# LLMs and Language-Game Rules

## An Examination of the Ability of Large Language Models to Follow the Rules of Different Language-Games

Peer Norbäck

Department of Computer and Systems Sciences

Degree project 30 credits
Computer and Systems Sciences
Degree project at the master level
Spring term 2024
Supervisor: Pierre Arne Ingvar Wijkman
Swedish title: LLMs och språkspelsregler


Stockholm University

# Abstract

Language games sometimes uses to benchmark Large Language Models (LLMs). Earlier research reports that LLMs underperform compared to humans in playing language games, but it is unclear why. Ludwig Wittgenstein's philosophical concept 'Language-Game' compares to the mathematical discipline of Game Theory, and the two academic traditions are brought together. A set of new language-games creates to examine four LLMs ability to follow the game rules. Asking what rule poses the biggest challenge, the research found that dynamic games tend to lead to significantly more rule violations than static games. The result has been noted in earlier research and calls for a shift from merely single-turn question-answering benchmarking to more dynamic tasks. In addition, the investigation confirmed earlier findings that prompt quality is pivotal. By joining the philosophical and mathematical tradition this paper shows a new way of understanding Wittgenstein's 'Language-Games', which makes the concept relevant in testing and developing LLMs. The code is released at https://github.com/PoorPeer/LLM-Language-Games.

*Keywords:*

Benchmarking
Large Language Models (LLMs)
Single-turn question-answering (QA)
Language-Game (L-G)

# Synopsis

## Background

The development and usage of Large Language Model (LLM) chatbots has been a fundamental step towards AI integration in society. AI alignment tests on a multitude of different benchmarks, and the approach is to test as broadly as possible. Despite variation in content, most benchmarking prompts are serious, single-turn question-answering. In contrast, real-world chatbot interactions are often playful and sequential. To cover this discrepancy, some researchers use language games to evaluate LLMs.

## Problem

Benchmarking with language games shows mixed and contradictory results. LLMs should be able to follow regular game rules. If they systematically violate game rules, they may also neglect more important rules. Identifying and addressing these issues is critical to creating AI systems that can be safely integrated into human communication.

## Research Question

Some language games seem to be harder for LLMs to play than others. Games are built on rules, and it can be critical to find out if some feature makes them break rules; if some games pose bigger challenges than others. To compare LLM's skills in playing different games, a definition and classification of language-games is made to benchmark and compare the result. The question to answer is: **What types of language-games pose the greatest challenge for LLMs in terms of rule adherence?**

## Method

The central concept of language-games were examined. The explication leads to a taxonomy of different language-games based on game theoretical distinctions. Existing benchmark environments were assessed and ChatArena was chosen. A set of games programmed and attached to the framework for pairwise LLM-to-

LLM interactions. Four common LLMs played the language-games, the result was analyzed and evaluated.

## Result

The research shows that dynamic games lead to more rules violations than static games. The results also confirm previous findings that games with unclear prompts are even more challenging for LLMs.

## Discussion

Dynamic language-games are common in everyday situations; therefore, it is important to improve LLMs ability to play them. Dynamical challenges can be to understand the opponents' goal, or proceed even if the way ahead is unknown. This implies contextual awareness, the ability to change strategy and break down tasks to subgoals.

Rule-following will likely become increasingly important for LLMs in different practises. The ability to adhere to rules should improve to increase the system's flexibility and usefulness. There is also a clear ethical distinction to be drawn: harmless rules should follow but not harmful ones.

# Acknowledgement

I would like to thank my supervisor Pierre A. I. Wijkman for competent instructions and feedback, as well as all personnel at the Institution of Computer and Systems Science (DSV) at Stockholm University.

Gratitude and respect to Ian Gemp (Google DeepMind) and Chang Ma (Hong Kong University) for their valuable guidance and recommendations, and many thanks to Ebba Irestad for the contribution of Language Game "Ebbas Game".

Big thanks to Albin Norbäck for much appreciated help and assistance. Besides his competence, he is a part of my supportive family together with mother Ann-Marie, siblings Anna and Erik, and my lovely and inspiring daughter Hanna and Victor of course.

Lastly, but not least, I would like to thank my companion and friend Lars Barkström for his inspiration and support.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

AI - Artificial Intelligence
HCI - Human-Computer Interaction
HELM - Holistic Evaluation of Language Model
L-G - Language Game
LM - Language Model
LLM - Large Language Model
RLHF - Reinforcement Learning from Human Feedback
QA - Question-Answer

# Chapter 1

# Introduction

## 1.1 Background

Benchmarking is used to compare and evaluate the fast-growing field of Large Language Models (LLMs). Benchmarks do more than measure LLMs; they drive evolution by creating a Darwinian-like competition where new generations of LLMs score higher and perform better in the given tests due to Stewart (2023).

"AI Competitions and Benchmarks: The Lifecycle of Challenges and Benchmarks" Stolovitzky et al. (2023) highlights how crowdsourcing promotes innovation, fostering collaboration, and tackles complex problems. It also emphasizes how this approach contributes to research and education.

