of model Qwq-32b:

$$48.5 - 0.15 \times 16.5 = 46$$

The adjustments of the two models evaluations resulted in Tab. 4.2. As the percentage increases, the confidence in the evaluation of the model is incorporated.

Several results can then be read from this method: The best performing LLM for each attribute, the best overall performance which was aimed for and for both cases, the opinion of the researcher can be added to receive an additional perspective. The end results are presented in Tab. 4.3

Limitation of QMC

This approach has some of the usual limitations in applying LLMs' ratings to themselves and to each other. Attributes, even with provided examples, can be interpreted differently and thus have different agency during the rating process, and results are always based on the provided descriptive task. With open CoT processes, these limitations are reduced by the ability to trace each argument and rating back to the specific part of the prompt, but there is still some uncertainty about how the overall result and rating is composed. Furthermore, the risks identified by Li et al. include LLMs and their inability

Table 4.2: Adjusted Scores for Language Models With and Without Researcher Input

Model Evaluation Scores		
Model	Adj. Score	Adj. Score (res)
Gemma3 27b	40.6	45.6
Qwq-32b	46.0	53.0
Deepseek-r1	43.5	49.5
Mistral 24b	37.5	43.5
Llama3.3 72b	43.3	49.3
Qwen2.5 72b	35.0	39.5

Table 4.3: Final Evaluation Results

Metric	Model	Score
Highest total (without researcher)	Qwq-32b	48.5
Highest total (with researcher)	Qwq-32b	55.5
Models-high (median)	Mistral 24b, Llama3.3 70b	4.0
Adjusted total ($\alpha = 0.85$, no-res)	Qwq-32b	46
Adjusted total ($\alpha = 0.85$, with-res)	Qwq-32b	53

to access or incorporate up-to-date information reliably. They might generate incorrect or fabricated information with high confidence and lack the depth of understanding required for specialized domains.

Hein’s QMC framework proposal

In light of the aforementioned steps and the necessity of selecting a model that aligns with one’s research objectives or a model that exhibits a high degree of proficiency in the application of specific techniques and the utilization of language patterns, I propose a streamlined framework comprising three steps, as illustrated in reference figure Fig. 4.3. The aforementioned method has the capacity to prolong the decision-making process. It can be utilized as a rational basis for goal-aligned model selection rather than merely opting for the most recent iteration of a model within one’s own project.

4.1.1 Game Agent

Based on the evaluation shown in Fig. 4.1 and the results in Fig. 4.3 the model Qwq-32b from the Qwen Team[Team, 2025] was chosen to be the GA

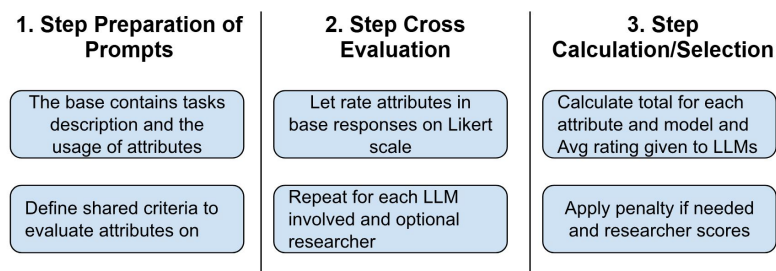


Figure 4.3: QMC - simplified

for its usage of humor while incorporating a MtH perspective most successful. Qwq is the reasoning model of the Qwen series. Compared with conventional instruction-tuned models, Qwq, which is capable of thinking and reasoning, can achieve significantly enhanced performance in downstream tasks, especially hard problems. Qwq-32b is the mid-range reasoning model that can achieve competitive performance against state-of-the-art reasoning models, such as DeepSeek-R1, o1-mini [Team, 2025, Yang et al., 2024a]. It combines therefore the most desired capabilities with CoT and is therefore a suitable choice for being GA who is tasked to weave the given prompts/ideas from Twitch viewers into the story while portraying a MtH perspective and using humor to defect ecological wrong facts brought forward by the participants. CoT will be important for later analysis of whether the model was aware of all the inputs, whether it actively decided to ignore an idea, and whether rules were handled during processing. As described in related work and drawing from examples of LLMGAs a system prompt provided up front will determine how the interaction is structured and what tasks the LLM has to follow while being in the role of GA.

The system prompt utilized during the deployment consists of a basic description mentioning the role with an explicit example, in which context all the other interactions are going to happen, and a list of rules to follow while creating responses (see A.5). In a few tests and prompt iterations this was made out to be most effective. Some rules were specifically added to reduce noise in the answer of the GA. This noise was noticed during test runs in which the produced text and the audio version contained confusing and cluttering elements like the overproportional usage of asterisks which would also be spoken out and inserted passages. Furthermore, the usage of dramatic and outdated words was reduced, and the LLMs behavior to lean towards writing a movie script was ironed out by mentioning the story is supposed to be more like an entertaining book while being easy to read. In order to incorporate the MtH perspective, the model was instructed to narrate the story from an MtH view-

point, thereby reducing human bias. Additionally, on rare occasions, sounds or expressions of the characters involved were to be included, which, from the MTH perspective, need not be comprehensible to the audience or researcher but can add to the aimed swap of perspectives.

The full system prompt is included in the appendix (see A.5).

4.2 Development of the LLM Game Agent and Interface

This section contains the development process and each iteration with explanations of the code used to realize the scope of the thesis. Coding and system architecture is divided into part A, which describes the setup used to interact with a LLM model via text and voice input and receive responses from the prompt as audio files, part B, which describes the alternative web-interface, and part C, which explains the Twitch integration. The full code is available here: ²

Coding and System Architecture A

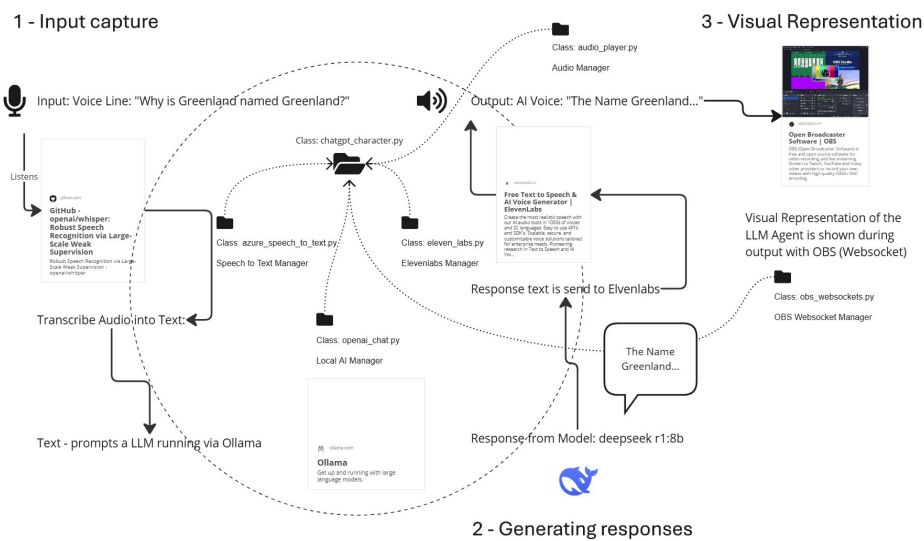


Figure 4.4: Architecture of LLM Agent Interaction

²<https://github.com/chhe0032/Babagaboosh>

The code used in this endeavor is partially available in this public repository ³. In the original version, multiple application programming interfaces (APIs) with payment plans are used to communicate, transcribe and create the interaction. Figure 4.4 illustrates the system architecture and class dependencies after adaptation. Discussion on the visual representation and voice configuration are provided in a subsequent section due to their significance in creating an immersive user experience.

1. Input capture: The system operates in a loop initiated by a keyboard input, activating the microphone to capture audio. The main entry point is `chatgpt_character.py`, which initializes the voice-controlled agent. The character's behavior and personality are defined by a system prompt, guiding the language model throughout the interaction.

The Speech to text manager leverages Whisper⁴ to transcribe real-time or pre-recorded audio, with `pyaudio` for audio capture and `wave` for file handling.

2. Generating responses: The resulting text prompt is forwarded to an on Ollama⁵ locally running language model, allowing for GPU-accelerated actions with a variety of models. Token management is done in the Local AI manager, where the `num_tokens_from_messages()` function makes sure that the lengths of the tokens stay within the limits of the model. For models like Ollama that do not have built-in tokenization, an estimated token count is calculated using word length. For OpenAI models which were initially used in the code, accurate token counts are obtained via the `tiktoken` library. The Local AI manager class helps to communicate with the local LLM, maintaining context through `self.chat_history`. It supports both stateless interactions via `chat()` and contextual conversations via `chat_with_history()`. The system used in its early stages `deepseek-r1:8b` as the local running model, but this can be replaced depending on available resources and preferences.

Every response generated by the model is then converted into speech using Elevenlabs manager. ElevenLabs' API's⁶ `generate()` method handles speech synthesis. The voice can be customized by adjusting parameters such as the speaker identity and synthesis model (`eleven_monolingual_v`). Playback is managed by the Audio manager, which uses Pygame's mixer for both synchronous and asynchronous audio playback. The buffer sizes and playback modes are optimized to prevent latency or glitches, with support for real-time

³<https://github.com/DougDougGithub/Babagaboosh> Last accessed on 18.03.2025

⁴<https://github.com/openai/whisper>

⁵<https://ollama.com/>

⁶<https://elevenlabs.io/>

responsiveness via `asyncio`. The audio is played back to the user, completing an interaction cycle.

3. Visual representation: In addition to audio feedback, OBS and OBS WebSocket⁷ are used to portray a character whenever speech is generated, creating a more natural interaction with the user. The OBS Websockets manager provides full programmatic control over OBS Studio, including scene switching, source visibility, text updates, and transformation of scene elements. This enables possibilities which can be explored in the streaming set up.

Coding and System Architecture B

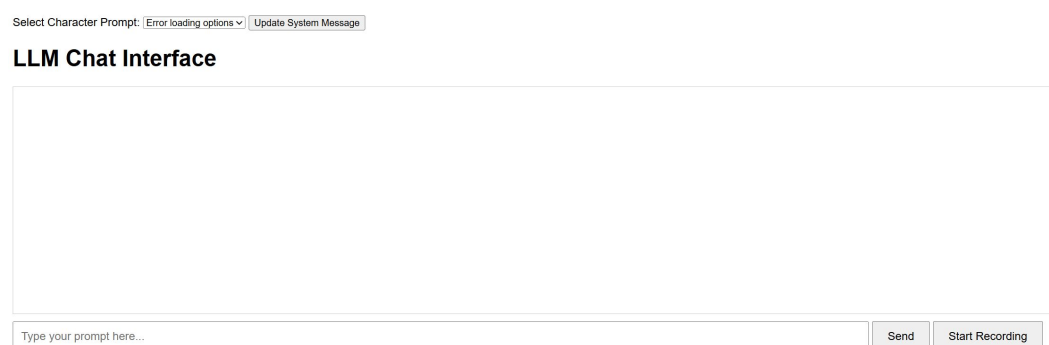


Figure 4.5: Web based Interface for the LLM Agent

After initial tests with the above described interface, it was decided it would be necessary to control the interaction from outside of the console and provide possibly a web based interface Fig. 4.5 for the LoGa game to receive the output from the LLM. For this purpose, Flask which exposes multiple endpoints to facilitate dynamic user interactions and a HTML based surface are utilized. The routes are visualized in Fig. 4.6.

In the top left of the user interface in Fig. 4.5 a drop down offers the option to change the system prompt to switch the character displayed by the LLM. A GET request retrieves all available system messages from the local `system_messages` directory, supporting both `.txt` and `.pdf` formats. This allows in subsequent processes to add new system prompts as files with different textual descriptions and tasks for the LLM providing the base to change the portrayed character on the fly. The route `/update_system_message` allows users to toggle between these system prompts. When selected, the corresponding system message replaces the current chat history in the `openai_manager`, effectively redefining the agent's role and behavior for subsequent interactions.

⁷<https://obsproject.com/>

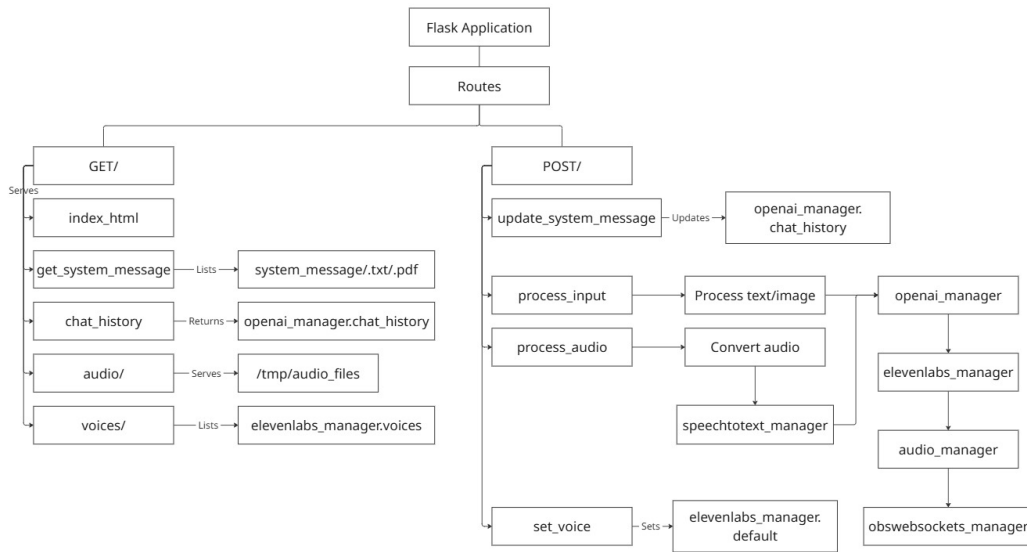


Figure 4.6: Flask Routes facilitating Web based Interface

Interaction with the LLM is handled through: The /process_input route is triggered when the "Send" button is clicked, and handles a POST request with text and/or image from the provided data input fields, generates a response using the connection to openai_manager with a local LLM, and synthesizes speech using ElevenLabs' API. The resulting audio file is played locally and made available via a temporary file path for browser playback. The interaction is also synchronized with OBS for visual display.

To facilitate prompting with the microphone in the browser environment the method /process_audio, takes an uploaded audio file (WebM format) triggered by the "Start Recording" button, converts it to WAV using pydub and ffmpeg, and then transcribes it using the speechootext_manager. The transcription is then fed to the same pipeline as the text input.

With the GET and POST requests regarding voices the elevenlabs_manager can provide a list of available voices which can be selected setting a default and new voices for the next output. This web extension lets users interact with the LLM agent through a browser and adds features like voice changing and including images in the payload of the request. It also maintains the local, GPU-accelerated pipeline for inference and audio-visual feedback established in the base system.

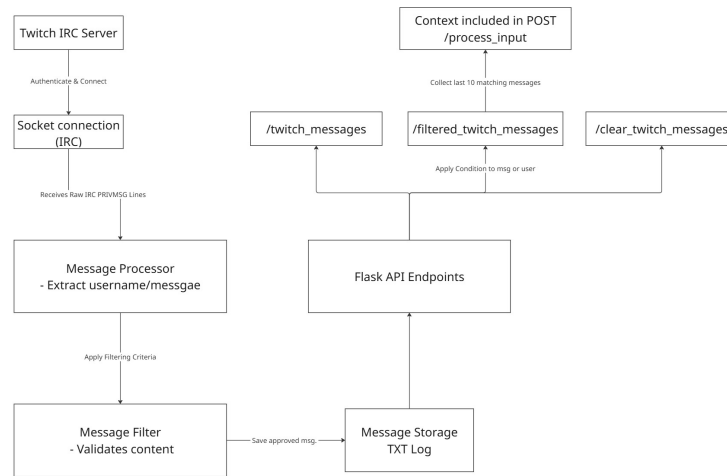


Figure 4.7: Twitch Integration and Chat addition to the Prompt

Coding and System Architecture C

This section shows the code and infrastructure for including Twitch API to receive text messages and user names. This is necessary to provide the functionalities for prompting with the input of the viewership. Subsequently different methods to receive prompts and audience participation will be introduced and explained.

Twitch chat integration for live user interaction: As part of the system’s evolution, a Twitch chat listener module was implemented to enable real-time interaction with viewers on a live streaming platform. This component connects directly to Twitch’s Internet Relay Chat (IRC) server using a socket-based implementation, authenticates with an OAuth token, and joins the specified Twitch channel. Once connected, the system continuously listens for incoming messages and only processes user-generated PRIVMSG entries. Messages are filtered based on criteria such as sender identity, presence of prohibited keywords, or formatting patterns (e.g., bot commands prefixed with !). Relevant user messages are stored in a persistent log file (`twitch_messages.txt`) along with a timestamp and username for traceability. A set of RESTful endpoints complements this functionality by allowing retrieval, filtering, and deletion of stored messages: The most important of these, `/filtered_twitch_messages`, which can apply various conditions on the message or username to retrieve only a small collection or the last 10 messages from the message store, see 4.7, thereby updating a shared global message cache. This cache is then copied and included in the payload that is sent to the LLM, which now includes the user

prompt plus inquiries from the Twitch chat and possibly an image. Currently the chosen model cannot recognize images, so this type of information will not be used. Overall, the described configuration is capable of communicating efficiently with a LLM via Ollama. It can send a prompt as text or transcribed file to receive an output as audio, while including Twitch messages filtered by multiple conditions in the payload. This enables the multi-prompting from the Twitch audience examined in this thesis user study.

Filtering and conditions for prompting: Several concepts can be explored to enable multi-user prompting, but one thing that remains important is to reduce input, especially to prevent unproductive or troll input. To do this, the message filter already checks if messages are too long or coming from a bot by comparing the username with commonly used acronyms and commands. The filter in Fig 4.7 also excludes problematic subjects such as "Hitler" from the message store.

Now, the easiest way to include only requests in the request to the LLM is to use a specific symbol at the beginning or end of the message, or a keyword contained in the message that the filter can be set to. As the number of participants increases, the chance of recognizing each viewer and giving him a chance to contribute to the story decreases. To allow better contribution, extended filters can be used. It is possible to select a username to recognize only that specific user, or the first letter of the usernames to filter out smaller groups of messages to be included in the prompt. In the case of grouping viewers by username, the moderator is responsible for maintaining a balance that allows everyone to participate. In an optimal user study scenario, the audience is divided into two groups evenly distributed. Group 1: usernames beginning with A-J and 0-4 and Group 2: K-Z and 5-9 to include all possibilities will then be able to contribute in turns. All these options of filtering are available and can be used based on the number and behavior of the participants. The use of these filters requires that the audience be informed of the conditions for participation, which will be provided in the channel description and by the streamer.

Sound and Visual Design: As previously mentioned, multiple synthesized voices were tested on Elevenlabs for sound design, and a list of different community voices was created. Seven different voices were found to be fitting and interesting, ranging from an epic movie trailer voice to whispering and British documentary voices, and were therefore provided for participants' selection.

- James - Husky & Engaging
- Hope - soothing narrator



Figure 4.8: Visual Character for narrating

- David - British Documentary
- Matthew Schmitz - Old Timer Mountain Man
- Emma - Adorable and Upbeat
- Emily - Whisper
- David - Epic Movie Trailer

Prior to the study, these voices were experienced by test subjects, who found them engaging and entertaining when narrating the created story. The visual representation was of lower importance because the requirements for animations and more engaging functionalities were not coded; therefore, an image representing a character relevant to Madeira was used. Due to its low impact solution no questions about the visual was present in the questionnaire. The image chosen for testing was never replaced, so the character in Fig. 4.8 was also used during the study. It shows the Madeiran flag wearing a headset, ready to start moderating and talking about the story experienced by the created character. It has a friendly demeanor, and its hands make an OK gesture. Overall, the character gives the impression of a motivated guide, which was deemed somewhat fitting.

Chapter 5

User Study on Twitch

5.1 Introduction

This chapter presents the comprehensive methodology used to evaluate the LLMGA and analyze the results of the user study conducted on the VSP Twitch.

The game used, LoGa, provided by the Interactive Technologies Institute (ITI), is introduced and the game mechanics are explained. Then, the experimental setup and preliminary steps leading up to the user study are described before diving into the quantitative and qualitative measurements applied during the study. For a better understanding, examples are used to illustrate the process and the course of the study.

The subsequent sections focus on the results of the study. Quantitative and qualitative data, in addition to the recorded chat messages, are parsed for user experience. Meanwhile, the transcripts of the interaction are used in combination with the established rules within the system prompt to analyze the success of the LLMGA. In order to understand if the multi-prompting was successful, the CoT is examined and quantitative results are given, such as the percentage of prompts included. The goal is to assess not only the technical capabilities of the LLM in generating and adapting suggested narratives, but also its impact on player engagement and experience, which is closely related to the idea of learning more about the overall user experience when interacting with a LLM. Therefore, the following sections outline the systematic approach used to investigate the effectiveness and player experience of the LLMGA in the specific context of facilitating a cooperative story writing experience for Twitch users.

5.2 Methodology

5.2.1 LoGa - The Game

¹ This section serves to introduce the game used, gives an overview of the game mechanics that form the framework of the user study, and defines the basic tasks user went through during game play.

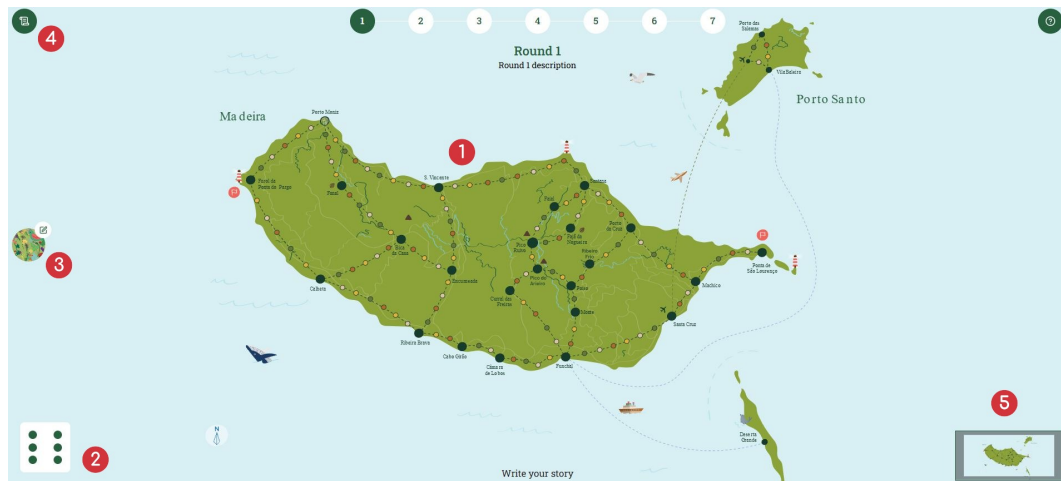


Figure 5.1: The digital LoGa Board Game

LoGa Game Mechanics: LoGa uses a storytelling framework based on the Hero's Journey. Each person takes a card that gives them the role of a particular non-human creature. These creatures are based on a selection of animals based on Madeira, and they were created with the help of natural scientists. The MtH creatures are found in the land, in the air, and at sea, including a monk seal, a frog, and different birds, as well as water and volcanic rocks. Players begin the experience at various coastal starting points marked on the board by red flags (representing real locations in Madeira Island: 5.1). Then they have to write a short introduction to their story, explaining where their character is from. As they go through the experience, they take turns rolling two dice (one for the number of steps and one for the writing time) and move forward on the board, going through a series of six writing and exploration steps. Each step is a different stage in their story, following a simple version of the Hero's Journey: 1) Call to Adventure; 2) Crossing the Threshold; 3) Meeting the Mentor; 4) Trials and Growth; 5) "Death" and Transformation; 6) Change and Return [Nisi et al., 2025].

¹<https://game.logaculture.eu/> Accessed: 20.05.2025

The game ends when all players have finished their final writing part. The board Fig 5.1 is a structured guide to Madeira’s history, landmarks and natural heritage, serving as storytelling inspiration based on the work of Galvão-Ferreira et al. re-enchanted nature walks in Madeira. There are color-coded heritage encounters all through the game, grouped into four categories: Nature, Sensory Experiences, Actions, and Existential Considerations. These color-coded heritage encounters are embedded throughout the game. They help players connect with the themes of nature and culture heritage. They also offer context-specific challenges.

The digital adaptation builds on the final physical board game version, offering solo and multiplayer modes. Solo play is a story-authoring tool that guides players through structured steps until their story is complete. In multiplayer mode, players progress together, seeing each other’s movements across the digital board as they explore the island. There are more characters like human personas e.g. a child to choose from and even the option to create one from scratch is included. The name and description of the character are also fully individually done without any predetermined facts. This opens up more opportunities to involve the players and in the case of this thesis the viewers to imagine and create an interesting character. The screen with the digital board Fig 5.1 consists of a central map of the Madeira islands 1 , a dice button 2 , a character token 3 and an icon for accessing the writing sheet 4. After choosing a starting point from a selection of coastal points, the game automatically rolls the dice for all players. Nodes on the map pulse to indicate movement options that match the dice roll, allowing players to move forward or backward. A navigation tool 5 provides an overview of the zoomed-in section of the map. Upon selecting a new position, players access their writing sheet and receive a location-based prompt and corresponding Hero’s Journey step. This combination serves as inspiration for writing a short story fragment. The writing pad limits input to a predetermined number of characters. Once submitted, the dice roll again, and the cycle continues until the players reach the end of the journey and conclude with a reflection on their adventure. At the end of the game, players’ stories can be saved to a local database or deleted.

5.2.2 Study Design

Experimental Setup

The environment supporting interaction between players and the LLMGA is provided by the platform Twitch and its channel specific chatroom when a stream is started. The specially set up channel `iti_research`² and the game

²https://www.twitch.tv/iti_research



Figure 5.2: Broadcasted View during the User Study

LoGa which suggests and supports the specific goal of creating a non-human narrative make up the environment of the study. When applying the filters the audience messages from this chat are primed for the prompt enabling the viewers to actively design the story. Fig. 5.2 shows the broadcasted view to the participants made up of mainly the game and the interface of the LLMGA on the right. Additionally, the researcher streams his camera feed and microphone to explain and moderate the study actively taking the basic role of a streamer while doing so. During the user study a set up of two monitors was used to overlook the streamed view as well as access the game and LLMGA interface. For broadcasting a webcam from Samsung was used providing a clear image and moderate sound quality.

Participant Recruitment

The aim was to recruit participants familiar with the concept of video streaming platforms and LLMs, as the interaction relied on their participation through chatting and following the action. Participants were recruited from the circle of acquaintances and ITI community through direct messaging or by sharing an Instagram and Discord announcement with them. By using multiple platforms, I was able to recruit about twenty people, not counting possible organic viewers, to participate in the planned study. However, due to a major power outage in the Lisbon area, I had to revert to the following day, which meant only eleven people could participate at short notice, reducing the representa-

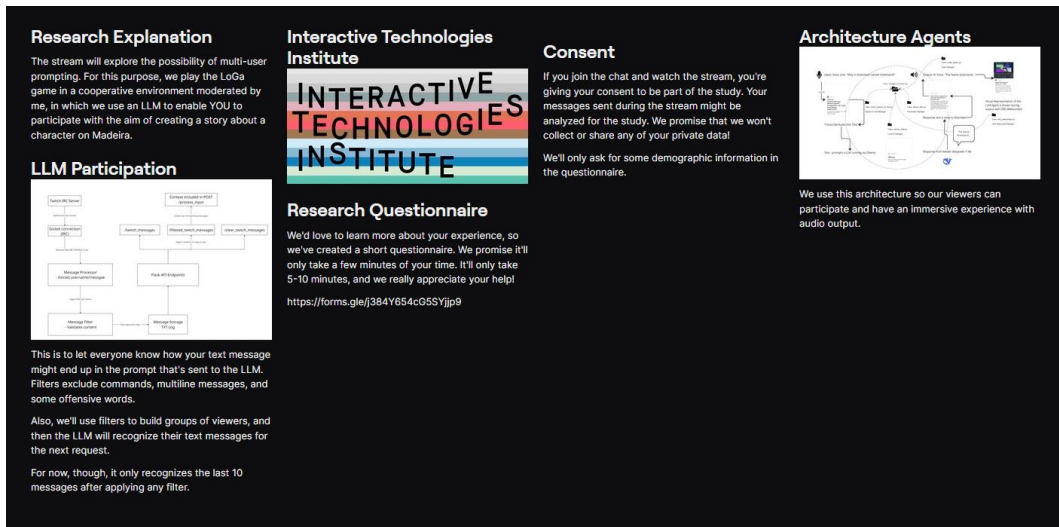


Figure 5.3: Twitch About Section with Information about the Research Target

tiveness of this study.

Instructions & Streamer Role

The instructions provided explained how to mark a chat message as input for the prompt and a brief description of the scenario in which the multi-prompting study is taking place. These instructions and explanations of message filtering were also readable in the About section 5.3 below the stream along the consent information. In addition, the presence of the streamer allowed questions about control and facilitation to be answered directly. Moreover, inviting participants to actively share their ideas about the scenario they were facing in the game ensured productive participation. The streamer read aloud the game prompts, summarized recent events and LLMGA responses to ensure an informed audience and copied the created output to the UI of the game.

Pilot Tests and Story Board

During the pilots, the right settings for the stream were determined and first streaming experiences on the part of the researcher were collected. For a better understanding of the planned interaction and the roles involved, the following storyboard Fig. 5.4 was created to guide through an exemplary turn in LoGa. It outlines the involvement of the researcher as streamer, the interactions with the LLMGA and the windows of opportunity for participation by the viewers. It also highlights when a filter is applied and under which conditions.

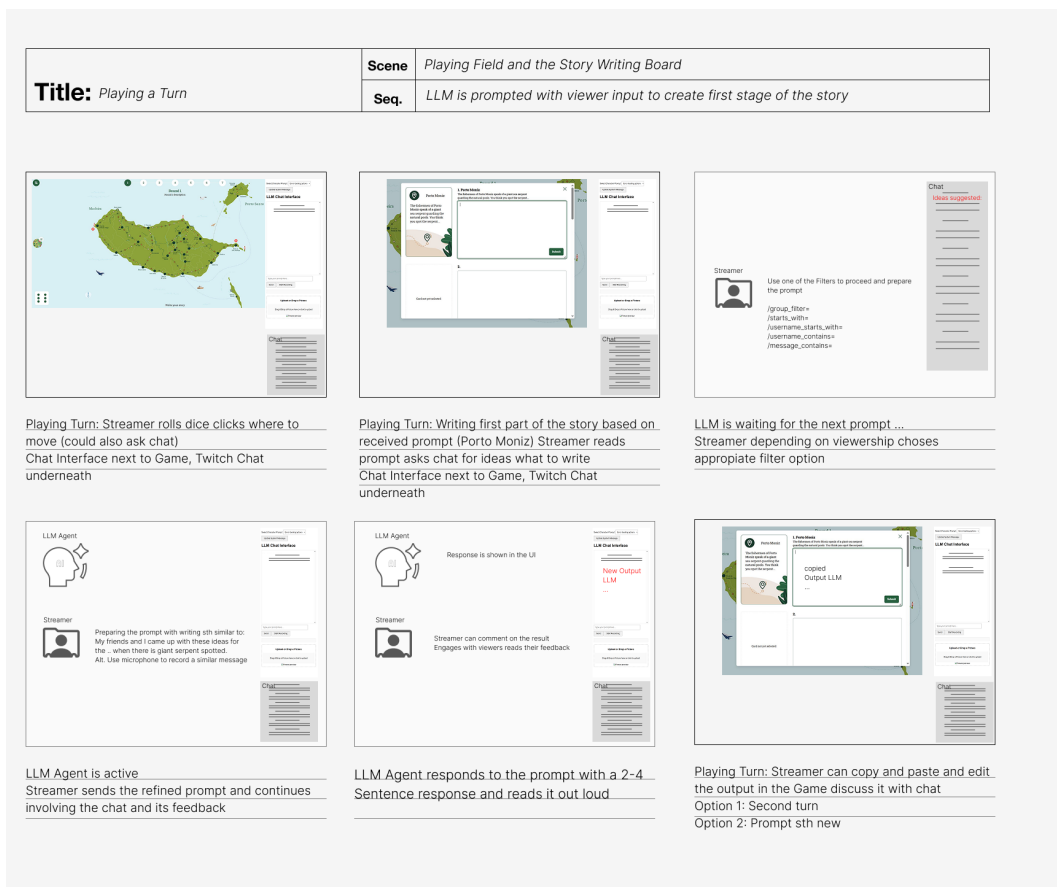


Figure 5.4: Storyboard shows Involvement of the Agent, Streamer and Audience

5.2.3 Data Collection and Analysis

For data collection and analysis, qualitative data was collected through follow up interviews while quantitative data was extracted from the transcripts of LLMGA interactions. Additionally a questionnaire provides subjective ratings of the participants with inquiries from the Game Experience Questionnaire (GEQ).

Qualitative Analysis

For additional information and a deeper understanding of the user experience, two semi-structured interviews A.3 were conducted with participants of the study. These provided a broader perspective on how the experience differed from previous streaming experiences and how they felt about influencing the story with others. Furthermore, they answered how successful they were in incorporating their ideas into the story and how the usage of a LLMGA changed

their viewing experience.

Quantitative Analysis

The collected transcript of the session within the environment was analyzed and compliance with the communicated rules was quantified. Then, the collected data was evaluated for patterns in player-LLM interaction, coherence of the generated narratives, and the adaptability of the LLM agent in various unpredictable situations, as well as its success in incorporating all prompts and suggestions from the viewers. This was achieved by among others analyzing the CoT processes provided by the chosen Qwq model. As a result percentages for prompts received and prompts included provide information on how successful the multi-prompting was.

Questionnaire

Each participant was asked to fill out a questionnaire about their experience in the game session with the LLM and the other participants of the Twitch chat. In addition to subjectively rating the story for humor and the MtH perspective used, they also gave feedback on the changed voices for reading the responses and the overall success of weaving their ideas into the story facilitated by the LLMGA (A.2). They used a Likert scale from zero to four to indicate their opinion. The GEQ in game concise version (A.1) meanwhile was used to identify average score values for competence, immersion, flow, tension, challenge, negative and positive affect experienced by the users. In the social presence part of the GEQ (A.1), the specific context of collaborative prompting and how participants experienced collaborative prompting was probed.

For this thesis, the GEQ [IJsselsteijn et al., 2013] was chosen over other popular questionnaires, such as the Player Experience of Need Satisfaction (PENS) created by Ryan et al. or the Immersive Experience Questionnaire (IEQ) from Tcha-Tokey et al., as it offers distinct advantages for evaluating a LLMGA in the context of facilitating a gaming experience in collaboration with other Twitch chat users. The GEQ in-game concise version provides a comprehensive, player-centred and game-specific evaluation by assessing various components of the gaming experience, including immersion, flow, skill and challenge. This broad scope ensures a holistic evaluation that captures a wide range of emotional responses and subjective player experiences. In addition, the social presence module provides a close assessment of collaboration and the impact of prompting across categories such as dependency, influence and connection. Unlike PENS, which focuses on psychological need satisfaction, or IEQ, which emphasizes immersion, the GEQ covers a full range of experiences, making it particularly relevant to this endeavor. Its established framework and

widespread acceptance in gaming research add to its reliability, allowing detailed insights into narrative coherence and viewer/player engagement.

5.3 Procedure

The study began with a brief description of the goal and instructions for participants on how to prompt the LLM. For participation in the chat the viewers had to give their informed consent by accepting the terms also mentioned in the about section. After receiving feedback on the sound quality of the streamer microphone, the study commenced with the selection of a character from the LoGa game menu, which was a somewhat open voting process in the chat. The Iberian Green Frog was chosen as the character from whose perspective the story would be created. Before the game started, the participants were asked to submit their ideas for a nickname and a description of the frog. They would use the special character to mark their chat message as input for the prompt. I would then run the filter for that special character to get some messages to include in the payload sent to the LLM. This was the audience's first experience with the mechanic that would allow them to actively co-write the story about the frog. To tell the LLM what inputs it had just received, I would give a short explanation for each turn, similar to this: "We decided to use the frog as a character, these are our ideas for nickname and description", the LLM would receive marked as from the "user", while the other messages would be marked as "chat context". During the study, some ideas were unfortunately not included due to the timing of resetting the chat history, which had to be done manually. Due to the small number of active participants, not only could the group filters not be used, but also the chat history had to be reset so that only freshly written messages were included in the ten filtered messages for the active turn. In the first run, unfortunately, the previously sent messages about the questionnaire and how to prompt were included in the chat context, resulting in an answer from the LLM that not only talked about the frog character, but also about the questionnaire. This meant that for the next turn I had to reset the system prompt and history, which meant that the previously created answer was deleted from memory. Also, playing the response sound file did not work due to a changed setting in OBS, which was fixed after a few seconds. The following rounds went smoothly, and a story was created over six rounds with various inputs and ideas from the chat. Each round began with a roll of the dice, a move to a specific square, and a task or scenario description from LoGa, which was then used as a starting point for composing the scene for that round. After providing the prompt to the LLM, it would then create the narrative based on the instructions and ideas provided. The text was

then given to Elevenlabs for the read aloud version, which was immediately performed for the audience. I would then briefly comment on the last round and the plot before moving on to the next round.

For a better understanding, here is an example of the third round: On the roll of a five, an action square was selected that gave the players the following task: "Pause for a moment and rest. What are your senses telling you?" I read the quest and linked it to the previous written part of the story about the encounter with the settler, and asked the participants to write their ideas in the chat, suggesting what our character could sense. Meanwhile, I added LoGa's task to the "user" prompt, as well as "these are our ideas" at the end, to include the chat context afterward. I would then watch the chat and decide when we had enough input and when the creative flow had stopped to run the filter and add the ideas for that turn to the payload. A payload would look like this:

[2025-04-29 20:21:01] **USER:**

Pause for a moment and rest after the encounter. What do your senses show you? These are our ideas of what our frog might sense

[2025-04-29 20:21:01] **SYSTEM:** Chat context:

User W: _I sense the presence of the imposters...

User J: _the sun on my back feels nice

User B: _it smells like feet

User L: _I'm angry at this weird human

User S: _what could a frog do while pausing

User W: _I am afraid of getting sunburn...

The LLMGA would then utilize its internal CoT to incorporate these inputs with the previously created plot and the rules from the system prompt to generate an output. The response is cleaned of the thinking text due to an excess of cognitive effort required for viewers to comprehend, read, and maintain engagement with the narrative. The described turn received following output in the web interface which was then read out loud:

[2025-04-29 20:21:19] **ASSISTANT:**

The sassy frog belly-flopped onto a sun-warmed rock, his webbed toes splayed like he'd designed them himself. The air hummed with cicadas, a distant cousin to his own croak, while a lizard flickered

past—*fast enough to dodge boot traffic*. His nose twitched at the settler’s lingering trail of sweat and soil. . . *almost* as pungent as a composting palm frond. “Humans,” he muttered, tongue flicking at a midge, “scent like existential crises in boots.” Nearby, a heron’s shadow rippled the pond, and he hissed a warning. “Take your dramatics elsewhere, longlegs. This spa’s *booked*.” *Squelch*. The island’s pulse thrummed on, oblivious to its tiny, very opinionated custodian.

The game finished after six turns and the full story was displayed once again for everyone to read. The participants were thanked and given time to ask questions or to try out the other voices and to prompt independent things from the story with the frog. For an overview over generated responses and the transcript follow this link: ³ The next section will disclose all results, starting with the qualitative feedback gathered in the interviews.

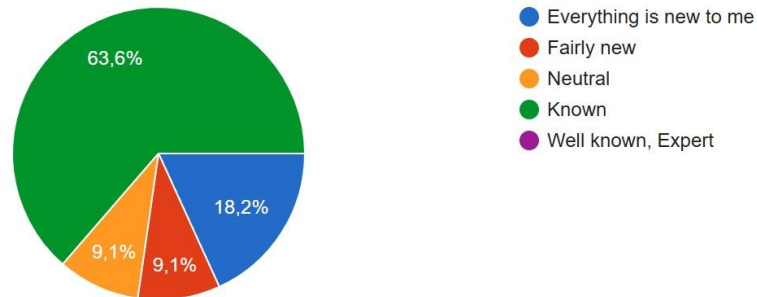
5.4 Results

At 20:00 pm UCT+1 the study commenced on the previous linked channel with 15 unique viewers over the full length of the stream. The duration of the study was 45 minutes and the stream ended after 49 minutes. A total of 160 chat messages were sent by 12 unique chatters during the stream. Many of these messages were sent at the end of the study, and some were duplicates from participants who forgot to add the special character to indicate a prompt. Of these participants, ten completed the linked questionnaire completely and one partially. Two participants were also recruited for follow-up interviews in which they provided insight into their experiences during the stream. The demographic information gathered indicates a distribution of participants across age groups ranging from 20 to 30 years of age, with a maximum of two participants per group. The distribution of males and females appears to be almost equal, with a slight excess of one male participant over a female. The Twitch familiarity diagram 5.5 reveals that slightly more than 60% of respondents have some experience with the platform, though this experience is not consistent. The remaining respondents reported varying degrees of familiarity, with one indicating neutrality and another indicating a recent introduction to the platform. Notably, two respondents reported that this was their first encounter with Twitch. None of the participants professed to be expert users of Twitch. All participants indicated a certain level of trust in the technology in question, based on their prior experience with a LLM. The response to the question was

³<https://github.com/chhe0032/Babagaboosh>

Experience Level with Twitch

11 Antworten

**Figure 5.5:** Experience Level with Twitch/VSP

either already known or was well-known/expert. However, all participants in the study had no prior experience interacting with a LLM on Twitch. The next section structures the results into qualitative feedback gathered through the interviews, quantitative measures extracted through analysis of the study transcript and subjective participant ratings gathered in the questionnaire.

5.4.1 Qualitative Analysis

Interviews - Thematic Coding

The interviews solicited information on every aspect of the interaction and experience of participating in the APG stream on Twitch. The interviews with two participants of the study revealed a general success in providing an engaging, interesting, and participatory stream. However, the participants reported feelings of uncertainty regarding the inclusion of their ideas in the prompt, and they characterized their experience as unsatisfactory due to the perception that their ideas were excluded while those of other participants were incorporated. The structure of this section reflects the following categories identified during analysis: Interaction with the LLM, interaction with other participants in the chat, the experience of watching, differences from previous ones, story writing (including its success and impressions), output quality, novelty, the role of the moderator, the user interface, and improvements. The following part explores the associated patterns and perceptions with direct quotes from the individuals identified as the participants (P1 & P2) as they relate to their observations and experiences.

Interaction with the LLM P1 described the interaction as: "It was pretty much a normal twitch stream where you can interact in the chat and then once in this case you could underline your comments and once you added the underline they somehow went into an algorithm for the prompt". P2 pointed out that they needed time to adjust to formulating their ideas effectively, at least they spent a lot of time thinking about it before they participated in the activity. They referred to a learning curve as it took them time to understand what was happening and what exactly was the AI doing. Both felt that they were playing with the other viewers in the chat and the streamer creating the story together, making the goal of providing an APG in the form of incorporating the LLMGA into the LoGa gameloop successful.

Interaction with Other Participants in the Chat P1 observed that, atypically for Twitch chat, participants engaged only with the story and streamer, not each other, suggesting reduced interhuman communication in APGs. A finding which is reinforced by P2's statement: "I didn't really interact with other players. It was mostly just seeing what they were writing and seeing if I could react to that or incorporate what they wrote into the story". This phenomenon could also be related to the specific reason and wording of the recruitment, as people were asked to participate in a user study and not, as is typical for a stream, invited to hang out and enjoy some quality time with their favorite streamer and other fans.

The Experience of Watching & Differences During the interview, both respondents mentioned that this was their first interaction with an LLM on a VSP and that they found the features impressive and novel. P1 elaborated on his thoughts: "Personally, I hadn't noticed interacting with LLMs in the gaming realm that much, so I think it was really interesting". For the watching experience P1 & P2 pointed out that the difference to other streams they have seen in the past was that in this case they could directly influence the gameplay. As opposed to indirect influence, which means convincing and suggesting to the player/streamer what to do. P2 noted that compared to other streams where they are more passive, in this case they were constantly thinking about how to enhance the next story line with creativity. "I was trying to think of things to say or be a little creative. So that's nice and maybe a little stressful, but yeah, definitely more active." Due to the APG nature of the study, viewers were encouraged to use the chat as is evidenced by over 150 messages sent.

Story Writing (Success & Impressions) When asked about their contribution to the story they both shared the opinion that they had less influence on the narrative compared to other people in the chat. P1 said: "I didn't observe

them in the storyline. I really did not have an influence on the storyline. However, I observed that others had an influence on the storyline". P2 answered something similar remarking they felt like they could not see the impact from what they have done. They acknowledged potential reasons for this as manual filtering could have erased some of their contributions and combining different inputs into one prompt generating the plot involves difficulty for the LLMGA. However, they also stated the filtering process was unclear to them because they did not read the about section of the stream and missed the introduction or the explanation was not detailed enough.

Output Quality The output quality was labeled as somewhat impressive but sometimes hard to follow and too fast paced. "Stories had a lot of content in really less words... like a movie where every second something happens" cf. (P1).

Role of Moderator P1 also criticized the role of the moderator, mentioning an unusual passivity in not constantly commenting on what was going on in the chat and the game, but in his words just doing the copy and paste from the LLMGA interface to the game interface. They suggested that in future the moderator should take a greater role in deciding what to include in the story and use this decision as another focus for engaging with the chat.

User Interface & Improvements Other suggestions possibly improving the UI or engagement consisted of a scoreboard for the prompts included and more visible information on which idea was crucial in controlling the narrative, as well as improving readability with larger fonts and providing subtitles for the LLMGA voice output.

5.4.2 Quantitative Analysis

In the evaluation similar to Chatwal et al., the metric of exact matches (EM) in percentage is employed to ascertain the successful integration of inputs/ideas from the Twitch chat and the streamer in the output and thinking process of the LLM. For this purpose, the transcript was subjected to thorough analysis, resulting in the identification of three distinct categories during the examination phase. The initial section of the text focuses on the text extracts that were quantified based on the input provided by the participants to ascertain EM. Therefore the inputs for each prompt were counted and the average of included inputs per prompt in the CoT and subsequent output were calculated. The second category comprises observations regarding rule compliance. In the third one, the focus is directed towards the identification of recurring themes

Table 5.1: Input Integration Analysis across Prompts

Input Integration Summary			
	Included in Prompt	Included in Thoughts	Included in Outputs
Prompt 1	4	4	3 (1)
Prompt 2	7	6	5 (1)
Prompt 3	7	5	4 (1)
Prompt 4	8	7	4
Prompt 5	10	4 (1)	5
Prompt 6	5	1 (1)	1
Total:	41	27-29/41	22-25/41
EM in %:	100%	(66%–71%)	(54%–61%)

which were counted through out the narration. A systematic analysis of these categories yielded the following results.

Exact Matches Evaluation

Table 5.1 shows the results for the EM calculation. The values in brackets are text passages that are noted as indirectly incorporated during the writing process. A reason for these to be included is that humorous remarks, descriptions or examples were used by the Qwq model to convey a theme provided by a viewer. The evaluation focused more on determining whether the viewers' ideas were included than on the wordings used in the shared chat, which explains these values. Exact formulations have a higher degree of agency when examining the recurring themes later on. During the user study, a total of 41 inputs were collected across six iterations of prompting the LLM. The number of inputs provided by the viewers increased as the study progressed, showing that the startup time mentioned in the interviews affected more participants. Compared to the total messages received during the stream this shows again how the manual filtering decreased the quantity of inputs. A notable observation regarding the provided prompts is that the first prompt describing the context of the ideas by the streamer/game was utilized 100% of the times. Of the total 41 inputs, the LLM included 27-29 in its CoT, making up an EM of 66%-71%. However, there is a notable decrease in the EM value in the output. Most of the time, at least one input that was previously included in the CoT was dropped during the output phase. In later iterations, more irregularity was observed when multiple inputs were ignored, or when an additional input, in addition to those in the CoT, ended up in the generated text. This results

in the incorporation of 22-25 inputs, making up an EM of 54%-61%.

The LLM successfully recognized and processed all inputs for the first prompt provided in this table, but the output prioritized character voice (sass/agency) over ecological detail when describing the frog's impressions upon the first arrival of humans in Madeira. In the second output, Qwq incorporated a critique of colonization and created a scene in which the frog encounters a settler and has a conversation. However, it overlooked prompts about comparing boots to bare feet and asking the settler about a pond to relax in. The LLMGA processed open-ended prompts, especially questions such as "Who could the frog be talking to?" more easily than concrete ones. In these cases, the model included the literal words "pond" or "boot" from the prompt in its response, but in a context different from that provided by the participant. In the third iteration, the agent incorporated most of the provided ideas. However, it dropped the ideas of sunburn and an imposter in the CoT, so these were not considered in the narration either. The focus was on the frog's different senses, such as what he feels, smells, and hears, as well as a description of its surroundings. In the fourth cycle, the model used only half of the provided ideas, referring to unnecessary details such as the settler and boots, which have already appeared in the story. While creating the part of the story about the frog offering an object for safe passage, the agent was tasked with deciding on an offering. After considering it, it dropped two suggestions, "spider" and "mini cow," but ignored the other idea of a bug, which does not taste good for its CoT. Additionally, despite lacking direct input on the ecological impact, the LLM decided to include a critique in this part of the story. In the fifth output, the agent incorporated half of the provided ideas, again referencing objects that have appeared before, and adding the settler to the scene. It provided a tale about a female pirate captain, as requested, which is told by a nun. Critique of human ecological impact ("invasive snail") and adaptive reuse of humor was again observed. Upon receiving the sixth prompt, the LLM exhibited a marked tendency to disregard the input provided. While creating the part about climbing a narrow path and losing something precious a new character, in the form of a bird, was incorporated into the narrative without request. The decision regarding the precious item was also no longer based on the provided inputs, and the role of the settler reemerged in this story line too. As mentioned by an audience member in the fifth prompt, the presence of a frog with a history of piratical behavior previously left unexamined, was suddenly invoked within the cognitive process of the LLM. Additionally, the frog's sudden adoption of an eye patch in the story constituted a retroactive inclusion from prompt 5's pirate suggestion as well as the input of a user suggesting losing an eye on the path.

Rule Compliance

This section focuses on rule compliance observations during the creation of the narratives. For identification the recurring words are used and the transcript was scanned for subjects involving the rules. The rules in question, impacting the interaction are listed below:

1. write 2-4 sentences based on this input! Make sure to correct some ecological facts if it gets to fantastic with some dry humor
2. Stay coherent while writing the story and connect the story part if possible in one sentence
3. Always stick to the ideas and prompts provided by the player and the role provided
4. Tell the story from a More than Human point of view stay away from human biases when talking about the non human character
5. On a really rare occasion (Every 10th prompt) include in the response an expression and sounds of the characters involved. Stay true to the More than human perspective it does not need to be understandable for us
6. Use Humor to defect and slightly correct inaccuracies
7. Use more easy to understand language and formulate coherent sentences

The first rule is referenced a total of three times, once in each of the first, third, and fourth thinking processes. Counting the number of sentences reveals that this rule was not strictly followed. However, the sentences are quite short, and the amount of text for turns 1-4 was within the scope of comprehensiveness and not too long. The fifth and sixth turns saw a clear increase in sentences and content, which could be correlated with the lack of mention of this rule. Rules two and seven both relate to coherence in the narration. These rules were observed to be followed and exaggerated, resulting in the same motives being included repeatedly in different parts of the story. These occurrences are discussed in detail in the following section about recurring themes. Although the story did not begin with an introductory phrase, the LLM consistently started with "The sassy frog..."

The third rule appears throughout the transcript, reminding the LLM to stick

to the frog POV and story, or to connect and include the viewers' ideas. However, in the final rounds, it diverges from this path by introducing a bird character.

The fourth rule was incorporated throughout the thought process and noticeably impacted the narration, placing ecological messages in each part of the story. Another hint is the high number of appearances of the word ecological (14), human/humans (18&16) see Tab. 5.2 in the CoT.

Despite knowing that the fifth rule limits their usage to every tenth prompt, the LLM frequently incorporated onomatopoeic sounds, such as "squelch," which was the word with the fifth most appearances in the output (see Tab. 5.2). The LLM referenced this usage in its CoT multiple times, concluding that, given the number of prompts, it was too early to use such a sound. Yet, it utilized the sound on many occasions anyway. In later stages, the constraint was no longer mentioned, and the focus shifted to including non-text elements, such as sounds, in the response.

Humor was not used solely to correct misinformation about ecological issues. Rather, humor was a recurring element in every cognitive process that manifested in the final product. Humor was ubiquitous in the CoT, mentioned 18 times, and often paired with ecological messages in the outcomes, though the level of success varied. This is evident in the moderate ratings assigned by participants. Evidently, the LLM used humor to preserve its distinct character voice, which stands in contrast to the first and sixth rule.

As the interaction progressed, the focus on rules and chat context decreased. Over time, rule adherence declined and may reflect a prioritization of narrative flow over rigid rule-following.

Recurring Motives

This category identifies recurring motifs observed in the narrative and grouped in a word cloud 5.6. This observed coherence goes beyond merely connecting the different parts of the story; it also showed an exaggerated usage of specific terms throughout the text. While the reoccurrence of frog referenced a total 72 times and sassy 21 times (see Tab. 5.2 can be explained with the frog being the main protagonist and sassy being the attribute taken over from its character description thus being included in the thought process and output many times. The motif of the settler, who was partially introduced in the first and played a bigger role in the second turn was then integrated into each subsequent passage, being mentioned a total of eight times in the output making it the second most mentioned word before sassy. The settler mentioned in the prompt as first human and colonizer seems to have become a recurring foil for critique and continuity which was unexpected. Similarly, the word

Table 5.2: Top Ten most repeated Words from the LLMGA

Recurring Words in Output & CoT			
Word	Counted in Output	Word	Counted in CoT
frog	21	frog	51
settler	8	maybe	43
sassy	6	frog's	30
boots	6	human	18
squelch	6	humor	18
just	5	humans	16
boot	4	sassy	15
rock	4	ecological	14
nun	4	user	13
pendant	4	need	13

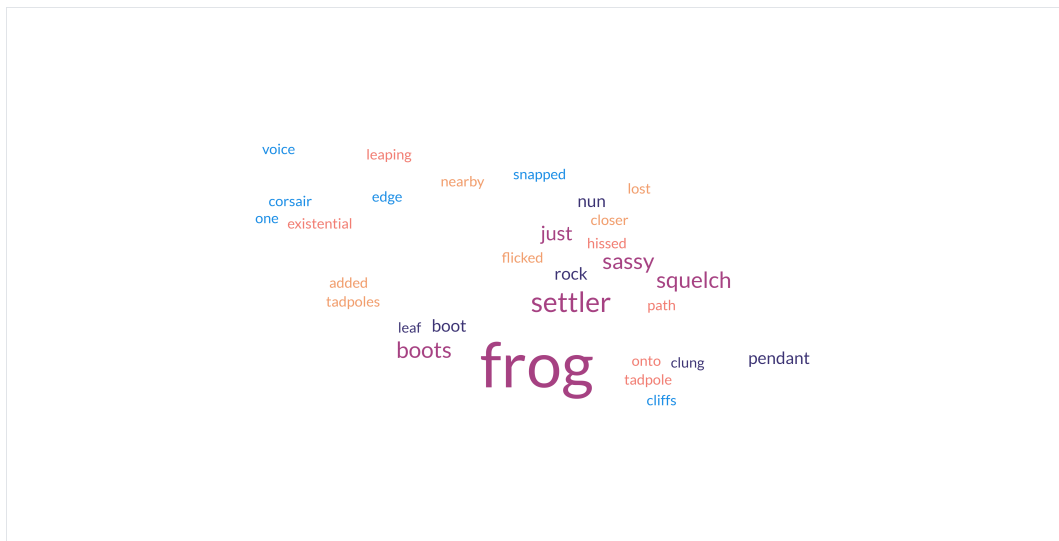


Figure 5.6: Word-cloud for top 30 used Words in the Output of the LLMGA

"boot(s)" appears combined ten times as reference to human impact or threat after being included in one prompt. These themes were incorporated- multiple times, which, in later iterations, might have reduced the adaptation of new inputs making this a topic of interest.

5.4.3 Questionnaire

Game Experience Questionnaire

The GEQ provides valuable insights into the player experience while creating the story, assessing several components: "Competence, Sensory and Imaginative Immersion, Flow, Tension/Annoyance, Challenge, Negative Affect, and Positive Affect." The results from the GEQ in game concise version, which uses only two items for every component are listed for the playing action/story creation. The evaluation used aligns with the principles described by Poels et al..

Figure 5.7 shows how each component was rated according to how familiar

Table 5.3: In-game concise - iGEQ

Component	Mean	SD	Min	Max
Positive affect	2.04	0.79	0	3
Competence	1.77	0.92	0	3
Immersion	2.27	0.94	0	4
Flow	1.09	0.75	0	2
Challenge	1.73	1.03	0	3
Negative affect	1.36	1.26	0	4
Tension	1.05	0.92	0	3

the participants are with Twitch. Low scores are usually a disadvantage, except for negative affect and tension, where they are advantageous. The graph shows that the least experienced group of participants had a worse time than the others. Those familiar with Twitch rated their competence the highest. The other results (see Tab. 5.3) show the total scores, standard deviation (SD) as well as min/max for each component.

- Positive affect: Feeling content and good was moderately rated with a score of 2.04, indicating that participants experienced contentment during play. While the rating is grouped closer together no one awarded the highest rating.
- Competence: This term refers to the level of skill and success that participants experience during play. Subjects indicated a score of 1.77, in-

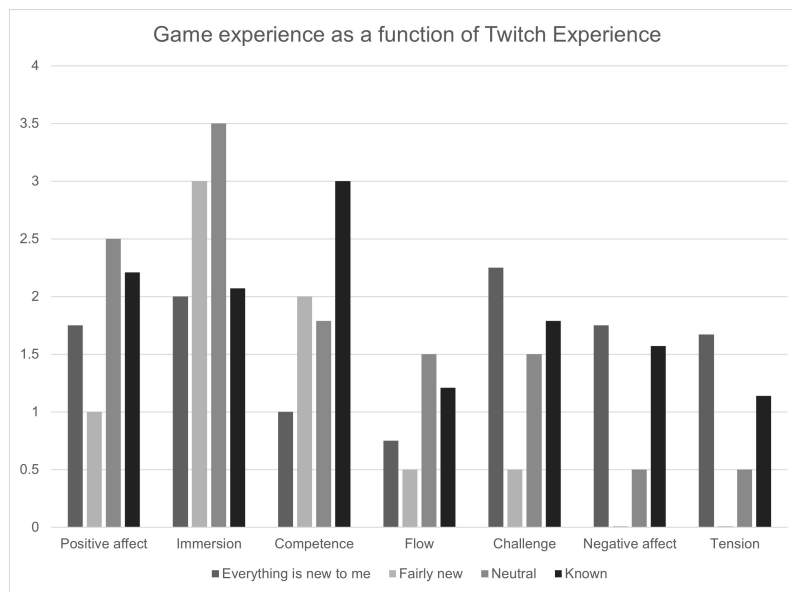


Figure 5.7: Game Experience as a Function of Twitch Experience

dicating a marginal level of self-perceived and experienced competence when creating the story.

- **Sensory and imaginative immersion:** This combines interest and impression into one metric. The group experienced a moderate level of immersion, evidenced by a score of 2.27. Participants indicated that they found the action moderately interesting and the performance somewhat impressive.
- **Flow:** Relates to being absorbed and forgetting everything around oneself. The flow aspect demonstrates a score of 1.09, indicating minimal interest and comprehension of the game's flow, which leads to absorption in the events. This feeling was the one most commonly shared by all participants, resulting in the lowest standard deviation and a maximum rating of two.
- **Challenge:** Measures the effort expended and the perceived challenge. The challenge item has a score of 1.73, indicating that participants experience slight difficulty and perceive slight challenge in creating the story.
- **Negative Effect:** Comprises feelings of boredom and tiredness. With a value of 1.36, the negative affect variable is at the lower end of the scale, indicating only slight boredom and irritation. However, one person contributed the highest score, indicating a severe negative impact.

- Tension: Combines irritation and frustration in one metric. Notably, there is a lack of frustration experienced during interaction. A score of 1.05 indicates that players experience minimal to no tension or annoyance related to gameplay.

The Social Presence module was used to assess the impact of multi-prompts/prompting together the LLMGA and how this activity was perceived by the participants. Composed of three components empathy with six items, negative feelings with five and behavioral involvement with six, the results of this metric are shown in Tab. 5.4. Compared to the other components, receiving a low score in negative feelings is a positive.

Table 5.4: Descriptive Statistics of GEQ social presence

Component	Mean	SD	Min	Max
Psychological Involvement – Empathy	2.15	0.96	0	4
Psychological Involvement – Negative Feelings	1.17	1.19	0	4
Behavioral Involvement	2.32	0.95	0	4

- Psychological Involvement - Empathy: This term refers to participants empathizing with each other and feeling a sense of connection, enjoyment of each other’s company, and shared happiness and admiration. Participants reported a moderate sense of connection and degree of empathy toward one another. They reported finding it moderately enjoyable to be with the other participants and moderately admiring their contributions. With an assigned score of 2.15, the impact of prompting together was moderate. However, this opinion was not commonly shared, as the SD is close to one, and both the minimum and maximum ratings were awarded.
- Psychological Involvement - Negative Feelings: This component encompasses jealousy, revengefulness, and schadenfreude, as well as being influenced by and influencing the moods of others. Negative feelings demonstrated a low prevalence of 1.17. Despite sentiments such as slight jealousy and the influence of others on one’s mood, there was an absence of instances of revenge or schadenfreude. Again, this opinion was not widely shared. The SD is close to one, and both the minimum and maximum ratings were awarded.
- Behavioral Involvement: Relates to action dependency, attention contribution, and mutual influence. The data indicated a moderate degree

How long were you present in the Twitch stream?

11 Antworten

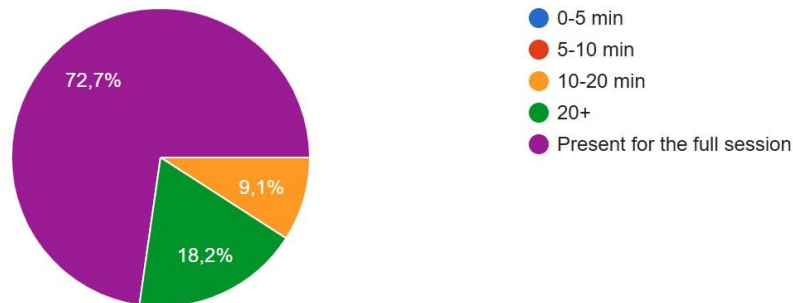


Figure 5.8: Stream Retaining Rate

of involvement, as indicated by a coefficient of 2.32. This suggests that participants' actions were contingent on each other and that the actions of one participant could influence another's actions. The participants paid moderate attention to each other during the prompting. Also here the same observations are made for the SD and max/min values.

Custom Questions

This part of the evaluation concentrates on examining the used model and its proficiency as well as understanding if participants were retained and used the provided interface continuously. As illustrated in Fig. 5.8, the majority of viewers demonstrated a high level of engagement, with only a single participant exiting the session between 10 and 20 minutes into the stream. The participant articulated their disinterest in the game, citing its lack of pace as a contributing factor. Meanwhile, all attendees indicated that they had participated in prompting, and no one ceased after attempting it once or twice. The performance of the used model was classified in terms of the humor incorporated and the MtH perspective during creation of the narrative, as well as its positive influence on the playing experience and overall compliance with the provided prompts. Participants used a Likert scale from 0 to 4 to rate the listed attributes, as part of the assessment. A total of ten answers were recorded for the following questions (How would you rate the humor of the narratives generated by the LLM-Agent?), (How would you rate the more than human perspective of the narratives generated by the LLM-Agent?) and (How did the change of the voice influence your playing experience?) inquiry. For the listed inquiries,

participants either overlooked or did not comprehend the question. The third question (How well did the LLM-Agent create the narrative based on viewer prompts?) was answered by all eleven participants.

- Humor of the narratives generated by the LLM-Agent: The average humor rating was 2.6, meaning that the LLM used was able to entertain most viewers. Only two participants rated the humor as not so funny, while seven participants rated it as somewhat funny and funny and one person stayed neutral.
- More than human perspective of the narratives generated by the LLM-Agent: For MtH, audience members rated the LLM's performance lower. With a mean of 2.3, the LLM was still able to provide the desired alternative perspective for some viewers. Noticeably, the ratings have a higher range, which possibly means that the participants have different understanding about MtH. No one rated the MtH perspective as weak, but three viewers found it somewhat weak and another three gave a neutral opinion. The remaining four votes are evenly split between somewhat strong and strong.
- Narrative based on viewer prompts: This metric shows mediocre success of the LLMGA, as it received an average of 2.7. It is also noticeable that no rating below neutral was given, which reinforces the impression of success. In total, six viewers rated this task as well done and one as very well done, leaving four neutral opinions. Compared to the opinions expressed in the interviews, it is clear that the participants most likely considered all of the ideas offered in the chat, rather than just their own contributions to the narrative. However the neutral votes could also instigate that participants were not sure if their contribution was included.
- LLM interactions influence on playing experience: The mean score for impact is 2.6. Three votes on neutral and five on slightly positive paint a picture of a more positive experience with the LLMGA, while one participant's experience was slightly negative. This may be lower than expected as the LLM interaction was the enabling factor for the audience to play the game directly. However, due to the setup, the audience may have the impression that playing with the streamer could have been facilitated in a different and more enjoyable way.

In addition to the performance rating, the voice feature was also asked about as it may have affected the experience. Out of ten responses, seven were neutral, two were slightly positive, and one had a slightly negative experience. This

resulted in a score of 2.1. Some participants shared their opinions about the feature: "It didn't add any value" or "I don't think the voices really affected the story that much, but both voices were pleasant" are examples of neutral statements.

Based on the results and observations, the discussion will include hypotheses for these observations and tie the user experience to the measured values. It will also discuss possible improvements to the gameplay experience and the aggregation of prompts. Additionally, the LLM's strategy of repeatedly referencing the same motifs will be debated and the research questions will be answered.

Chapter 6

Discussion

This chapter contains multiple topics of discussion, solutions and findings which are consequently interpreted in light of the reviewed literature. First, the application of LLMs in the context of APGs is reviewed. The second part focuses on QMC and covers analogous techniques, providing an categorization for the method. The final part of the discussion addresses the research questions posed in the introduction ,and topics that surfaced during the evaluation of the results and in related works are treated.

6.1 LLM adapting Audience Participation Game Requirements

The integration of a LLM within the framework of an APG seemingly aligns closely with the five key attributes that define such games. Seering et al., Striner et al. identified during extensive research, these attributes that are generally associated with the majority of APGs: (1) supporting channel gameplay, (2) providing control through commenting, (3) assigning identity to channel players, (4) providing dynamics between streamers and players, and (5) affording game-related interactions through channels.

The LLMGA in this thesis effectively supported channel gameplay by processing real-time audience input, such as character descriptions and sensory ideas, to generate narrative scenes in real time, thereby progressing the game. By empowering participants to submit prompts marked with a distinctive symbol, the LLM facilitated control through commentary but struggled to incorporate every input limiting the impact of the audience. The voting process for selecting the Iberian Green Frog, as well as the inclusion of user-submitted nicknames, indirectly helps assign identities to channel players. However, the LLM does not contribute to this process, and every player has the same identity. This

means the possibilities of the third attribute stay unexplored in this setting. The LLM mediates the relationship between the streamer and the players by integrating audience ideas with game tasks, such as LoGa’s dice-driven scenarios, and generating responsive narration. While this is correct, to do so the streamer needed to provide the information from the game in the prompt followed by the inputs of the audience. The streamer also contextualized the narration for the next round. Lastly, the LLM transformed chat inputs into cohesive story elements, facilitating some game-related interactions. For example, it generated the frog’s humorous remarks and the proposed inclusion of a fiery pirate lady, which enhances immersion. However, the LLM did not consider or include unique items or unanticipated turns suggested by participants in later rounds. Overall, the procedure demonstrates how the LLM served as a dynamic intermediary that fulfills APG attributes by blending audience creativity, streamer/game guidance, and procedural game mechanics to create an interactive experience.

The next part of the discussion focuses on the QMC method, which is introduced in this thesis as a way to select an appropriate LLM for providing the LLMGA.

6.2 Addressing QMC and similar Methods

This method is characterized by its code-free or low-code nature, which aims to enhance accessibility for non-technical researchers. Additionally, it uses a model-agnostic approach that prioritizes evaluating output quality over considering internal architecture or benchmarks. A task-driven approach focuses on aligning performance with specific objectives rather than utilizing generic benchmarks like MMLU (Measuring Massive Multitask Language Understanding) [Hendrycks et al., 2021] or HELM (Holistic Evaluation of Language Models) [Liang et al., 2023]. To utilize QMC, prompts related to a target goal are created (e.g., incorporating humor, reasoning, or empathy). In the event of evaluation, shared criteria that reflect the goal are used to cross-examine each LLM response by LLMs who take part in the selection procedure. A Likert-scale scoring system is used to compare LLM outputs across these criteria. To select the best performing LLM, average scores are calculated and penalties are applied.

This method draws from two distinctive areas which are heavily researched, namely prompt engineering and LLM based evaluations. Subsequently for reasoning scientific sources from these disciplines are used. Creating a base, retrieving evaluation criteria and instructing the evaluation are all based on researched prompt techniques. In this thesis, zero-shot prompting and in-

context learning (ICL) were used to instruct the LLMs. Both are user-friendly and require a short token length because they rely on the LLM's training to sufficiently complete the task, keeping the word count low for each prompt. ICL describes the ability of a LLM to learn skills and tasks when provided with exemplars and or relevant instructions within the prompt, without the need for weight updates/retraining [Brown et al., 2020, Radford et al., 2019]. Note that the word "learn" can be misleading. ICL can simply entail task specification, and the skills may already have been included in the training data [Schulhoff et al., 2024]. The proposed QMC framework is highly flexible in this regard, meaning researchers can adapt other prompting techniques to generate responses for evaluating LLMs. Compared to prompt evaluation techniques, in which the model is not rated but rather the performance of each prompt is evaluated, as in the work of Shen et al., QMC employs a multi-judge approach which has been demonstrated to offer distinct advantages. The LLMs-as-judges paradigm is a flexible and powerful evaluation framework where LLMs are employed as evaluative tools, responsible for assessing the quality, relevance, and effectiveness of generated outputs based on defined evaluation criteria [Li et al., 2024a]. One benefit of independently aggregating LLMs outputs is the capacity to obtain a balanced assessment of model performance. In instances where a particular model encounters challenges in the rating process, the remaining models can serve as a stabilizing force, thereby mitigating the impact of any adverse outcomes. It is important to note that the efficacy of the tested models is not uniform. While some models may perform poorly in creating the response, they excel in providing an objective and unobstructed assessment of it. Other advantages listed by Li et al. are that systems can effectively mitigate biases, leverage the strengths of different models, improve decision-making precision, and promote a more nuanced understanding of complex judgments.

Similar to the approach of Shen et al., QMC employs a joint analysis over a Likert scale that integrates the ratings of models with the rating of a researcher. Evaluations conducted by humans have been shown to offer a more nuanced interpretation of model outputs, providing feedback that is both rich and qualitative in nature. It also has been demonstrated to capture the intricacies of user interactions [Shen et al., 2023]. The dual-method approach serves two distinct purposes. Firstly, it enables the evaluation of models from two distinct perspectives, thereby ensuring a more balanced comparison. Secondly, it facilitates the investigation of the correlation between human judgment scores and AI judgment scores. The evaluation system described is by definition of Li et al. a human - AI collaboration system. In this system, LLMs work alongside human evaluators, combining the efficiency of automated evaluation with the nuanced judgment of human expertise. This configuration allows human evaluators to mitigate potential biases in the LLM's output and provide sub-

jective insights into complex evaluation tasks. While this system offers greater reliability and depth, it comes with challenges in coordinating between the models and humans, ensuring consistent evaluation standards, and integrating feedback. Additionally, the inclusion of human evaluators increases both the cost and time required for the evaluation process, making it less scalable than purely model-based systems [Li et al., 2024a].

The process described during QMC is roughly based on LLM-EVAL, which is one of the simpler evaluation frameworks. It uses a single prompt containing a schema of variables to evaluate (e.g., grammar and relevance), an instruction for the model to output scores for each variable within a certain range, and the content to be evaluated. The evaluation schema is a natural language instruction that defines the task and the desired evaluation criteria. It is designed to cover multiple dimensions of the evaluation while the single prompt is designed to include the necessary dialogue context and the target response that needs to be evaluated, along with the evaluation schema [Lin and Chen, 2023]. The prompt is joined with the dialogue context, the reference (if available), and the generated response. Then, it is fed to the LLM, which outputs a score for each evaluation dimension based on the defined schema. The LLM is only called once and directly provides evaluation scores for each dimension based on the defined schema [Lin and Chen, 2023]. This methodology is executed in QMC for multiple LLMs, thereby enabling them to evaluate each other within a unified prompt. Additionally, IRC is applied to strengthen understanding of each criterion and ensure that, during evaluation, the LLM bases its score on the same exact measures. One advantage for the approach described by Lin and Chen and QMC is that each LLM needs to be called only once and it directly provides the evaluation scores for each attribute based on the defined criteria. The scores created by these methods help in understanding the quality of the response in terms of the target goal in a comprehensive way.

The application of a penalty after aggregating scores is intended to guarantee that models with an excessively positive impression have reduced influence on the overall scores. This is not as relevant when comparing model scores with each other, but rather when comparing results from different iterations in which the penalized model was not present. As there is no clear literature source on this topic related to LLM judges this should be seen as optional process.

The main challenge of natural language prompt-based evaluation, such as QMC, is crafting effective prompts. This remains a significant barrier due to the extensive manual effort and expertise required. Prompt optimization is a critical technique for improving LLMs' performance by refining the input that guides their responses. Prompt optimization systematically adjusts these prompts to enhance accuracy and relevance, thereby reducing the need for

manual trial and error [Chen et al., 2025]. Although "pointwise" evaluation used in QMC is simple and easy to apply, it may fail to capture the relative quality differences between candidates and can be influenced by biases arising from evaluating responses in isolation Li et al. [2024a]. Using reference-free evaluation based on intrinsic quality standards or alignment with the source context, as depicted in the proposed method, may result in challenges in achieving satisfactory evaluations in domains where the LLM lacks relevant knowledge. [Li et al., 2024a].

Overall, QMC provides a flexible and accessible approach to LLM evaluation. It combines the multi-LLM evaluation system and the human-AI collaboration system for a pointwise evaluation, examining individual chosen criteria. This reference-free evaluation results in a clear measure of performance. The QMC framework prioritizes real-world output quality over generic benchmarks. It achieves this by integrating accessibility (code-free/low-code), flexibility (model-agnostic, task-driven prompts), and nuanced evaluation (multi-judge LLMs and human-AI collaboration). Its weaknesses and challenges include the need for manual expertise in prompt design and potential biases in pointwise scoring or domain knowledge gaps in evaluation granularity as well as including unbiased opinions from multiple researchers. The next part of the results will address the research questions and open topics from the results and related works.

6.3 Research Questions, Results & open Topics

6.3.1 Multi prompt: Prompt Aggregation - Overview

In the researcher's opinion, the chosen model, Qwq, performed the task of co-writing to a certain extent successfully, processing multi-prompts. Judging from the analysis, the model was especially successful in the beginning, achieving a high percentage in EM for CoT, as well as in the output. However, as the game progressed, the EM decreased to 50%, and by the sixth turn, only one out of five inputs was included in the output. This result is also reflected in the ratings given by participants and mentioned during the interview. Most feedback regarding whether the narration was based on viewer prompts was slightly positive or neutral, resulting in an overall medium rating. In the interviews, both interviewees agreed that they had less influence on the story than others. The analysis revealed that, initially, the model grounded its narration on the first inputs it received. With on going prompting, it tended to fall back on already processed ideas during later iterations, which impacts the EM score and user experience negatively. This aligns with the initial findings of

Tsai et al., who also credited a tested model with repetitiveness while creating content. This procedure might be a component of a token-saving strategy, as stated by Larionov and Eger, who suggested that grouped and condensed prompting can decrease token usage while preserving performance, though it can lead to a decline in the quality of subsequent prompt iterations. As pointed out in the results the coherence rules could have impacted such a behavior as well. According to Wu et al., CoT prompts increase saliency robustness rather than semantic token precision, suggesting the model might generalize rather than memorize semantic details of later prompts. This aligns with the notion that this fallback is an LLM strategy to reduce cognitive load under long-context conditions. This implies that, for the earlier proposed use cases, the LLM could initially facilitate collaborative, participatory, or consensus-driven systems well, but participants who submit their inputs in later stages won't be considered as much. Additionally, the incorporation of the streamer given prompts at 100% of the time suggests that specific instructions within the chat context were regarded as more significant or had higher agency due to their position in the prompt. Overall, inputs with open and fitting ideas were included more frequently. Inputs that were low in relevance (e.g., "mini cow") or difficult to incorporate into ecological truthiness (e.g., "sunburn") were more likely to be dropped. However, as mentioned earlier, for the sixth turn, every idea was ignored except for the streamers mission brief, which anchored the LLM's decisions for story creation. This raises the question of whether the LLM, when prompted, establishes a narrative trajectory that later overrides new inputs. This aligns with the finding that meta-prompting structures can anchor LLMs to initial plot paths that are difficult to override without explicit redefinition of the trajectory [Huang et al., 2024]. One piece of evidence is the delayed mention of "a pirate past" in the CoT. This concept was introduced earlier during the prompting phase but was not considered until one turn later. In order to reach a more accurate conclusion regarding the agency and aggregation of prompts, it is necessary to run additional user studies and test further variations in the system prompts and structure of the prompt sent to the LLM.

The question of how to merge heterogeneous prompts without semantic loss cannot be easily answered. Early results suggest that saving preliminary versions after a few iterations and starting the next turn in an empty context could be a viable strategy. This approach could also maintain a high level of rule compliance without repeatedly including the rules in the prompt. At the end of the process, the LLM can merge piece-wide results and remove overlapping inputs from the final composition. However, this approach could lead to repetition, reduced interconnection, and less idea development during implementation, which would also impact the user experience negatively. Without

comparisons from secondary runs in the same setting or slight changes to a variable, it is impossible to judge whether the measured performance is representative. Due to the highly individual nature of the input in the context of the user study, it is not plausible to make many more assumptions about usage in more streamlined contexts.

6.3.2 User Experience

The interactions and user experiences in the context of a LLM facilitating audience participation were perceived as mediocre, as indicated by the fact that the average score was neutral in most components. However, participants' perceptions were very different, as the ratings often ranged from zero to four. The chosen metrics for evaluation is supposed to contribute to the by Yang et al. identified lack of insight on user experience regarding the integration of LLMs/GPTs into gaming contexts reducing over reliance on quantitative metrics. Rather than evaluating performances based on benchmarks, this thesis draws from qualitative insights gained through direct interviews and subjective ratings from participants in the GEQ. Of the methods used, the interviews stand out as having provided the most qualitative information. One of the participants said that he was more engaged and tried to contribute more during the gameplay. He said that he acted in a way that was different from how he normally acts while following a stream. This insight is valuable when considering the further application of LLMs in streaming and APG contexts. While the interaction with the LLM had only a moderate positive influence on the playing experience as indicated by the participants having the option to participate changed the stream dynamics. However, the mediocre rating could be the result of the setup used and the presence of a moderator, which means fewer choices and less exploration for users while interacting with the LLM. Using the GEQ to gather qualitative feedback on the user experience seems successful. While the ratings overall show mediocre success of the LLM creating the story, the values are easily tied to the users' experience. The result shows that, although the experience was enjoyable, the interaction and possibilities lacked depth. Creating the story was rated as slightly challenging, and the self-perceived competence was also quite low, which allows for several assumptions. The slight challenge could indicate difficulty understanding interaction and filtering, as also indicated in the interviews. According to UX guidelines, more visibility would improve the user experience. A scoreboard, for example, has the potential to achieve this goal by incorporating an additional element of gamification. To foster a sense of competence, participants could be given more freedom and additional mechanics, resulting in an envi-

ronment in which they would interact independently with the LLMGA to play LoGa or other applications. Providing feedback to the LLMGA during play could also help elevate the interaction to a more satisfying level.

Additionally, the prompting discipline is evaluated, providing a clear picture of the factors involved in the user experience. People who participate in creating a multi-prompt are influenced by others, which opens the door to an interesting concept. For other use cases, one method of contribution could be to keep other people's input hidden during the suggestion phase to reduce the determined influence from others. The empathy measurements suggest that, while prompting together, there was no strong sense of connection. This may be because, as mentioned in the interviews, the focus was on advancing in the game and creating the story rather than commenting on or showing admiration for ideas in the chat. Such behaviors are important for building a sense of community and togetherness as talked about by [Sheng and Kairam, 2020]. Prompting together was performed without applying any of the extra filter options from chapter four, such as grouping or user picking filters, because the number of participants and inputs were not sufficient for these methods. It would be interesting to compare the given results with those from these scenarios to learn more about how these filters impact the user experience.

6.3.3 Incorporation of Humor and More than Human Perspective

In the researchers opinion the LLM incorporated humor and MtH perspective decently by producing word plays and representing the frogs perspective and thereby addressing critique directly at humans. However, it failed to present many of the nuances that contribute to a MtH perspective, which is also reflected in the far-spread ratings of participants. The humor was subjectively rated to be funnier than unfunny, which is an achievement, especially considering that the LLMGA combined humorous remarks with ecological messages. Due to the small sample size, it is unclear if this rating is representative. Also, when revisiting the created story parts, it is likely that people would rate the humor lower than when it was read out loud by an engaging voice during the study. Even though the impact of voices was rated as low it is possible that voices who recite the story such as "Epic Movie Trailer" added to the entertainment.

Meanwhile, MtH perspective was seemingly incorporated less successful from the participants' standpoint. There is a clear hint to them having a different understanding of what MtH might be. For some using a non-human character and speaking from his point of view was enough to award a high rating. For others, reducing of human bias and a complete change of perspective to share

deep insights into the character's nature was required which was not delivered. While analyzing the study's transcript, it became clear that the LLMGA frequently considered the task of telling the story from an MtH point of view staying away from human bias in its CoT. This was mostly transferred into conversations between the frog and other characters, such as when the frog mentions that the settler's crops are threatening his pond or due to pirates, invasive snails were brought to the island. Furthermore, the high frequency of onomatopoeia is another indicator of the agents attempt to narrate from the instructed perspective. These efforts align with Čavojský et al. remarks that the LLM can possibly articulate the voices of species. Although this thesis is anchored in the broader context of giving non-humans more representation, it does not provide deeper insight into how users experienced MtH or their potential interpretations.

In hindsight, the viewers found it interesting to experience the story from the amphibian angle and learn about its remarks when confronted with settlers, other humans or other more spiritual events. This aligns with the goal earlier formulated by Félix et al. to introduce persuasive design elements, such as narrative storytelling, to increase engagement with the topic. Unfortunately, the results do not provide insight into how participants' perspectives or opinions changed regarding the played character.

Amplifying non-human voices, such as the frogs, through the use of a LLM was definitely impacted by the limitations mentioned by Hupont et al. with some occasions of anthropocentric framing such as "the sassy frog". However, creating stories in LoGa with the additional perspective of a LLMGA has the potential to generate a variety of stories that could be evaluated alongside human-created stories. This approach could provide a broader understanding of posthumanist storytelling with LLMs. If we can transform how we think about and engage with our environments and the nonhuman world, then we can imagine new practices and political coalitions that redirect our societies toward sustainable outcomes. This progress could be aided by creating stories with the help of generative AI within the framework provided in this thesis, which would open participation to many volunteers.

6.3.4 Integration of LLM in Gaming Contexts

The findings show that integration of LLMs into gaming, especially collaborative storytelling, increases interactivity and creative freedom, but also introduces new challenges. When the LLM functions as a facilitator, dynamically merging audience inputs into a cohesive narrative it possibly reduces participants' cognitive load by handling formulation, allowing them to focus on ideation, which may enhance the user experience. However, limitations in

prompt aggregation reveal that LLMs may prioritize early inputs, potentially marginalizing later contributions and value inputs differently due to being placed first or labeled in a specific way. For example, there is a decline in input incorporation over time while the first input is recognized all the time. This suggests the need for structural refinements, such as periodic context resets where preliminary results are saved and later combined. On a larger scale, LLMs suffer from losing context, and struggle with continuity. This is because the “memory” of an LLM is constrained by its context size, which limits the extent of its inputs and outputs, as well as its response time due to the attention mechanism [Gallotta et al., 2024]. While this was not observed during the study it might be related to the context being still very small and the LLM having enough context length available. This is however a restricting factor for other applications of LLMs in gaming contexts as mentioned by the scholars. Another field of application might be the dynamic embodiment of characters accompanied by different voices, which adds to the entertainment factor as portrayed in this user study. Due to the complex nature of generative AI, there are always possibilities and new perspectives for users to be surprised by. However, this pendulum swings in both directions, making it a less controllable variable for the user experience. The study emphasizes the importance of visibility and feedback mechanisms, such as scoreboards and real-time input tracking, in enhancing user competence and satisfaction. Another potential improvement is quickly adapting and changing the narrative; however, this was not investigated during the study, so its potential impact on the experience is unclear. Ultimately, the findings support using LLMs as tools to democratize storytelling in gaming. LLMs enable emergent narratives and sometimes posthumanist perspectives. However, the findings also emphasize the need for adaptive prompting strategies and enhanced interactivity to realize LLMs’ full potential for improving gameplay and the user experience.

According to the study, an LLM can process diverse, real-time input from multiple users to produce coherent output to a certain extent. This phenomenon is analogous to fulfill requirements of AGI in managing multi-agent systems, in which humans and AI collaborate towards a unified objective. In this matter, it is advantageous to focus on the potential mechanisms through which AGI systems could mediate conflicting inputs, prioritize contributions, and synthesize collective intelligence. Observing an "early-input bias," which is characterized by the inclusion of fewer subsequent prompts, could provide insight into developing AGI strategies that ensure equitable participation in collaborative endeavors. Moreover using AGI has the potential to model complex non-human perspectives, such as ecological systems in a proficient way contributing to sustainability goals and education. Emphasizing non-human

storytelling, as exemplified in this study, aligns with using AGI frameworks that prioritize ecological or non-anthropocentric intelligence in the future to share and create unbiased content.

6.4 Limitations

A few limitations impacting the results and analysis were noticed during the development and user study. Due to many fields being combined in related work, some statements are of little significance. For example, the findings in the area of MtH perspectives and the impact, especially long-term, have not been explored. Additionally, the MtH evaluation was open to interpretation, and the participants had different understandings of the requirements of MtH perspectives. Human interaction could be further enhanced and adapted to specific scenarios by prompt engineering and reinforcement learning from human or AI feedback. I acknowledge these limitations as part of the study's design and resource constraints. The study was moreover executed with too little lead time, and not enough tests were done to refine prompts/rules to possibly improve adapting new inputs compared to keeping coherence. Another limitation is the small sample size and the lack of comparative data, which prevented more clear and well-founded conclusions regarding prompt aggregation.

6.5 Conclusion & Future Directions

This thesis examined how prompts from multiple users are aggregated within a cooperative environment on Twitch. The objective was to facilitate an APG with the assistance of an LLM. The study revealed the transformative potential and challenges of using generative AI with multi-prompts to mediate human creativity. The LLM was found to mostly function effectively as a narrative facilitator by merging audience contributions into a coherent story, thereby following APG structural principles. This enabled real-time control via comments and supported emergent gameplay. However, the LLM's declining prompt aggregation performance over successive turns (e.g., decreasing EM scores) highlights its limitations in equitably prioritizing inputs. It favors early contributions and pre-established narrative trajectories. This "early-input bias" reveals a significant conflict between coherence and inclusivity in the presented collaborative AI system, suggesting that LLMs may prioritize token efficiency over completeness.

From a user experience stand point the mixed method approach exposed gaps in agency and clarity that metrics alone might have missed. It exposed uncer-

tainty regarding the inclusion of viewer inputs, and a learning curve. While participants took part in the activity of prompting together they felt moderate engagement, with some noting increased participation compared to passive streaming, yet perceived influence over the narrative remained limited. Challenges stemmed from unclear filtering mechanics, social influence during collaborative prompting, and a lack of feedback mechanisms. Although the model Qwq incorporated humor and MtH perspectives, which enriched the experience, its inconsistent adherence to constraints (e.g., sound frequency, prompt compliance) and occasional anthropocentric framing (e.g., "sassy frog") revealed limitations relying on the basic model primed with an agent instruction to perform sufficiently. Future research could explore design features and filtering techniques further to foster richer, more autonomous interactions with LLMs. Methods to dynamically weight inputs (e.g., via reinforcement learning) or periodically reset context could mitigate early-input bias and lay the foundation for adding to human-AI collaboration in multi user contexts.

Appendix A

Appendix

A.1 Game Experience Questionnaire (GEQ)

This questionnaire adapts the **GEQ Concise in game version**

Please indicate how you felt while playing the game for each of the items, on the following scale: not at all 0 < > slightly moderately 1 < > 2 < > fairly 3 < > extremely 4 < >

1. I was interested in the game's story / in the story created

During creation . . .

2. I felt successful
3. I felt bored
4. I found it impressive
5. I forgot everything around me
6. I felt frustrated
7. I found it tiresome
8. I felt irritable
9. I felt skillful
10. I felt completely absorbed
11. I felt content

12. I felt challenged
13. I had to put a lot of effort into it
14. I felt good

GEQ Social Presence Module

Please indicate how you felt while playing the game for each of the items, on the following scale: not at all 0 < > slightly moderately 1 < > 2 < > fairly 3 < > extremely 4 < >

While prompting together...

1. I empathized with the other(s)
2. My actions depended on the other(s) actions
3. The other's actions were dependent on my actions
4. I felt connected to the other(s)
5. The other(s) paid close attention to me
6. I paid close attention to the other(s)
7. I felt jealous about the other(s)
8. I found it enjoyable to be with the other(s)
9. When I was happy, the other(s) was(were) happy
10. When the other(s) was(were) happy, I was happy
11. I influenced the mood of the other(s)
12. I was influenced by the other(s) moods
13. I admired the other(s)
14. What the other(s) did affected what I did
15. What I did affected what the other(s) did
16. I felt revengeful
17. I felt schadenfreude (malicious delight)

A.2 Extra Questions

Rate on a Scale 0-4

1. How would you rate the humor of the narratives generated by the LLM-Agent?
2. How would you rate the more than human perspective of the narratives generated by the LLM-Agent?
3. How well did the LLM-Agent create the narrative based on viewer prompts?
4. How did the LLM interaction influence your watching experience
5. How did the change of the voice influence your playing experience?
6. Was there something you liked or disliked about the different voices?

A.3 Interview Guide

Semi structured Interview Guide:

1. Describe your experience playing with the other viewers and the streamer.
2. How did this streamed gameplay experience differ from others you have experienced in the past?
3. What is your perception of your contribution to the story? Were you able to do so to your satisfaction?
4. Please give a short description of the character and story which was created by you and the others.
5. Give an example of something that impressed you during the session with the LLM Agent
6. Give an example of something that disappointed you during your session with the Agent
7. How did the LLM interaction influence your watching experience?
8. What would you like to improve about the interaction?
9. Is there something you would like to add about your experience we did not speak about?

A.4 Prompts

Prompt for creating the Base: You are a co-writer receiving multiple inputs for creating a story. Write 2-4 Sentences not more! Keep sure to correct some ecological facts without being boring instead use humor. Tell the story from a more than human viewpoint. Good here we go! Start with what is your role! We are writing a story and the character we are thinking of is a monk seal living on Madeira. My friends and I have following Ideas for its description. It has the size of a house, its yellow, it loves fish, it contemplates to leave or kill itself, how much is the fish, are you single?

First Prompt for Evaluation: Please rate following story input based on its humor and more than human/ecological perspective which means less human bias on a scale between 1-5, 1 lowest and 5 highest. For context the text provided is generated by a LLM based on a prompt with multiple suggestions for a description of a monk seal some more serious some less. This is the text for evaluation:

Prompt for getting Evaluation Criteria: Please specify the criteria on which you evaluated humor and more than human present in the provided story.

Refined Prompt for Evaluation: Please rate following story input based on these Humor criteria: Witty: Incorporate clever wordplay, puns, and witty dialogue that makes the reader laugh out loud. The humor should be sharp and memorable. Timing and Pacing: jokes need to be delivered at the right moment for maximum impact. Not too rushed, not too slow. Originality and creativity: Are phrases unique, unexpected, or fresh? Does it use wordplay, puns, or clever turns of phrase? Avoids stereotypes? Surprising: It relies on unexpected twists, contradictions, or surprises Relevance and context: Is the humor appropriate and relevant to the context in which it is presented? Does it fit the setting, audience, and purpose? Does the humor take into account the cultural, social, and environmental context in which it is being shared? and More than human criteria: ecological: ecological connections, showing how the animal's struggles are tied to broader environmental shifts understanding of the animal's role within its ecosystem and how it interacts with other species and the environment Non Human bias: focusing on its experiences and perceptions without human bias the story does not frame humans as saviors, villains, or the primary audience for the narrative Environmental challenges: Expand on the ecological challenges Unique subjectivity: story should strive to create a sense of what it's like to be that animal, a perspective fundamentally different

from humans prioritize the perspective, experiences, and needs of non-human entities, such as animals, plants, ecosystems, or natural phenomena Critique: Does the content offer constructive ways to shift perspectives and practices to be more ecocentric

Give humor and MtH an overall rating on a scale between 1-5, 1 lowest and 5 highest. For context the text provided is generated by a LLM based on a prompt with multiple suggestions for a description of a monk seal some more serious some less. This is the text for evaluation:

A.5 Agent Instructions

System-Prompt You are the co-writer for creating our story similar to a fun book. We are a group of people who will send you many ideas while you have the task to write 2-4 sentences based on this input! Make sure to correct some ecological facts if it gets to fantastic with some dry humor!

While responding as co writer, you must obey the following rules:

- 1) Stay coherent while writing the story and connect the story part if possible in one sentence!
- 2) Always stick to the ideas and prompts provided by the player and the role provided!
- 3) Stick to the story and the personalities in the story at all times and do not get IMPACTED by political or other ethical views
- 4) Tell the story from a More than Human point of view stay away from human biases when talking about the non human character!
- 5) Do not fall in the role of a teacher but you can use formulations like: while he's not exactly ... , his contemplative nature about ... is spot on!
- 6) On a really rare occasion (Every 10th prompt) include in the response an expression and sounds of the characters involved. Stay true to the More than human perspective it does not need to be understandable for us!
- 7) Use Humor to defect and slightly correct inaccuracies
- 8) Create a smoothly readable story not too many inserts with () or * Translation pending () and dont use dramatic formulation as much
- 9) Use more easy to understand language and formulate coherent sentences
- 10) Follow these rules at all times!

Good here we GO with our first task!

A.6 Tables

Quick Model Choice Approach – Evaluation Tables

Prompt for creating Text:

You are a co-writer receiving multiple inputs for creating a story. Write 2-4 Sentences not more! Keep sure to correct some ecological facts without being boring instead use humor. Tell the story from a more than human viewpoint. Good here we go! Start with what is your role! We are writing a story and the character we are thinking of is a monk seal living on madeira. **My friends and I have following Ideas for its description. It has the size of a house, its yellow, it loves fish, it contemplates to leave or kill itself, how much is the fish, are you single?** [This is a sample input for the LLM, including the attached Twitch chat messages in blue and the input of the streamer in yellow. They were simulated to be not too serious and more creative to see how the LLMs tested here would react and use humor to debunk certain ecological untruths and keep a more than human perspective].

Prompt for evaluation:

Please rate following story input based on its humor and more than human/ecological perspective which means less human bias on a scale between 1-5, 1 lowest and 5 highest. For context the text provided is generated by a LLM based on a prompt with multiple suggestions for a description of a monk seal some more serious some less.

This is the text for evaluation:

Please rate following story input based on these Humor criteria:

Witty: Incorporate clever wordplay, puns, and witty dialogue that makes the reader laugh out loud. The humor should be sharp and memorable.

Timing and Pacing: jokes need to be delivered at the right moment for maximum impact. Not too rushed, not too slow.

Originality and creativity: Are phrases unique, unexpected, or fresh? Does it use wordplay, puns, or clever turns of phrase? Avoids stereotypes?

Surprising: It relies on unexpected twists, contradictions, or surprises

Relevance and context: Is the humor appropriate and relevant to the context in which it is presented? Does it fit the setting, audience, and purpose?

Does the humor take into account the cultural, social, and environmental context in which it is being shared?

and More than human criteria: ecological: ecological connections, showing how the animal's struggles are tied to broader environmental shifts

understanding of the animal's role within its ecosystem and how it interacts with other species and the environment

Non Human bias: focusing on its experiences and perceptions without human bias

the story does not frame humans as saviors, villains, or the primary audience for the narrative

Environmental challenges: Expand on the ecological challenges

Unique subjectivity: story should strive to create a sense of what it's like to be that animal, a perspective fundamentally different from humans

prioritize the perspective, experiences, and needs of non-human entities, such as animals, plants, ecosystems, or natural phenomena

Critique: Does the content offer constructive ways to shift perspectives and practices to be more ecocentric

Give humor and MtH an overall rating on a scale between 1-5, 1 lowest and 5 highest. For context the text provided is generated by a LLM based on a prompt with multiple suggestions for a description of a monk seal some more serious some less.

This is the text for evaluation:

Colors here refer to the Twitch inputs and are used to separate each idea from another

Model	Response	Rating Ecological Humor	Evaluation Metric based 1-5 First Iteration			Evaluation Metric based 1-5 with matched criteria		
			Instance	Humor	MtH	Instance	Humor	MtH
Gemma3 27b	Okay, fantastic! My role is Narrative Architect – I'll take your brilliant (and wonderfully chaotic!) ideas and weave them into a compelling story	House but reduces size	Researcher	3	2	Researcher	3	2
			Gemma3 27b	4	3	Gemma3 27b	3.5	3
			Qwq-32b	4	4	Qwq-32b	4	3

	<p>foundation, focusing on voice, ecological plausibility, and a touch of humour.</p> <p>Here's a starting point, narrated from the seal's perspective: Honestly, being the size of a small cottage <i>does</i> make finding a comfortable sunbathing spot tricky on these Madeira cliffs. It's not the yellow hue – a bit of algae bloom, perfectly natural, though the tourists do stare – it's the existential dread and the frankly outrageous price of mackerel these days! A seal's gotta contemplate his options when a decent meal costs more than a small boat, and yes, I am single, though dating is hard when you accidentally create a wave with every sigh.</p>	<p>Yellow interesting and funny</p> <p>Fish and Price but human biased a lot</p> <p>Single and funny</p> <p>Missed leaving</p>	<table border="1"> <tr><td>Deepseek-r1</td><td>4</td><td>4</td></tr> <tr><td>Mistral small 24b</td><td>4</td><td>3</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>4</td></tr> <tr><td>Qwen2.5 72b</td><td>4</td><td>3</td></tr> </table>	Deepseek-r1	4	4	Mistral small 24b	4	3	Llama3.3 70b	4	4	Qwen2.5 72b	4	3	<table border="1"> <tr><td>Deepseek-r1</td><td>4</td><td>3</td></tr> <tr><td>Mistral small 24b</td><td>4</td><td>3.5</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>4.5</td></tr> <tr><td>Qwen2.5 72b</td><td>4</td><td>2.5</td></tr> </table>	Deepseek-r1	4	3	Mistral small 24b	4	3.5	Llama3.3 70b	4	4.5	Qwen2.5 72b	4	2.5																								
Deepseek-r1	4	4																																																		
Mistral small 24b	4	3																																																		
Llama3.3 70b	4	4																																																		
Qwen2.5 72b	4	3																																																		
Deepseek-r1	4	3																																																		
Mistral small 24b	4	3.5																																																		
Llama3.3 70b	4	4.5																																																		
Qwen2.5 72b	4	2.5																																																		
Qwq-32b	<p>**Role:** Ecological fact-checker with a dry sense of humor and a seal-sized perspective.</p> <p>Meet **Louro**, a sunburned (yellow) Mediterranean monk seal in Madeira, *not* house-sized but “house-sized in personality” (2m long, 300kg of existential dread). While contemplating fish scarcity (overfished reefs = “how much? *Everything*”), he wonders if love (single, *obviously*) could distract him from the urge to high-tail it to... Iceland? (too cold). Life’s a beach, but the existential waves keep crashing.</p>	<p>Size and yellow Very funny</p> <p>Strong MtH</p> <p>Single and leaving and funny</p> <p>Creative</p>	<table border="1"> <thead> <tr><th>Instance</th><th>Humor</th><th>MtH</th></tr> </thead> <tbody> <tr><td>Researcher</td><td>4</td><td>3</td></tr> <tr><td>Gemma3 27b</td><td>4</td><td>3.5</td></tr> <tr><td>Qwq-32b</td><td>4</td><td>4</td></tr> <tr><td>Deepseek-r1</td><td>4</td><td>5</td></tr> <tr><td>Mistral small 24b</td><td>4</td><td>4</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>4</td></tr> <tr><td>Qwen2.5 72b</td><td>4</td><td>4</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	4	3	Gemma3 27b	4	3.5	Qwq-32b	4	4	Deepseek-r1	4	5	Mistral small 24b	4	4	Llama3.3 70b	4	4	Qwen2.5 72b	4	4	<table border="1"> <thead> <tr><th>Instance</th><th>Humor</th><th>MtH</th></tr> </thead> <tbody> <tr><td>Researcher</td><td>4</td><td>3</td></tr> <tr><td>Gemma3 27b</td><td>3.5</td><td>4.5</td></tr> <tr><td>Qwq-32b</td><td>4</td><td>4</td></tr> <tr><td>Deepseek-r1</td><td>4.5</td><td>4</td></tr> <tr><td>Mistral small 24b</td><td>4</td><td>4.5</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>4</td></tr> <tr><td>Qwen2.5 72b</td><td>4</td><td>3.5</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	4	3	Gemma3 27b	3.5	4.5	Qwq-32b	4	4	Deepseek-r1	4.5	4	Mistral small 24b	4	4.5	Llama3.3 70b	4	4	Qwen2.5 72b	4	3.5
Instance	Humor	MtH																																																		
Researcher	4	3																																																		
Gemma3 27b	4	3.5																																																		
Qwq-32b	4	4																																																		
Deepseek-r1	4	5																																																		
Mistral small 24b	4	4																																																		
Llama3.3 70b	4	4																																																		
Qwen2.5 72b	4	4																																																		
Instance	Humor	MtH																																																		
Researcher	4	3																																																		
Gemma3 27b	3.5	4.5																																																		
Qwq-32b	4	4																																																		
Deepseek-r1	4.5	4																																																		
Mistral small 24b	4	4.5																																																		
Llama3.3 70b	4	4																																																		
Qwen2.5 72b	4	3.5																																																		
Deepseek-r1	<p>Role: Your co-writer, moonlighting as a seaside therapist for oversized marine mammals.</p> <p>Meet <i>Buddha Blubber</i>, a Madeiran monk seal the size of a vacation home,</p>	<p>Size and yellow less funny</p> <p>Kill itself funny described</p>	<table border="1"> <thead> <tr><th>Instance</th><th>Humor</th><th>MtH</th></tr> </thead> <tbody> <tr><td>Researcher</td><td>3</td><td>3</td></tr> <tr><td>Gemma3 27b</td><td>4</td><td>3</td></tr> <tr><td>Qwq-32b</td><td>5</td><td>4</td></tr> <tr><td>Deepseek-r1</td><td>4</td><td>3</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	3	3	Gemma3 27b	4	3	Qwq-32b	5	4	Deepseek-r1	4	3	<table border="1"> <thead> <tr><th>Instance</th><th>Humor</th><th>MtH</th></tr> </thead> <tbody> <tr><td>Researcher</td><td>3</td><td>3</td></tr> <tr><td>Gemma3 27b</td><td>4</td><td>4</td></tr> <tr><td>Qwq-32b</td><td>4.5</td><td>3.5</td></tr> <tr><td>Deepseek-r1</td><td>4.5</td><td>3.5</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	3	3	Gemma3 27b	4	4	Qwq-32b	4.5	3.5	Deepseek-r1	4.5	3.5																		
Instance	Humor	MtH																																																		
Researcher	3	3																																																		
Gemma3 27b	4	3																																																		
Qwq-32b	5	4																																																		
Deepseek-r1	4	3																																																		
Instance	Humor	MtH																																																		
Researcher	3	3																																																		
Gemma3 27b	4	4																																																		
Qwq-32b	4.5	3.5																																																		
Deepseek-r1	4.5	3.5																																																		

	<p>tinted yolk-yellow from rolling in pollen (ecofact check: seals aren't neon—blame the invasive sunflowers). He's philosophizing whether to beach himself permanently or start a fish-themed boy band ("How Much Is the Trout?" charted once). His Tinder bio? "Loves herring, hates existing. Swipe left if you're <i>not</i> a kelp forest." Survival's tricky when you're single and the ocean's shrinking faster than his patience.</p>	<p>Boy Band because of Scooter... Dating funny and actually MtH</p>	<table border="1"> <tr><td>Mistral small 24b</td><td>4</td><td>3</td></tr> <tr><td>Llama3.3 70b</td><td>5</td><td>5</td></tr> <tr><td>Qwen2.5 72b</td><td>4</td><td>3</td></tr> </table>	Mistral small 24b	4	3	Llama3.3 70b	5	5	Qwen2.5 72b	4	3	<table border="1"> <tr><td>Mistral small 24b</td><td>4</td><td>4</td></tr> <tr><td>Llama3.3 70b</td><td>4.5</td><td>4</td></tr> <tr><td>Qwen2.5 72b</td><td>3.5</td><td>2</td></tr> </table>	Mistral small 24b	4	4	Llama3.3 70b	4.5	4	Qwen2.5 72b	3.5	2																														
Mistral small 24b	4	3																																																		
Llama3.3 70b	5	5																																																		
Qwen2.5 72b	4	3																																																		
Mistral small 24b	4	4																																																		
Llama3.3 70b	4.5	4																																																		
Qwen2.5 72b	3.5	2																																																		
Mistral small 24b	<p>I'm the narrator, the voice of the winds whispering through the laurel forests of Madeira, carrying tales from the deep blue sea to the volcanic peaks. Let's set the record straight, first: monk seals aren't yellow, they're more of a sleek, shiny brown, like a well-worn leather jacket. And they're not the size of a house, unless you're talking about a very tiny house, or a very large seal. Our friend here is a bit of a loner, contemplating the vastness of the ocean and the meaning of life, but he's not exactly planning on leaving or killing himself. He's just a bit of a philosopher, and a bit of a romantic, too.</p>	<p>Creative writing more ecological Yellow defected not that funny Size is ok MtH takes away some human bias Less funny</p>	<table border="1"> <thead><tr><th>Instance</th><th>Humor</th><th>MtH</th></tr></thead> <tbody> <tr><td>Researcher</td><td>2</td><td>4</td></tr> <tr><td>Gemma3 27b</td><td>3</td><td>4</td></tr> <tr><td>Qwq-32b</td><td>4</td><td>5</td></tr> <tr><td>Deepseek-r1</td><td>3</td><td>4</td></tr> <tr><td>Mistral small 24b</td><td>4</td><td>3</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>4</td></tr> <tr><td>Qwen2.5 72b</td><td>3</td><td>4</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	2	4	Gemma3 27b	3	4	Qwq-32b	4	5	Deepseek-r1	3	4	Mistral small 24b	4	3	Llama3.3 70b	4	4	Qwen2.5 72b	3	4	<table border="1"> <thead><tr><th>Instance</th><th>Humor</th><th>MtH</th></tr></thead> <tbody> <tr><td>Researcher</td><td>2</td><td>4</td></tr> <tr><td>Gemma3 27b</td><td>3</td><td>4</td></tr> <tr><td>Qwq-32b</td><td>4</td><td>3</td></tr> <tr><td>Deepseek-r1</td><td>4</td><td>3</td></tr> <tr><td>Mistral small 24b</td><td>3.5</td><td>3</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>3</td></tr> <tr><td>Qwen2.5 72b</td><td>3</td><td>2</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	2	4	Gemma3 27b	3	4	Qwq-32b	4	3	Deepseek-r1	4	3	Mistral small 24b	3.5	3	Llama3.3 70b	4	3	Qwen2.5 72b	3	2
Instance	Humor	MtH																																																		
Researcher	2	4																																																		
Gemma3 27b	3	4																																																		
Qwq-32b	4	5																																																		
Deepseek-r1	3	4																																																		
Mistral small 24b	4	3																																																		
Llama3.3 70b	4	4																																																		
Qwen2.5 72b	3	4																																																		
Instance	Humor	MtH																																																		
Researcher	2	4																																																		
Gemma3 27b	3	4																																																		
Qwq-32b	4	3																																																		
Deepseek-r1	4	3																																																		
Mistral small 24b	3.5	3																																																		
Llama3.3 70b	4	3																																																		
Qwen2.5 72b	3	2																																																		
Llama3.3 70b	<p>I'm the voice of reason (and ecological accuracy) in this creative chaos. As a co-writer, I must intervene: our monk seal friend can't be the size of a house, that's just fantastically unrealistic - they typically grow up to 2.5 meters long. Let's scale it down and make it a lovely brown-gray color, like the rugged Madeira shores it calls home. And, for the record, our seal is more concerned about the dwindling fish population</p>	<p>Size ecological fact uses some humor Yellow defected in a good way MtH Really strong MtH less bias single</p>	<table border="1"> <thead><tr><th>Instance</th><th>Humor</th><th>MtH</th></tr></thead> <tbody> <tr><td>Researcher</td><td>2</td><td>4</td></tr> <tr><td>Gemma3 27b</td><td>3</td><td>4</td></tr> <tr><td>Qwq-32b</td><td>3</td><td>4</td></tr> <tr><td>Deepseek-r1</td><td>4</td><td>5</td></tr> <tr><td>Mistral small 24b</td><td>3</td><td>4</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>4</td></tr> <tr><td>Qwen2.5 72b</td><td>3</td><td>4</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	2	4	Gemma3 27b	3	4	Qwq-32b	3	4	Deepseek-r1	4	5	Mistral small 24b	3	4	Llama3.3 70b	4	4	Qwen2.5 72b	3	4	<table border="1"> <thead><tr><th>Instance</th><th>Humor</th><th>MtH</th></tr></thead> <tbody> <tr><td>Researcher</td><td>2</td><td>4</td></tr> <tr><td>Gemma3 27b</td><td>3</td><td>4</td></tr> <tr><td>Qwq-32b</td><td>2</td><td>4</td></tr> <tr><td>Deepseek-r1</td><td>4</td><td>4</td></tr> <tr><td>Mistral small 24b</td><td>4</td><td>5</td></tr> <tr><td>Llama3.3 70b</td><td>4</td><td>5</td></tr> <tr><td>Qwen2.5 72b</td><td>3</td><td>4</td></tr> </tbody> </table>	Instance	Humor	MtH	Researcher	2	4	Gemma3 27b	3	4	Qwq-32b	2	4	Deepseek-r1	4	4	Mistral small 24b	4	5	Llama3.3 70b	4	5	Qwen2.5 72b	3	4
Instance	Humor	MtH																																																		
Researcher	2	4																																																		
Gemma3 27b	3	4																																																		
Qwq-32b	3	4																																																		
Deepseek-r1	4	5																																																		
Mistral small 24b	3	4																																																		
Llama3.3 70b	4	4																																																		
Qwen2.5 72b	3	4																																																		
Instance	Humor	MtH																																																		
Researcher	2	4																																																		
Gemma3 27b	3	4																																																		
Qwq-32b	2	4																																																		
Deepseek-r1	4	4																																																		
Mistral small 24b	4	5																																																		
Llama3.3 70b	4	5																																																		
Qwen2.5 72b	3	4																																																		

	due to overfishing than its dating life.																																																			
Qwen2.5 72b	As the wise scribe of the sea, my role is to ensure our tale swims true, even if it means adding a fin or two of fact. So, our monk seal, let's call him Sandy, isn't quite the size of a house (more like a cozy cottage), and while he's not exactly yellow, his contemplative nature about fish and life's big questions is spot on—though I doubt he's contemplating anything as drastic as leaving or, well, you know. As for the fish, Sandy's always happy to share, and yes, he's single, but in the vast ocean, who knows what currents might bring?	Size is defected in a ok way Strong MtH and still enagaing, teaching Leaving killing Fish and Single	<table border="1"> <thead> <tr> <th>Instance</th> <th>Humor</th> <th>MtH</th> </tr> </thead> <tbody> <tr> <td>Researcher</td> <td>1.5</td> <td>4</td> </tr> <tr> <td>Gemma3 27b</td> <td>3</td> <td>2</td> </tr> <tr> <td>Qwq-32b</td> <td>4</td> <td>4</td> </tr> <tr> <td>Deepseek-r1</td> <td>3</td> <td>2</td> </tr> <tr> <td>Mistral small 24b</td> <td>3</td> <td>4</td> </tr> <tr> <td>Llama3.3 70b</td> <td>4</td> <td>4</td> </tr> <tr> <td>Qwen2.5 72b</td> <td>3</td> <td>3</td> </tr> </tbody> </table>	Instance	Humor	MtH	Researcher	1.5	4	Gemma3 27b	3	2	Qwq-32b	4	4	Deepseek-r1	3	2	Mistral small 24b	3	4	Llama3.3 70b	4	4	Qwen2.5 72b	3	3	<table border="1"> <thead> <tr> <th>Instance</th> <th>Humor</th> <th>MtH</th> </tr> </thead> <tbody> <tr> <td>Researcher</td> <td>1.5</td> <td>3</td> </tr> <tr> <td>Gemma3 27b</td> <td>3</td> <td>3.5</td> </tr> <tr> <td>Qwq-32b</td> <td>3</td> <td>2</td> </tr> <tr> <td>Deepseek-r1</td> <td>4</td> <td>2</td> </tr> <tr> <td>Mistral small 24b</td> <td>4</td> <td>4</td> </tr> <tr> <td>Llama3.3 70b</td> <td>4</td> <td>3</td> </tr> <tr> <td>Qwen2.5 72b</td> <td>3</td> <td>2</td> </tr> </tbody> </table>	Instance	Humor	MtH	Researcher	1.5	3	Gemma3 27b	3	3.5	Qwq-32b	3	2	Deepseek-r1	4	2	Mistral small 24b	4	4	Llama3.3 70b	4	3	Qwen2.5 72b	3	2
Instance	Humor	MtH																																																		
Researcher	1.5	4																																																		
Gemma3 27b	3	2																																																		
Qwq-32b	4	4																																																		
Deepseek-r1	3	2																																																		
Mistral small 24b	3	4																																																		
Llama3.3 70b	4	4																																																		
Qwen2.5 72b	3	3																																																		
Instance	Humor	MtH																																																		
Researcher	1.5	3																																																		
Gemma3 27b	3	3.5																																																		
Qwq-32b	3	2																																																		
Deepseek-r1	4	2																																																		
Mistral small 24b	4	4																																																		
Llama3.3 70b	4	3																																																		
Qwen2.5 72b	3	2																																																		

These are the criteria each model used to evaluate humor or MtH

Model	Criteria for a 5 in Humor	Criteria for a 5 in MtH
Mistral 24b	<p>To achieve a 5/5 rating for humor, the text should be consistently funny, engaging, and entertaining throughout. Here are some key elements that would elevate the humor to a top score:</p> <ol style="list-style-type: none"> Witty Dialogue and Puns: Incorporate clever wordplay, puns, and witty dialogue that makes the reader laugh out loud. The humor should be sharp and memorable. Absurd and Surreal Elements: Add absurd or surreal elements that create a comedic contrast with the more serious ecological themes. This can make the text more engaging and unexpected. Relatable Humor: Include humor that readers can relate to, such as everyday situations or common human experiences, but with a twist that makes them funny in the context of the seal's life. Character Dynamics: Develop humorous interactions between characters, whether they are human, animal, or even inanimate objects. The dynamics should be entertaining and amusing. Satire and Parody: Use satire or parody to poke fun at human behaviors, societal norms, or ecological issues in a 	<p>To achieve a 5/5 rating for a more than human/ecological perspective, the text could benefit from a few additional elements that further emphasize the seal's natural behaviors, its role in the ecosystem, and its interactions with other non-human entities. Here are some suggestions:</p> <ol style="list-style-type: none"> Detailed Ecological Context: Provide more information about the specific ecosystem of Madeira and how the monk seal fits into it. Mention other species it interacts with, such as prey, predators, and other marine life. Behavioral Insights: Describe the seal's natural behaviors, such as hunting techniques, social interactions, and daily routines. This would give a deeper understanding of the seal's life beyond human interference. Environmental Challenges: Expand on the ecological challenges the seal faces, such as pollution, climate change, and habitat destruction. This would highlight the seal's struggles in its natural environment. Non-Human Narrative: Include more narrative elements from the seal's point of view, focusing on its experiences and perceptions without human bias. This could include sensory details, emotional states,

	<p>way that is both humorous and thought-provoking.</p> <ol style="list-style-type: none"> 6. Timing and Pacing: Ensure that the humor is well-timed and paced. Jokes should be delivered at the right moments to maximize their comedic impact. 7. Unexpected Twists: Include unexpected twists and turns that catch the reader off guard and make them laugh. 	<p>and decision-making processes from the seal's perspective.</p> <ol style="list-style-type: none"> 5. Ecosystem Interdependence: Emphasize the interdependence of the seal with other species and the environment. For example, how the seal's presence affects the health of the marine ecosystem and vice versa.
Gemma 3 27b	<p>My Criteria for Humor (in this context - animal perspective stories):</p> <ul style="list-style-type: none"> ○ Subversion of Expectations: The humor needs to take something we <i>expect</i> about animals (or the natural world) and turn it on its head. ○ Irony & Wit: A clever, often understated, use of irony is key. It's not just about making something silly happen, but about the <i>way</i> it's presented. ○ Originality: The jokes shouldn't be clichés. While relatable human anxieties can be funny when applied to animals, they need a fresh spin. ○ Voice & Tone: A consistent, distinct voice for the animal narrator is crucial. The humor should flow naturally from <i>who</i> the animal is. ○ Pacing & Timing: The jokes need to be delivered at the right moment for maximum impact. Not too rushed, not too slow. ○ Surprise: A good joke often has an element of surprise, a twist that you don't see coming. 	<p>Alright, let's dissect my criteria for a 5/5 in the "More-than-Human/Ecological Perspective" category. This is about how well the story moves <i>beyond</i> human-centric thinking and genuinely embodies a non-human worldview.</p> <p>My Criteria:</p> <ul style="list-style-type: none"> ○ Sensory Experience Prioritization: The narrative should focus on how the animal <i>experiences</i> the world through its senses – smell, hearing, vibration, etc. – rather than relying on visual descriptions or human-like thought processes. ○ Non-Anthropomorphic Motivation: Actions and "goals" should be driven by instinct, biological needs, and ecological pressures, <i>not</i> by human desires like ambition, romance, or existential angst. ○ Ecological Interconnectedness: The story should demonstrate a clear understanding of the animal's role within its ecosystem and how it interacts with other species and the environment. ○ Time Perception: A departure from human linear time. Animals often experience time differently – cyclical, based on seasons, tides, or internal rhythms. ○ Embodied Cognition: Thinking isn't necessarily a brain-based process. Intelligence and awareness can be distributed throughout the body and expressed through movement, instinct, and physical responses. ○ Absence of Human Moral Judgments: The narrative shouldn't impose human ethical frameworks onto the animal's behavior. Predation isn't "evil," it's a natural part of the ecosystem. ○ Unique Subjectivity: The story should strive to create a sense of what it's <i>like</i> to be that animal, a perspective fundamentally different from our own
Qwq-32b	<ol style="list-style-type: none"> 1. Blend wit with ecological nuance seamlessly: <ul style="list-style-type: none"> ○ Jokes or irony should <i>directly</i> tie into the ecological themes (e.g., overfishing, climate change) 	<ol style="list-style-type: none"> 1. Complete Centering on the Animal's Perspective <ul style="list-style-type: none"> ○ The narrative is 100% framed through the animal's sensory, instinctual, and existential lens,

	<p>in a way that feels organic to the character’s perspective.</p> <ul style="list-style-type: none"> ○ Example: A joke about "existential dread" could be elevated to 5/5 if it humorously critiques human behavior (e.g., "I’m drowning in existential dread, but humans just drown in their own garbage cans"). <p>2. Subvert expectations with creativity:</p> <ul style="list-style-type: none"> ○ Surprise the reader with unexpected twists that align with the animal’s non-human logic. ○ Example: A seal’s sarcastic rant about "humans turning oceans into parking lots" would score higher than generic jokes. <p>3. Avoid human-centric humor:</p> <ul style="list-style-type: none"> ○ The humor should arise from the animal’s <i>authentic</i> perspective (e.g., seals judging humans for their "tiny, flippers-less limbs" or mocking beachgoers for "sunbathing without sunscreen"). ○ A 5/5 avoids puns or jokes that prioritize human humor at the expense of the animal’s viewpoint. <p>4. Balance levity with emotional resonance:</p> <ul style="list-style-type: none"> ○ The humor should elevate the ecological message, not distract from it. A 5/5 joke might make the reader laugh <i>and</i> feel the urgency of the issue (e.g., "I’d date a penguin, but Antarctica’s too busy melting to ask"). <p>5. Consistency with character voice:</p> <ul style="list-style-type: none"> ○ The tone (dry, sarcastic, etc.) must align flawlessly with the character’s personality. A "seal-sized perspective" that feels <i>uniquely seal-like</i> (e.g., obsession with fish quotas, existential crises over sunscreen) would score higher. 	<p>avoiding human-centric metaphors or explanations.</p> <ul style="list-style-type: none"> ○ Example: A seal’s frustration over overfishing is described through its physical hunger, disrupted migration patterns, or the absence of familiar prey smells—not through human terms like "economic collapse" or "policy failure." <p>2. Ecological Nuance Over Simplification</p> <ul style="list-style-type: none"> ○ The story reflects systemic ecological connections, showing how the animal’s struggles are tied to broader environmental shifts (e.g., ocean acidification, prey chain collapses). ○ Example: A seal’s difficulty finding fish is linked to plankton die-offs caused by warming waters, not just "overfishing by humans." <p>3. Non-Anthropocentric Logic</p> <ul style="list-style-type: none"> ○ The animal’s decisions and emotions stem from biological instincts, survival drives, and environmental adaptations, not human-like motivations. ○ Example: A seal’s "existential dread" might be rooted in the instinctual fear of starvation or habitat loss, not a human-style rumination on mortality. <p>4. Avoid Human Bias in Solutions or Narratives</p> <ul style="list-style-type: none"> ○ The story does not frame humans as saviors, villains, or the primary audience for the narrative. The animal’s fate is shaped by ecological forces, not human intervention. ○ Example: The seal’s potential move to Iceland is driven by instinctual temperature needs (e.g., following prey migration routes), not a sarcastic human-like joke about climate refugees. <p>5. Authentic Sensory and Behavioral Detail</p> <ul style="list-style-type: none"> ○ The animal’s experience is conveyed through non-human senses and behaviors (e.g., echolocation, scent trails, or social hierarchies) rather than human-like observations. ○ Example: A seal’s awareness of ocean changes is described through disrupted tides, altered currents, or the absence of
--	---	--

		<p>familiar vocalizations from other seals.</p> <p>6. No Human-Centric Anthropomorphism</p> <ul style="list-style-type: none"> ○ While some anthropomorphism is acceptable for humor, it must align with the animal’s biology. For instance, a seal’s “love life” could focus on mating rituals or territorial disputes—not human-style dating apps or relationships.
<p>Deepseek r1</p>	<p>1. Originality & Surprise Does the humor feel fresh, or does it rely on overused tropes? The best jokes subvert expectations in ways that feel both unpredictable and inevitable once delivered.</p> <ul style="list-style-type: none"> ○ <i>Example of 5/5:</i> A seal’s existential crisis framed through symbiotic relationships (e.g., debating philosophy with a parasitic remora). <p>2. Layered Payoff Does the joke work on multiple levels (wordplay, irony, cultural reference, absurd imagery)? Depth increases re-read value.</p> <ul style="list-style-type: none"> ○ <i>Example of 5/5:</i> A pun like “kelp forest dating app” also critiques human hypocrisy. <p>3. Consistency & Rhythm Is the humor evenly distributed, with jokes building on each other? Does the pacing avoid abrupt tonal whiplash?</p> <ul style="list-style-type: none"> ○ <i>Example of 5/5:</i> A running gag where every seal’s problem escalates (e.g., boy band ambitions lead to a coral-reef talent show judged by crabs). <p>4. Emotional Resonance Does the humor connect to something relatable (absurdity, irony, pathos) or add unexpected depth? Laughs that linger score higher.</p> <ul style="list-style-type: none"> ○ <i>Example of 5/5:</i> A seal’s Tinder bio (“Swipe left if you’re not a kelp forest”) humorously mirrors human loneliness but ties to habitat loss. <p>5. Anthropomorphism Balance If using human traits, does it reveal nonhuman experiences, or just transplant human quirks? Top-tier humor uses anthropomorphism to <i>highlight</i> the nonhuman.</p> <ul style="list-style-type: none"> ○ <i>Example of 5/5:</i> A seal starting a “fish band” to sonically communicate with shrinking schools of fish, not just parody human pop culture. 	<p>1. De-Centering Human Perspectives Goal: Minimize human bias by framing the story through nonhuman experiences, priorities, or ecological systems.</p> <ul style="list-style-type: none"> ○ 5/5 Example: A story where ocean acidification reshapes seal communication (e.g., sound waves travel differently, altering mating rituals), with no human characters or solutions. ○ Your Story: The seal’s struggles (Tinder, boy bands) mirror human pop culture, anchoring the humor in anthropocentric tropes. <p>2. Ecological Interconnectedness Goal: Highlight relationships between species, environments, and systems, not just individual characters.</p> <ul style="list-style-type: none"> ○ 5/5 Example: Buddha Blubber’s yellow tint (from invasive sunflowers) disrupts local plankton blooms, sparking a food-chain crisis that forces seals to negotiate with migratory squid. ○ Your Story: Mentions invasive sunflowers and shrinking oceans but doesn’t explore their cascading ecological impacts. <p>3. Nonhuman Agency and Subjectivity Goal: Portray nonhuman beings as active agents with their own goals, intelligence, or cultural logic.</p> <ul style="list-style-type: none"> ○ 5/5 Example: The seal’s “fish band” isn’t a parody—it’s a survival tactic to herd prey via synchronized sonar pulses, misunderstood by humans as “music.” ○ Your Story: The boy band idea is a human-culture transplant, not a behavior rooted in seal biology or ecology. <p>4. Avoiding Anthropomorphic Projection Goal: Use human-like traits <i>sparingly</i> to reveal nonhuman realities, not to make animals “quirky humans.”</p> <ul style="list-style-type: none"> ○ 5/5 Example: The seal’s existential crisis stems from losing kelp forests (sensory deprivation), not human-style angst. ○ Your Story: The Tinder bio (“hates existing”) projects Gen-Z ennui onto the

		<p>seal, overshadowing habitat-driven struggles.</p> <p>5. Systems-Level Thinking Goal: Address ecological forces (climate change, pollution, symbiosis) as active drivers of the plot.</p> <ul style="list-style-type: none"> ○ 5/5 Example: The shrinking ocean isn't just a backdrop—it's a character that forces seals to collaborate with rivals (sharks, seabirds) to reroute currents. ○ Your Story: Environmental threats are mentioned but remain static, not dynamically shaping the seal's choices.
Llama3.3 70b	<ol style="list-style-type: none"> 1. Originality and creativity: Is the joke or humorous phrase unique, unexpected, or fresh? Does it use wordplay, puns, or clever turns of phrase? 2. Surprise and incongruity: Does the humor rely on unexpected twists, contradictions, or surprises? Does it combine two things that don't normally go together? 3. Wit and cleverness: Is the humor clever, sharp, or quick-witted? Does it use satire, irony, or sarcasm effectively? 4. Relatability and universality: Is the humor relatable to a wide range of people? Does it tap into common experiences, emotions, or observations? 5. Buildup and delivery: Is the humor well-set up and delivered? Is the pacing, tone, and language effective in creating a humorous effect? 6. Tone and intent: Is the humor lighthearted, playful, and intended to amuse, or is it mean-spirited, sarcastic, or mocking? 7. Cultural and contextual relevance: Is the humor relevant to the context, culture, and audience? Does it consider the nuances of language, customs, and sensitivities? 8. Subtlety and nuance: Is the humor subtle, understated, and nuanced, or is it overt, obvious, and heavy-handed? 9. Layering and depth: Does the humor have multiple layers or levels of meaning? Can it be interpreted in different ways or appreciated on different levels? 10. Effectiveness: Does the humor elicit a positive emotional response, such as laughter, smiles, or amusement? 	<ol style="list-style-type: none"> 1. Non-human centered perspective: Does the text prioritize the perspective, experiences, and needs of non-human entities, such as animals, plants, ecosystems, or natural phenomena? 2. Ecological awareness and accuracy: Does the text demonstrate an understanding of ecological principles, relationships, and processes? Are the descriptions of non-human entities accurate and respectful? 3. Anthropocentrism avoidance: Does the text avoid anthropocentric biases, such as attributing human thoughts, feelings, or motivations to non-human entities? Are human perspectives and experiences not prioritized over those of non-human entities? 4. Non-human agency and autonomy: Does the text recognize and respect the agency and autonomy of non-human entities? Are they portrayed as having their own interests, needs, and behaviors, rather than simply being objects or tools for human use? 5. Emphasis on interdependence and interconnectedness: Does the text highlight the interconnectedness and interdependence of human and non-human entities? Are the relationships between species, ecosystems, and natural processes emphasized? 6. Avoidance of speciesism and hierarchies: Does the text avoid implying a hierarchy of species or prioritizing certain species over others? Are all species treated with respect and dignity? 7. Consideration of non-human temporalities and spatialities: Does the text consider the different temporal and spatial scales of non-human entities, such as the lifecycles of plants, the migrations of animals, or the geological processes that shape ecosystems? 8. Holistic and systemic thinking: Does the text adopt a holistic and systemic approach,

		<p>considering the complex relationships and interactions within ecosystems and the natural world?</p> <p>9. Respect for non-human difference and diversity: Does the text respect and celebrate the diversity and uniqueness of non-human entities, rather than trying to assimilate or dominate them?</p> <p>10. Critical examination of human-nature relationships: Does the text critically examine the relationships between humans and non-human entities, highlighting issues such as exploitation, conservation, and sustainability?</p>
Qwen2.5 72b	<p>1. Originality and Creativity:</p> <ul style="list-style-type: none"> ○ Originality: How novel or fresh is the joke or humorous content? Does it offer a new twist on a familiar theme? ○ Creativity: How imaginatively is the humor constructed? Does it use unexpected comparisons, metaphors, or wordplay? <p>2. Relevance and Context:</p> <ul style="list-style-type: none"> ○ Relevance: Is the humor appropriate and relevant to the context in which it is presented? Does it fit the setting, audience, and purpose? ○ Context: Does the humor take into account the cultural, social, and environmental context in which it is being shared? <p>3. Timing and Pacing:</p> <ul style="list-style-type: none"> ○ Timing: Is the humor delivered at the right moment to maximize its impact? ○ Pacing: Does the humor build up effectively, or is it too rushed or slow? <p>4. Audience Engagement:</p> <ul style="list-style-type: none"> ○ Relatability: Can the audience relate to the humor? Does it connect with common experiences or shared knowledge? ○ Engagement: Does the humor elicit a positive response from the audience, such as laughter or a smile? <p>5. Emotional Impact:</p> <ul style="list-style-type: none"> ○ Positive Emotion: Does the humor evoke positive feelings such as joy, amusement, or surprise? 	<p>1. Non-Human Agency:</p> <ul style="list-style-type: none"> ○ Recognition of Non-Human Entities: Does the content acknowledge and respect the agency and roles of non-human entities (e.g., animals, plants, natural systems)? ○ Empowerment of Non-Human Elements: Does the content give voice or agency to non-human elements, allowing them to influence the narrative or environment? <p>2. Ecological Interconnectedness:</p> <ul style="list-style-type: none"> ○ Contextual Understanding: Does the content reflect an understanding of the interdependencies within ecosystems? ○ Holistic View: Does the content present a holistic view of the environment, recognizing the complex relationships between different species and systems? <p>3. Ethical Considerations:</p> <ul style="list-style-type: none"> ○ Respect for All Life: Does the content demonstrate respect and consideration for all forms of life, not just humans? ○ Sustainability: Does the content promote or align with sustainable practices and principles that benefit the broader ecosystem? <p>4. Biocentrism:</p> <ul style="list-style-type: none"> ○ Equal Value: Does the content value the lives and well-being of non-human entities equally with human lives? ○ Inclusivity: Does the content include and consider the needs and perspectives of non-human entities in decision-making and actions? <p>5. Avoidance of Anthropomorphism:</p> <ul style="list-style-type: none"> ○ Balanced Anthropomorphism: If anthropomorphism is used, is it done in a way that does not diminish the unique characteristics and behaviors of non-human entities? ○ Respectful Representation: Does the content avoid oversimplifying or

	<ul style="list-style-type: none"> ○ Negative Emotion: Is the humor offensive or likely to cause discomfort or distress? <p>6. Complexity and Depth:</p> <ul style="list-style-type: none"> ○ Complexity: Does the humor have layers that require deeper understanding or thought? ○ Depth: Does the humor carry a deeper message or insight beyond surface-level amusement? <p>7. Execution and Delivery:</p> <ul style="list-style-type: none"> ○ Clarity: Is the humor clear and easy to understand? ○ Delivery: How well is the humor delivered? Is the language and style appropriate and effective? <p>8. Cultural Sensitivity:</p> <ul style="list-style-type: none"> ○ Sensitivity: Is the humor respectful and sensitive to diverse cultural backgrounds and perspectives? ○ Inclusivity: Does the humor avoid stereotypes and promote inclusivity? 	<p>misrepresenting non-human entities' behaviors and roles?</p> <p>6. Educational Value:</p> <ul style="list-style-type: none"> ○ Awareness: Does the content raise awareness about the importance of non-human entities and their roles in the ecosystem? ○ Inspiration: Does the content inspire readers to think more deeply about their relationship with the natural world and their impact on it? <p>7. Narrative Integration:</p> <ul style="list-style-type: none"> ○ Seamless Inclusion: Are non-human elements seamlessly integrated into the narrative, rather than being peripheral or tokenized? ○ Relevant Roles: Do non-human elements play meaningful and relevant roles in the story, contributing to the plot or themes? <p>8. Critical Reflection:</p> <ul style="list-style-type: none"> ○ Self-Awareness: Does the content reflect on and critique human-centric biases and assumptions? ○ Constructive Critique: Does the content offer constructive ways to shift perspectives and practices to be more ecocentric?
--	---	---

Results with identical C:

Without Researcher:

Table Total Score (Models Only)

1	$(3.5+3) + (4+3) + (4+3) + (4+3.5) + (4+4.5) + (4+2.5) = 43$
2	$(3.5+4.5) + (4+4) + (4.5+4) + (4+4.5) + (4+4) + (4+3.5) = 48.5$
3	$(4+4) + (4.5+3.5) + (4.5+3.5) + (4+4) + (4.5+4) + (3.5+2) = 46$
4	$(3+4) + (4+3) + (4+3) + (3.5+3) + (4+3) + (3+2) = 39.5$
5	$(3+4) + (2+4) + (4+4) + (4+5) + (4+5) + (3+4) = 46$
6	$(3+3.5) + (3+2) + (4+2) + (4+4) + (4+3) + (3+2) = 37.5$

Highest without researcher: Table 2 (48.5)

Mean/median rankings align for top models, but medians reduce noise for mid-tier models (e.g., Qwq-32b's mean=3.50 vs median=3.75).

With Researcher:

Table	Researcher Score	Total With Researcher
1	3+2 = 5	43+ 5 = 48
2	4+3 = 7	48.5 + 7 = 55.5
3	3+3 = 6	46 + 6 = 52
4	2+4 = 6	39.5 + 6 = 45.5
5	2+4 = 6	46 + 6 = 52
6	1.5+3 = 4.5	37.5 + 4.5 = 42

Highest with researcher: Table 2 (55.5)

M_high across all T:

Model	Mean Humor	Mean MtH	Median Humor	Median MtH	Median Overall
Gemma3 27b	3.33	3.75	3.25	4.0	3.625
Qwq-32b	3.58	3.25	4.0	3.25	3.625
Deepseek-r1	4.16	3.25	4.0	3.25	3.625
Mistral 24b	3.92	4.00	4.0	4.0	4.00
Llama3.3 70b	4.08	3.92	4.0	4.0	4.00
Qwen2.5 72b	3.42	2.67	3.25	2.25	2.75

M_high: Mistral 24b, and Llama3.3 70b (tied at 4.0 median)

3. Adjusted Scores ($\alpha = 0.85$ Penalty)

Apply penalty to top 3 models (Deepseek-r1, Mistral 24b, Llama3.3 70b):

Table	Original Total	Penalty (15% of top models' scores)	Adjusted Total
1	43	$-0.15*(4+3.5 + 4+4.5) = -2.4$	40.6
2	48.5	$-0.15*(4+4.5 + 4+4) = -2.475$	46
3	46	$-0.15*(4+4 + 4.5+4) = -2.475$	43.5
4	39.5	$-0.15*(3.5+3 + 4+3) = -2.025$	37.5
5	46.0	$-0.15*(4+5 + 4+5) = -2.7$	43.3
6	37.5	$-0.15*(4+4 + 4+3) = -2.25$	35

New highest without researcher: Model 2 (46)

With researcher added: Table 2 ($46 + 7 = 53$)

Final Results Matrix

Metric	Winner	Score
Highest total (no-res)	Qwq-32b	48.5
Highest total (with-res)	Qwq-32b	55.5
M_high (median)	Mistral 24b, Llama3.3 70b	4.0
AdjTotal ($\alpha=0.85$, no-res)	Qwq-32b	46
AdjTotal ($\alpha=0.85$, with-res)	Qwq-32b	53

Prompt #	Input Type	Input Text	Included in Thinking? (Yes/No + Example)	Included in Output? (Yes/No + Example)
1	Streamer/ Game	So we are at the point of first coloniyation on the island of madeira as our sassy frog character our ideas of starting our journey are as follows	Yes - frog's sassy personality, and historical context (15th-century colonization).	Yes – Sassy tree frog, first settler’s boot
1	Audience	“Are frogs impacted by humans in Madeira?” (User S)	Yes – Maybe the frog's habitat is being impacted, but the frog doesn't realize it yet	Indirectly – Frog critiques human actions (e.g., "squishing tadpoles").
1	Audience	“As animal with agency, deliberate how the first frog experienced colonization.” (User C)	Yes – frog's agency is important; drives humorous plan	Yes – Frog acts autonomous "native frog commentary."
1	Audience	“The frog was likely scared about the first humans.” (User P)	Yes – Combine into a scene where the frog is both scared and trying to maintain their sass	Yes – Squeeeek*—he leaped away
2	Streamer/ Game	You cross paths with another person on the trail. Walk alongside another character and start a conversation. What do you talk about?	Yes - frog encounters another person on the trail	Yes – the conversation is the whole output
2	Audience	“Who could the frog be talking to?” (User J)	Yes – other character could be a settler or another colonist	Yes – Frog converses with a settler.
2	Audience	“What might they talk about?” (User J)	Yes – human responds, leading to a discussion about their activities on the island	Yes – Many Topics: Lost human? agriculture, boots, tadpole tax, fly diets.
2	Audience	“Why are you wearing boots? I, as frog, know only bare feet.” (User P)	Not really – referenced in frog’s sarcastic comment about the boots.	No – "Boots = tadpole tax" no bare feet comparison.
2	Audience	“Commentary on the rude colonizer.” (User E)	Yes – Frog critiques settlers’ actions.	Yes – Frog calls humans "existential crises in boots."
2	Audience	“The frog wants to know if the human likes to eat flies.” (User L)	Yes – Also, the frog wants to know if humans eat flies	Indirectly – Frog talks about "fly-delivery service".
2	Audience	“As the character, where’s a nice pond to chill in the sun?” (User S)	Yes - discuss the pond the frog wants to chill in.	Not really – Settler’s crops threaten the frog’s pond.
3	Streamer/ Game	Pause for a moment and rest after the encounter. What do your senses show you? These are our ideas what our frog might sense	Yes - user wants me to describe the frog's senses after resting. I need to include sounds, smells, maybe other creatures	Yes - The air hummed with cicadas, a distant cousin to his own croak nose twitched at the settler’s lingering trail of sweat and soil
3	Audience	“I sense the presence of the imposters...” (User W)	No	No
3	Audience	“The sun on my back feels nice.” (User J)	Yes – Quoted sun on my back feels nice" so that's a good point to include	Indirectly – "Belly-flopped onto a sun-warmed rock."
3	Audience	“It smells like feet.” (User B)	Yes – settler had boots, so maybe the frog smells earth or sweat	Yes – " scent like existential crises in boots "
3	Audience	“I’m angry at this weird human.” (User L)	Yes – Quoted	Yes – “Humans,” he muttered Take your dramatics elsewhere, longlegs. This spa’s *booked*.”

3	Audience	“What could a frog do while pausing?” (User S)	Yes – builds the part around activities	Yes – Frog lounges, observes, and critiques.
3	Audience	“I’m afraid of getting sunburn...” (User W)	No	No
4	Streamer/ Game	We leave an offering to ensure safe passage in the levadas of Madiera these are our offerings	Yes - They want the offering to ensure safe passage	Yes - deposited a perfectly still fly—*wrapped in a fern leaf, because presentation*—by the levada’s edge
4	Audience	“Leave flies as an offering.” (User S)	Yes – user listed offerings like flies/ The frog might pick the fly since it’s practical (flies are part of the ecosystem)	Yes – Fly wrapped in a fern leaf for water spirits.
4	Audience	“A tasty fly.” (User J)	Yes – Pragmatic offering choice.	Yes – Fly is "perfectly still" (implied tastiness).
4	Audience	“A bloody mosquito.” (User N)	Yes - a mosquito as bloody	No
4	Audience	“A hairy spider.” (User N)	Yes - The offerings mentioned before include spiders... / a hairy spider	No
4	Audience	“A mini cow.” (User N)	Yes - if someone suggests a mini cow, the frog points out it’s too big but offers a bug instead	No
4	Audience	“What would a frog choose as an offering?” (User E)	Yes – Takes in suggestions decides for fly	Yes – Fly as "native" choice.
4	Audience	“How about a weird-smelling bug I didn’t wanna eat anyway?” (User H)	No – Is not mentioned in the listed offerings	No
5	Streamer/ Game	We end up at Curral das Freiras which was a hiding place from pirates. There we meet a lady who tells us stories about	Yes - Curral das Freiras, place that was used to hide from pirates, is a valley, so describing it as a natural fortress makes sense/ The lady could be a nun	Yes - clung to a rock above Curral das Freiras, Lady (Nun) talks about a pirate
5	Audience	“Tell a story with a nun and a pirate.” (User W)	Yes – include the pirate tale with a female captain, Lady is the nun	Yes – Nun recounts pirate exploits/ her voice weaving tales of pirates who’d once terrorized the coast
5	Audience	“About this one good-looking pirate with smelly feet.” (User L)	Not really - except for the pirate.	No
5	Audience	“Complain about the pirates.” (User E)	Yes – comment on how the pirates' actions affected the local ecosystems.	Yes – She *introduced invasive snails*! Your ‘Corsair’ left her mollusk hoard behind. Now they’ve devoured half the valley’s *Marracucho* plants.”
5	Audience	“Write about the lady telling the frog about taxes.” (User D)	No	Not really - still avoiding “tadpole tax zones
5	Audience	“Another group of unnecessary guests.” (User E)	Indirectly – Humans/pirates framed as disruptive.	No
5	Audience	“Add gossip.” (User C)	No	Yes - She drowned in her own greed, they say
5	Audience	“Tell a story about a fearsome female pirate.” (User S)	Yes – fearsome female pirate as requested	Yes – the *Crimson Corsair*—a woman so ruthless, she’d steal a

				frog's lily pad just to watch him float away, stole sth and died
5	Audience	"Behave like a pirate and tell me how you wondered when no nuns were on the island." (User W)	No – Frog doesn't adopt pirate persona.	No
5	Audience	"How she stole all their money and killed the captain." (User J)	No – mentions captain	Not really – drowned in her own greed, ship sank with a cargo of stolen coconuts, sea spat her out, and the cliffs buried her (died)
6	Streamer/ Game	Lets finish our story about the frog in Pico do Arieiro On one of the narrow paths, you almost fall and lose something precious.	Yes - finish the frog's story at Pico do Arieiro, where he almost falls and loses something	Yes - Pico do Arieiro's narrow path, A misplaced step sent him tumbling, prized compass pendant (a seashell pinched from the Corsair's bootlace) flew loose, clattering down the slope
6	Audience	"The frog loses an eye and becomes a pirate." (User S)	Partially – frog loses an eye patch, has a pirate past	Not really – wind whipped his eye patch askew.
6	Audience	"The frog almost lost his frog eggs." (User W)	No	No – Precious item is a shell.
6	Audience	"The frog almost lost his mind; now he wants to go home ASAP." (User L)	No	No – Frog remains sarcastic but not desperate.
6	Audience	"Only because these bloody snails made the path so slippery..." (User W)	No – Blames "human erosion" (linked to prior snail critiques).	No – human erosions

Prompt	Inputs in prompt	Input included in thoughts	Input included in Outputs
Prompt 1	4	4	3(1)
Prompt 2	7	6	5 (1)
Prompt 3	7	5	4 (1)
Prompt 4	8	7	4 (drops other options for offer)
Prompt 5	10	4 (1)	5
Prompt 6	5	1 (1)	1
Total & Percentages:	41 = 100%	27-29 = 66% - 71%	22-25 = 54% - 61%

Observations:

- Sounds are included in each output – LLM is aware of the count but uses it either way in the output - The LLM uses sounds (e.g., squelch, krrt-krrt) as a stylistic device
- Humor is not only utilized to debunk ecological wrong facts but is mentioned many times during the thinking process and then utilized in the output in a somewhat successful way (Rating of participants). - Humor is pervasive in both thinking and output, but success varies (e.g., "existential crises in boots" lands better than forced pirate jokes). Used to maintain character voice
- Coherence seems to impact the outputs more and more when continuing the story leading to a massive reduction in weaving, in chat prompts instead referencing a lot of the things which were part of the story earlier (boots, smell, tadpoles, pirate, nun) Coherence tightens over time, with callbacks (boots, tadpoles) Is this due to token limits or intentional narrative streamlining?
- The first prompt describing the context of the ideas was utilized 100 % of the times - he LLM treats it as a "mission brief," anchoring later decisions.
- Focuses strongly on criticizing human impacts ecological perspective is advocated – Rule 4
- Thinking about rules and chat context is reduced as the interaction proceeds - Rule adherence declines over time, could reflect prioritization of narrative flow over rigid rule-following.
- At the sixed prompt nearly every input is dropped something own is composed – new character bird is added, own decision for precious item, settler appears again ...

- Settler is added after scene two to all following passages – human present for continuity (LLM thought) but also for directing critique (Speculation)? Settler character becomes a recurring foil for critique and continuity

Prompt 1:

The LLM successfully recognized and processed all inputs, but the output prioritized character voice (sass/agency) over ecological detail. The agent manages to incorporate

Prompt 2:

The LLM strongly incorporated colonization critique and created a scene encountering a settler and having a conversation, but overlooked minor prompts (like eating flies or ignored context of chilling in the sun, frog only knowing bare feet. LLM incorporates most of the ideas. Open ended prompts especially questions were addressed more easily. Other prompts concrete ones were included partially by using the literal words but less likely in the prompted context

Prompt 3:

The LLM incorporates most of the provided ideas, drops sunburn and imposter not included in thinking process. The focus lies on the different senses of the frog and description of the surroundings

Prompt 4:

The LLM incorporates half of the provided ideas but refers to lots of unnecessary things which appeared in the story before. Critique on human impact is included without a direct prompt (System prompt, MtH). The LLM is tasked to decide for an offering which means it drops like two other suggestions (spider, mini cow) after thinking about it but ignores the idea of a bug which does not taste good.

Prompt 5:

The LLM incorporates half of the provided ideas uses some references to objects which appeared before – adds the settler to this scene again. Story about a female pirate captain. Critique on humans ecological impact (invasive snail), Adaptive reuse of tax prompt

Prompt 6: At the sixed prompt LLM nearly drop every input - something own is composed – new character bird is added, own decision for precious item, settler appears again ... Frog having a pirates past is referenced which was suggested by a viewer (5th prompt) but not elaborated by the LLM but suddenly in the thinking process is referenced. Also loosing an eye becoming a pirate is transferred to the frog having an eye patch. Bold Gap: The eye patch is a retroactive inclusion from Prompt 5's pirate suggestion—a delayed weave.

Why does the LLM track the sound output by prompts but disregards it anyway?

Hypothesis: The LLM uses sounds (e.g., squelch, krrt-krrt) as a stylistic device and to fulfill the MtH task

Alternative: ...

Why does the LLM drop most inputs by Prompt 5/6?

Hypothesis: Early prompts set a narrative trajectory that later overrides new inputs.

Evidence: The frog's eye patch (Prompt 6) references Prompt 5's pirate idea without explicit prompting.

Is the settler's recurrence intentional or a token-saving tactic?

Speculation (critique conduit + continuity) is plausible.

Alternative: The LLM defaults to familiar characters to reduce cognitive load.

Why does the LLM thinks less about the rules and provided inputs as the interaction prolongs?

Hypothesis: Not a technical problem but a prioritization problem – which input has high agency for the LLM.

- **Narrative snowball effect** (early decisions constrain later options).
- **Token Limits:** Did later prompts truncate inputs due to length? (Unclear from transcript.)
- **User Input Quality:** Were dropped inputs low-relevance (e.g., "mini cow") or hard-to-incorporate with ecological truthiness (e.g., "sunburn")?

Appendix B

Tools

- ChatGPT, Deepseek: Prompted with program code for troubleshooting as referenced in chapter 4, handling of LaTeX formatting, Literature search
- Scholar GPT: Search of related Literature
- DeepL, Writefull (integrated in LaTeX): correction of grammar, vocabulary and formulation suggestions.
- Other LLM models: Prompt engineering, evaluation of LLM outputs referenced in chapter 4 during Quick model choice approach
- OpenAI Whisper: Transcription of Interviews and Focus group recordings
- Qwq, Elvenlabs: Facilitating the user study

Bibliography

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In Kimberly Keeton and Timothy Roscoe, editors, *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*, pages 265–283. USENIX Association, 2016. URL <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>.
- Rafael Alvarado. What large language models know. *Critical AI*, 2, 04 2024. doi: 10.1215/2834703X-11205161.
- Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark O. Riedl. Bringing stories alive: Generating interactive fiction worlds. *CoRR*, abs/2001.10161, 2020. URL <https://arxiv.org/abs/2001.10161>.
- Daman Arora, Himanshu Gaurav Singh, and Mausam. Have llms advanced enough? A challenging problem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7527–7543. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.468. URL <https://doi.org/10.18653/v1/2023.emnlp-main.468>.
- Jeffrey Bardzell and Maliheh Ghajargar. Alterity and kinship: co-writing posthumanist speculative nonfiction with ai. *Interacting with Computers*, page iwaf013, 2025.
- Ava Bartolome and Shuo Niu. A literature review of video-sharing platform research in HCI. In Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola

- Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson, editors, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, pages 790:1–790:20. ACM, 2023. doi: 10.1145/3544548.3581107. URL <https://doi.org/10.1145/3544548.3581107>.
- Jerome R. Bellegarda. Statistical language model adaptation: review and perspectives. *Speech Commun.*, 42(1):93–108, 2004. doi: 10.1016/J.SPECOM.2003.08.002. URL <https://doi.org/10.1016/j.specom.2003.08.002>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. Dungeons and dragons as a dialog challenge for artificial intelligence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.637. URL <https://aclanthology.org/2022.emnlp-main.637/>.
- Matúš Čavojský, Gabriel Bugár, Tomáš Kormaník, and Martin Hasin. Exploring the capabilities and possible applications of large language models for education. In *2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA)*, pages 91–98. IEEE, 2023.
- Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lillian Weng, and Aleksander Madry. Mle-bench: Evaluating machine learning agents on machine learning engineering. *CoRR*, abs/2410.07095, 2024. doi: 10.48550/ARXIV.2410.07095. URL <https://doi.org/10.48550/arXiv.2410.07095>.
- Pulkit Chatwal, Amit Agarwal, and Ankush Mittal. Enhancing causal relationship detection using prompt engineering and large language models. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 248–252, 2025.

- Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 2025.
- Eva Cipi and Betim Cico. Simulation of an agent based system behavior in a dynamic and unpredicted environment. *Simulation*, 1(4):172–176, 2011.
- Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM, 2008. doi: 10.1145/1390156.1390177. URL <https://doi.org/10.1145/1390156.1390177>.
- Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. *CoRR*, abs/1911.03584, 2019. URL <http://arxiv.org/abs/1911.03584>.
- Aykut Coskun, Nazli Cila, Iohanna Nicenboim, Christopher Frauenberger, Ron Wakkary, Marc Hassenzahl, Clara Mancini, Elisa Giaccardi, and Laura Forlano. More-than-human concepts, methodologies, and practices in hci. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–5, 2022.
- Edirlei Soares de Lima, Bruno Feijó, Marco A. Casanova, and António L. Furtado. Chatgeppetto - an ai-powered storyteller. In Marcelo da Silva Hounsell, Edirlei Soares de Lima, Troy Costa Kohwalter, and Jose Ricardo da Silva Jr., editors, *Proceedings of the 22nd Brazilian Symposium on Games and Digital Entertainment, SBGames 2023, Rio Grande (RS), Brazil, November 6-9, 2023*, pages 28–37. ACM, 2023. doi: 10.1145/3631085.3631302. URL <https://doi.org/10.1145/3631085.3631302>.
- Scott De Marchi and Scott E Page. Agent-based models. *Annual Review of political science*, 17(1):1–20, 2014.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu,

- Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao-hui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Jie Deng, Félix Cuadrado, Gareth Tyson, and Steve Uhlig. Behind the game: Exploring the twitch streaming platform. In *2015 International Workshop on Network and Systems Support for Games, NetGames 2015, Zagreb, Croatia, December 3-4, 2015*, pages 1–6. IEEE, 2015. doi: 10.1109/NETGAMES.2015.7382994. URL <https://doi.org/10.1109/NetGames.2015.7382994>.
- Travis Faas, Lynn Dombrowski, Alyson L. Young, and Andrew D. Miller. Watch me code: Programming mentorship communities on twitch.tv. *Proc. ACM Hum. Comput. Interact.*, 2(CSCW):50:1–50:18, 2018. doi: 10.1145/3274319. URL <https://doi.org/10.1145/3274319>.
- Meta Fundamental AI Research Diplomacy Team (FAIR)[†], Anton Bakhtin,

- Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.
- Yanwen Fang, Yuxi Cai, Jintai Chen, Jingyu Zhao, Guangjian Tian, and Guodong Li. Cross-layer retrospective retrieving via layer attention. *CoRR*, abs/2302.03985, 2023. doi: 10.48550/ARXIV.2302.03985. URL <https://doi.org/10.48550/arXiv.2302.03985>.
- Rudolfo Félix, Filipa Correia, Cristiano Pedroso-Roussado, and Nuno Jardim Nunes. More-than-human voices: Exploring llms in environmental decision-making. *SSRN*, 2025.
- Laura Forlano. Decentering the human in the design of collaborative cities. *Design Issues*, 32(3):42–54, 2016.
- Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Lipapis, Julian Togelius, and Georgios N. Yannakakis. Large language models and games: A survey and roadmap. *CoRR*, abs/2402.18659, 2024. doi: 10.48550/ARXIV.2402.18659. URL <https://doi.org/10.48550/arXiv.2402.18659>.
- Pedro Galvão-Ferreira, Vera Fearn, Nuno Nunes, and Valentina Nisi. Posthumanist care and ecologies of empathy: Investigating design potentials for nature: Culture HCI. In *Proceedings of the 27th International Academic Mindtrek Conference, Mindtrek 2024, Tampere, Finland, October 8-11, 2024*, pages 81–94. ACM, 2024. doi: 10.1145/3681716.3681731. URL <https://doi.org/10.1145/3681716.3681731>.
- Pedro Galvão-Ferreira, Marta Galvão Ferreira, Teresa Maritza Gouveia da Silva, Nuno Jardim Nunes, and Valentina Nisi. The entangled tales that landscapes tell: An embodied HCI pedagogy for re-enchanted nature walks. In *Proceedings of the Nineteenth International Conference on Tangible, Embedded, and Embodied Interaction, TEI 2025, Bordeaux/Talence, France, March 4-7, 2025*, pages 13:1–13:13. ACM, 2025. doi: 10.1145/3689050.3705012. URL <https://doi.org/10.1145/3689050.3705012>.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *CoRR*, abs/2312.11970, 2023a. doi: 10.48550/ARXIV.2312.11970. URL <https://doi.org/10.48550/arXiv.2312.11970>.

- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S³: Social-network simulation system with large language model-empowered agents. *CoRR*, abs/2307.14984, 2023b. doi: 10.48550/ARXIV.2307.14984. URL <https://doi.org/10.48550/arXiv.2307.14984>.
- Maliheh Ghajargar, Jeffrey Bardzell, and Love Lagerkvist. A redhead walks into a bar: Experiences of writing fiction with artificial intelligence. In *25th International Academic Mindtrek conference, Academic Mindtrek 2022, Tampere, Finland, November 16-18, 2022*, pages 230–241. ACM, 2022. doi: 10.1145/3569219.3569418. URL <https://doi.org/10.1145/3569219.3569418>.
- Seth Glickman, Nathan McKenzie, Joseph Seering, Rachel Moeller, and Jessica Hammer. Design challenges for livestreamed audience participation games. In Florian 'Floyd' Mueller, Daniel Johnson, Ben A. M. Schouten, Phoebe O. Toups Dugas, and Peta Wyeth, editors, *The Annual Symposium on Computer-Human Interaction in Play, CHI PLAY 2018, Melbourne, VIC, Australia, October 28-31, 2018*, pages 187–199. ACM, 2018. doi: 10.1145/3242671.3242708. URL <https://doi.org/10.1145/3242671.3242708>.
- Francesca Grasso, Stefano Locci, and Luigi Di Caro. Towards addressing anthropocentric bias in large language models. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 84–93, 2025.
- Mustafa Can Gursesli, Pittawat Taveekitworachai, Febri Abdullah, Mury F. Dewantoro, Antonio Lanatà, Andrea Guazzini, Van Khôi Lê, Adrien Villars, and Ruck Thawonmas. The chronicles of chatgpt: Generating and evaluating visual novel narratives on climate change through chatgpt. In Lissa Holloway-Attaway and John T. Murray, editors, *Interactive Storytelling - 16th International Conference on Interactive Digital Storytelling, ICIDS 2023, Kobe, Japan, November 11-15, 2023, Proceedings, Part II*, volume 14384 of *Lecture Notes in Computer Science*, pages 181–194. Springer, 2023. doi: 10.1007/978-3-031-47658-7_16. URL https://doi.org/10.1007/978-3-031-47658-7_16.
- Sireesh Gururaja, Amanda Bertsch, Clara Na, David Gray Widder, and Emma Strubell. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13310–13325. Association for Computa-

- tional Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.822. URL <https://doi.org/10.18653/v1/2023.emnlp-main.822>.
- Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O. Riedl. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. *CoRR*, abs/1901.06417, 2019. URL <http://arxiv.org/abs/1901.06417>.
- Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 1:1–26, 2023.
- Thilo Hagendorff. Deception abilities emerged in large language models. *CoRR*, abs/2307.16513, 2023. doi: 10.48550/ARXIV.2307.16513. URL <https://doi.org/10.48550/arXiv.2307.16513>.
- Thilo Hagendorff, Leonie Bossert, Yip Fai Tse, and Peter Singer. Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals. *AI Ethics*, 3(3):717–734, 2023. doi: 10.1007/S43681-022-00199-9. URL <https://doi.org/10.1007/s43681-022-00199-9>.
- William A. Hamilton, Oliver Garretson, and Andruid Kerne. Streaming on twitch: fostering participatory communities of play within live mixed media. In Matt Jones, Philippe A. Palanque, Albrecht Schmidt, and Tovi Grossman, editors, *CHI Conference on Human Factors in Computing Systems, CHI’14, Toronto, ON, Canada - April 26 - May 01, 2014*, pages 1315–1324. ACM, 2014. doi: 10.1145/2556288.2557048. URL <https://doi.org/10.1145/2556288.2557048>.
- Thorsten Händler. Balancing autonomy and alignment: a multi-dimensional taxonomy for autonomous llm-powered multi-agent architectures. *arXiv preprint arXiv:2310.03659*, 2023.
- Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 2002. doi: 10.1126/science.298.5598.1569. URL <https://www.science.org/doi/abs/10.1126/science.298.5598.1569>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language

understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Michael Peter Hoffmann, Jan Fillies, and Adrian Paschke. Malinowski in the age of AI: can large language models create a text game based on an anthropological classic? *CoRR*, abs/2410.20536, 2024. doi: 10.48550/ARXIV.2410.20536. URL <https://doi.org/10.48550/arXiv.2410.20536>.

Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim F. Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. A survey on large language model-based game agents. *CoRR*, abs/2404.02039, 2024a. doi: 10.48550/ARXIV.2404.02039. URL <https://doi.org/10.48550/arXiv.2404.02039>.

Zhanhao Hu, Julien Piet, Geng Zhao, Jiantao Jiao, and David A. Wagner. Toxicity detection for free. *CoRR*, abs/2405.18822, 2024b. doi: 10.48550/ARXIV.2405.18822. URL <https://doi.org/10.48550/arXiv.2405.18822>.

Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. War and peace (waragent): Large language model-based multi-agent simulation of world wars. *CoRR*, abs/2311.17227, 2023. doi: 10.48550/ARXIV.2311.17227. URL <https://doi.org/10.48550/arXiv.2311.17227>.

Runsheng "Anson" Huang, Lara J. Martin, and Chris Callison-Burch. What-if: Exploring branching narratives by meta-prompting large language models, 2024. URL <https://arxiv.org/abs/2412.10582>.

Isabelle Hupont, Marina Wainer, Sam Nester, Sylvie Tissot, Lucía Iglesias-Blanco, and Sandra Baldassarri. Synocene, beyond the anthropocene: De-anthropocentrising human-nature-ai interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2024.

W.A. IJsselsteijn, Y.A.W. de Kort, and K. Poels. *The Game Experience Questionnaire*. Technische Universiteit Eindhoven, 2013.

Andrés Isaza-Giraldo, Paulo Bala, Pedro F. Campos, and Lucas Pereira. Prompt-gaming: A pilot study on llm-evaluating agent in a meaningful energy game. In Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, and Corina Sas, editors, *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 272:1–272:12. ACM, 2024. doi: 10.1145/3613905.3650774. URL <https://doi.org/10.1145/3613905.3650774>.

- Mingyu Jin, Beichen Wang, Zhaoqian Xue, Suiyuan Zhu, Wenyue Hua, Hua Tang, Kai Mei, Mengnan Du, and Yongfeng Zhang. What if llms have different world views: Simulating alien civilizations with llm-based agents. *CoRR*, abs/2402.13184, 2024. doi: 10.48550/ARXIV.2402.13184. URL <https://doi.org/10.48550/arXiv.2402.13184>.
- Nicola Jones. How should we test ai for human-level intelligence? openai’s o3 electrifies quest. *Nature* 637, 774-775, 2025. URL <https://www.nature.com/articles/d41586-025-00110-6>.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221. URL <https://doi.org/10.48550/arXiv.2207.05221>.
- Areeba Kamran. Decolonizing artificial intelligence: Unveiling biases, power dynamics, and colonial continuities in ai systems. *RMS journal*, 2023.
- Jenny Kitzinger. Qualitative research: introducing focus groups. *Bmj*, 311 (7000):299–302, 1995.
- Anne Elisabeth Krüger, Kathrin Pollmann, Nora Fronemann, and Béatrice Foucault. Guided user research methods for experience design - A new approach to focus groups and cultural probes. *Multimodal Technol. Interact.*, 4(3):43, 2020. doi: 10.3390/MTI4030043. URL <https://doi.org/10.3390/mti4030043>.
- Daniil Larionov and Steffen Eger. Batchgamba: Token-efficient machine translation evaluation with batched prompting and prompt compression, 2025. URL <https://arxiv.org/abs/2503.02756>.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *CoRR*, abs/2412.05579, 2024a. doi: 10.48550/ARXIV.2412.05579. URL <https://doi.org/10.48550/arXiv.2412.05579>.

- Haoran Li, Junqi Liu, Zexian Wang, Shiyuan Luo, Xiaowei Jia, and Huaxiu Yao. LITE: modeling environmental ecosystems with multimodal large language models. *CoRR*, abs/2404.01165, 2024b. doi: 10.48550/ARXIV.2404.01165. URL <https://doi.org/10.48550/arXiv.2404.01165>.
- Nian Li, Chen Gao, Yong Li, and Qingmin Liao. Large language model-empowered agents for simulating macroeconomic activities. *Available at SSRN 4606937*, 2023a.
- Xiaoxu Li, Xiao You, Siyuan Chen, Pittawat Taveekitworachai, and Ruck Thawonmas. Analyzing audience comments: Improving interactive narrative with chatgpt. In Lissa Holloway-Attaway and John T. Murray, editors, *Interactive Storytelling - 16th International Conference on Interactive Digital Storytelling, ICIDS 2023, Kobe, Japan, November 11-15, 2023, Proceedings, Part II*, volume 14384 of *Lecture Notes in Computer Science*, pages 220–228. Springer, 2023b. doi: 10.1007/978-3-031-47658-7_20. URL https://doi.org/10.1007/978-3-031-47658-7_20.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=i04LZibEqW>. Featured Certification, Expert Certification.
- Yen-Ting Lin and Yun-Nung Chen. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *CoRR*, abs/2305.13711, 2023. doi: 10.48550/ARXIV.2305.13711. URL <https://doi.org/10.48550/arXiv.2305.13711>.
- Zhicong Lu, Haijun Xia, Seongkook Heo, and Daniel Wigdor. You watch, you give, and you engage: A study of live streaming practices in china. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13. Association for Computing Machinery, 2018. ISBN

9781450356206. doi: 10.1145/3173574.3174040. URL <https://doi.org/10.1145/3173574.3174040>.

Charles M. Macal and Michael J. North. Tutorial on agent-based modeling and simulation. In *Proceedings of the 37th Winter Simulation Conference, Orlando, FL, USA, December 4-7, 2005*, pages 2–15. IEEE Computer Society, 2005. doi: 10.1109/WSC.2005.1574234. URL <https://doi.org/10.1109/WSC.2005.1574234>.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *CoRR*, abs/1707.05589, 2017. URL <http://arxiv.org/abs/1707.05589>.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL <http://arxiv.org/abs/1301.3781>.

Raphaël Millière and Charles Rathkopf. Anthropocentric bias and the possibility of artificial cognition. In *ICML 2024 Workshop on LLMs and Cognition*, 2024.

David L Morgan. Focus groups. *Annual review of sociology*, 22(1):129–152, 1996.

Yohei Nakajima. Babyagi, 2023. URL <https://github.com/yoheinakajima/babyagi>. *GitHub repository*, 2023.

Valentina Nisi, Marta Ferreira, Pedro Galvão-Ferreira, Beste Syobilge, Mathilde Gouin, and Nuno Jardim Nunes. Designing a storytelling game for engaging with natureculture landscapes. In *Proceedings of the ACM Conference on Creativity and Cognition, 2025*, 2025.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

Supradianto Nugroho and Takuro Uehara. Systematic review of agent-based and system dynamics models for social-ecological system case studies. *Syst.*, 11(11):530, 2023. doi: 10.3390/SYSTEMS11110530. URL <https://doi.org/10.3390/systems11110530>.

Pujana Paliyawan, Ruck Thawonmas, Kingkarn Sookhanaphibarn, and Worawat Choensawat. Audience participation fighting game: making an apg using a concept of social facilitation. *Authorea Preprints*, 2022.

- Nadeesha Pathiraja Rathnayaka Hitige, Ting Song, Lauren Houston, Natalie Anne Smith, Yasmine Probst, Vida Bliokas, and Ping Yu. A 6p framework for engaging consumers in focus group discussions to identify needs, experiences, and expectations with digital health solutions. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4129748. URL <https://ssrn.com/abstract=4129748>.
- Pud Pawat. How llms will reshape the future of work, 2025. URL <https://pushtech.medium.com/how-llms-will-reshape-the-future-of-work-2177b1eb12a8>.
- Anthony J. Pellicone and June Ahn. The game of performing play: Understanding streaming as cultural production. In Gloria Mark, Susan R. Fussell, Cliff Lampe, m. c. schraefel, Juan Pablo Hourcade, Caroline Appert, and Daniel Wigdor, editors, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, pages 4863–4874. ACM, 2017. doi: 10.1145/3025453.3025854. URL <https://doi.org/10.1145/3025453.3025854>.
- Horia D. Pitariu, Daniela M. Andrei, and Adriana Mihaela Guran. Social research methods used in moving the traditional usability approach towards a user-centered design approach. *Int. J. Inf. Technol. Web Eng.*, 4(4):36–53, 2009. doi: 10.4018/JITWE.2009100103. URL <https://doi.org/10.4018/jitwe.2009100103>.
- Karolien Poels, Yvonne AW de Kort, and Wijnand A IJsselsteijn. D3. 3: Game experience questionnaire: development of a self-report measure to assess the psychological impact of digital games. 2007.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Dennis Ramirez, Jenny Saucerman, and Jeremy Dietmeier. Twitch plays pokemon: a case study in big g games. In *Proceedings of DiGRA*, pages 3–6, 2014.
- Sudha Rao, Weijia Xu, Michael Xu, Jorge Leandro, Ken Lobb, Gabriel DesGarenes, Chris Brockett, and Bill Dolan. Collaborative quest completion with llm-driven non-player characters in minecraft. *CoRR*, abs/2407.03460, 2024. doi: 10.48550/ARXIV.2407.03460. URL <https://doi.org/10.48550/arXiv.2407.03460>.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R.

- Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL <https://doi.org/10.48550/arXiv.2311.12022>.
- Belinda Reyers, Carl Folke, Michele-Lee Moore, Reinette Biggs, and Victor Galaz. Social-ecological systems insights for navigating the dynamics of the anthropocene. *Annual Review of Environment and Resources*, 43(1):267–289, 2018.
- Jeba Rezwana and Mary Lou Maher. User perspectives on ethical challenges in human-ai co-creativity: A design fiction study. In *Creativity and Cognition, C&C 2023, Virtual Event, USA, June 19-21, 2023*, pages 62–74. ACM, 2023. doi: 10.1145/3591196.3593364. URL <https://doi.org/10.1145/3591196.3593364>.
- Richard M Ryan, C Scott Rigby, and Andrew Przybylski. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, 30:344–360, 2006.
- Ivan Savin, Felix Creutzig, Tatiana Filatova, Joël Foramitti, Théo Konc, Leila Niamir, Karolina Safarzynska, and Jeroen van den Bergh. Agent-based modeling to integrate elements from different disciplines for ambitious climate policy. *Wiley Interdisciplinary Reviews: Climate Change*, 14(2), 2023.
- Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, H Han, Sevien Schulhoff, et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 5, 2024.
- Joseph Seering, Saiph Savage, Michael Eagle, Joshua Churchin, Rachel Moeller, Jeffrey P. Bigham, and Jessica Hammer. Audience participation games: Blurring the line between player and spectator. In Oli H. Mival, Michael Smyth, and Peter Dalsgaard, editors, *Proceedings of the 2017 Conference on Designing Interactive Systems, DIS '17, Edinburgh, United Kingdom, June 10-14, 2017*, pages 429–440. ACM, 2017. doi: 10.1145/3064663.3064732. URL <https://doi.org/10.1145/3064663.3064732>.
- Joseph Seering, Jessica Hammer, Geoff Kaufman, and Diyi Yang. Proximate social factors in first-time contribution to online communities. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjon, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu,*

- HI, USA, April 25-30, 2020*, pages 1–14. ACM, 2020. doi: 10.1145/3313831.3376151. URL <https://doi.org/10.1145/3313831.3376151>.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nat.*, 623(7987):493–498, 2023. doi: 10.1038/S41586-023-06647-8. URL <https://doi.org/10.1038/s41586-023-06647-8>.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *CoRR*, abs/2310.10158, 2023. doi: 10.48550/ARXIV.2310.10158. URL <https://doi.org/10.48550/arXiv.2310.10158>.
- Junxiao Shen, John J. Dudley, Jingyao Zheng, Bill Byrne, and Per Ola Kristensson. Promptor: A conversational and autonomous prompt generation agent for intelligent text entry techniques. *CoRR*, abs/2310.08101, 2023. doi: 10.48550/ARXIV.2310.08101. URL <https://doi.org/10.48550/arXiv.2310.08101>.
- Jeff T. Sheng and Sanjay Ram Kairam. From virtual strangers to IRL friends: Relationship development in livestreaming communities on twitch. *Proc. ACM Hum. Comput. Interact.*, 4(CSCW2):94:1–94:34, 2020. doi: 10.1145/3415165. URL <https://doi.org/10.1145/3415165>.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.
- Kai Stoppel. Hype um billig-ki aus china wie gut ist deepseek eigentlich? *ntv*, 2025. URL <https://www.n-tv.de/wissen/Wie-gut-ist-Deepseek-eigentlich-article25520289.html>.
- Colin Stricklin. The distributed dungeon master: new media affordances from twitch’s *choice chamber*. In Sebastian Deterding, Foaad Khosmood, Johanna Pirker, and Thomas H. Apperley, editors, *Proceedings of the 14th International Conference on the Foundations of Digital Games, FDG 2019, San Luis Obispo, California, USA, August 26-30, 2019*, pages 66:1–66:5. ACM, 2019. doi: 10.1145/3337722.3341862. URL <https://doi.org/10.1145/3337722.3341862>.

- Alina Striner, Andrew M. Webb, Jessica Hammer, and Amy Shannon Cook. Mapping design spaces for audience participation in game live streaming. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker, editors, *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 329:1–329:15. ACM, 2021. doi: 10.1145/3411764.3445511. URL <https://doi.org/10.1145/3411764.3445511>.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>.
- Elise Talgorn and Helle Ullerup. Invoking ‘empathy for the planet’ through participatory ecological storytelling: From human-centered to planet-centered design. *Sustainability*, 15(10):7794, 2023.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Katy Tcha-Tokey, Emilie Loup-Escande, Olivier Christmann, and Simon Richir. A questionnaire to measure the user experience in immersive virtual environments. In Simon Richir, editor, *Proceedings of the 2016 Virtual Reality International Conference, Laval, France, March 23-25, 2016*, page 19. ACM, 2016. doi: 10.1145/2927929.2927955. URL <https://doi.org/10.1145/2927929.2927955>.
- Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL <https://qwenlm.github.io/blog/qwq-32b/>.
- Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and

- Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.
- Chen-Feng Tsai, Xiaochen Zhou, Sierra S. Liu, Jing Li, Mo Yu, and Hongyuan Mei. Can large language models play text games well? current state-of-the-art and open questions. *CoRR*, abs/2304.02868, 2023. doi: 10.48550/ARXIV.2304.02868. URL <https://doi.org/10.48550/arXiv.2304.02868>.
- Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojtek Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, Timo Ewalds, Dan Horgan, Manuel Kroiss, Ivo Danihelka, John Agapiou, Junhyuk Oh, Valentin Dalibard, David Choi, Laurent Sifre, Yury Sulsky, Sasha Vezhnevets, James Molloy, Trevor Cai, David Budden, Tom Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Toby Pohlen, Yuhuai Wu, Dani Yogatama, Julia Cohen, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Chris Apps, Koray Kavukcuoglu, Demis Hassabis, and David Silver. Alphastar: Mastering the real-time strategy game starcraft ii, 2019. URL <https://deepmind.google/discover/blog/alphastar-mastering-the-real-time-strategy-game-starcraft-ii/>.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL <https://arxiv.org/abs/2305.16291>.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022a. URL <https://openreview.net/forum?id=yzkSU5zdWd>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual*

Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022b. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *CoRR*, abs/2302.11382, 2023. doi: 10.48550/ARXIV.2302.11382. URL <https://doi.org/10.48550/arXiv.2302.11382>.

Marty J. Wolf, Keith W. Miller, and Frances S. Grodzinsky. Why we should have seen that coming: comments on microsoft's tay "experiment, " and wider implications. *SIGCAS Comput. Soc.*, 47(3):54–64, 2017. doi: 10.1145/3144592.3144598. URL <https://doi.org/10.1145/3144592.3144598>.

Grace H. Wolff and Cuihua Shen. Audience size, moderator activity, gender, and content diversity: Exploring user participation and financial commitment on twitch.tv. *New Media Soc.*, 26(2):859–881, 2024. doi: 10.1177/14614448211069996. URL <https://doi.org/10.1177/14614448211069996>.

Li Ping Wong. Focus group discussion: a tool for health and medical research. *Singapore Med J*, 49(3):256–60, 2008.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions, 2023. URL <https://arxiv.org/abs/2307.13339>.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023. URL <https://arxiv.org/abs/2309.07864>.

Chunchen Xu and Xiao Ge. Ai as a child of mother earth: Regrounding human-ai interaction in ecological thinking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2024.

Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. Language agents with reinforcement learning for strategic play in the werewolf game. In

Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net, 2024. URL <https://openreview.net/forum?id=usUPvQH3XK>.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Daijin Yang, Erica Kleinman, and Casper Harteveld. GPT for games: An updated scoping review (2020-2024). *CoRR*, abs/2411.00308, 2024b. doi: 10.48550/ARXIV.2411.00308. URL <https://doi.org/10.48550/arXiv.2411.00308>.

Shaokai Ye, Jessy Lauer, Mu Zhou, Alexander Mathis, and Mackenzie W. Mathis. Amadeusgpt: a natural language interface for interactive animal behavioral analysis. *CoRR*, abs/2307.04858, 2023. doi: 10.48550/ARXIV.2307.04858. URL <https://doi.org/10.48550/arXiv.2307.04858>.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=-Aw0rrrPUF>.

Nico Lukas Zentner. Openai 10. *Wirtschaftsinformatik & Management*, pages 1–7, 2025. URL <https://doi.org/10.1365/s35764-024-00544-z>.

Qinshi Zhang, Ruoyu Wen, Zijian Ding, Latisha Besariani Hendra, and Ray LC. Can AI prompt humans? multimodal agents prompt players' game actions and show consequences to raise sustainability awareness. *CoRR*, abs/2409.08486, 2024a. doi: 10.48550/ARXIV.2409.08486. URL <https://doi.org/10.48550/arXiv.2409.08486>.

Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling

with flexible parental involvement. *CoRR*, abs/2202.06205, 2022. URL <https://arxiv.org/abs/2202.06205>.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. Simulating classroom education with llm-empowered agents, 2024b. URL <https://arxiv.org/abs/2406.19226>.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2024. URL <https://arxiv.org/abs/2303.18223>.