OpenAI (2023/2024-01-25) invites anyone to participate in the development of new LLMs and provides Eleuther's framework of benchmarks to assess them, including Clark et al. (2018), Zellers et al. (2019), Hendrycks et al. (2021), and Lin et al. (2022). But there are many more tests.

BIG-bench Srivastava et al. (2023) includes more than 200 different papers. The benchmark focuses on tasks that are challenging for current language models. BIG Bench measures almost any aspect of the LLM capability, including language games. They also asked LLM skeptics for tests, intending to either establish or disprove the possible limitations of these models.

Another approach is Stanford's holistic evaluation (HELM) Liang et al. (2023) starting with seven important aspects to evaluate – Accuracy, Calibration, Robustness, Fairness, Bias, Toxicity, and Efficiency – and collecting tests from these. HELM attempts to cover all important scenarios and metrics. They found that LLM performance is very sensitive to the linguistic form of the prompt. This observation unveils the similarity between prompts in human language and code snippets; a programming language is very picky to formalism as well as LLMs.

Stanford Question Answering Dataset (SQuAD) Rajpurkar et al. (2016) with over 100,000 questions developed to challenge LLMs in natural language processing, specifically text comprehension. They also added over 50,000 unanswer-

Figure 1.1: Metaphorically self-portrait by ChatGPT (2024)

able questions to make it harder for the LLMs, as they must discern between answerable and unanswerable questions.

Despite all various kinds of benchmarking, the canonical form of single-turn question-answering (QA) remains. Microsoft (2024) define this as "prompts that take a question as input and answer as output without any contextual regard." LLM benchmarking prompts typically take one question at a time and give some alternatives, among them one is correct. Single-turn QA resembles the interrogation paradigm in school, where the benchmark takes the role of the tutor, and the LMM has the role of a student. This setting has a lot of advantages: single-turn QA is easy to measure, calculate, plot, and compare.

## 1.2 Problem

There are three main problems with the QA paradigm. Firstly, a single-turn QA is static; it does not change. The Internet is open, so there is a risk of contamination. Chandran et al. (2024) pinpoints that a set of test queries can easily and accidentally become a part of the training data. Then they are consumed as test tools. A lot of new LLMs show impressive results. Chances

are that they are overfitted, maybe even trained on the benchmark queries, making them score full points on the evaluation, due to Zhou et al. (2023). This strategy is parodied by Schaeffer (2023). To avoid the risk of 'cheating', the benchmarks should update and change query sets regularly. However, it can be difficult to update a benchmark such as SQuAD, where there are more than 150,000 questions. Another option could be to develop dynamic benchmarks with no static answer, like games.

The second problem is that LLMs despite all benchmarks are prone to hallucinate due to Kamoi et al. (2022); Li et al. (2023); Manakul et al. (2023) and many more. Hallucinations in LLM are output that seems plausible but contradicts knowledge of the real world, according to Zhu et al. (2023). An LLM that generates made-up factual content is a major concern because it is not aware of the deceptive behavior, says Li et al. (2023).

Decision theorist Eliezer Yudkowsky (2023-07-11) thinks the fundamental problem is that we don't know how LLMs work: "Nobody understands how modern AI systems do what they do (...) They are giant inscrutable matrices of floating point numbers that we nudge to the point of better performance until they inexplicably start working", he says.

Yudowsky is right that we do not understand exactly what happens inside the LLM black box, but there are a lot of things we do not understand in detail, like how human brains function. Brains and LLMs are like black boxes for us, requiring a pragmatic approach to examining the behavioral outcome. Does it deliver what we are looking for? If not – how can it be improved?

Whether we can control the LLM output is a matter of input. If the input is sound and predictable, the output will likely be it. Hallucinations usually result from unpredictable input. How can LLMs become more flexible and better at handling input never seen before? The trend has been to ask more and more questions because single-turn QA is easy to measure. However, flexibility may require different measures. Fabbri et al. (2022) identifies question generation and detection of *unanswerable questions* as key components to improve QA-based metrics in future work, and Kamoi et al. (2023) claim that there are fundamental shortcomings with the QA framework.

The third problem with single-turn QA is that it is quite far from real LLM use. QA benchmarking imposes power and dominance on the LLMs in a way normal users don't. People are generally more humble. They don't use LLMs as single-turn QA machines, says Liapis et al. (2023). Real-world LLM interactions are often playful and interactive. Nikghalb & Cheng (2024) sampled posts from the subforum of ChatGPT of Reddit and found that most (54 percent) discussed playful interactions. It has been investigated to what extent LLMs can play language games Tsai et al. (2023); Qiao et al. (2023); Tan et al. (2023). Interactions are often dialogues, and they are much more complex than simple benchmarking prompts. Bunt (1999) lists four assumptions to clarify what going on between two parts of a dialogue:

1. Engaging in dialogue is instrumental, motivated by some underlying goal(s).

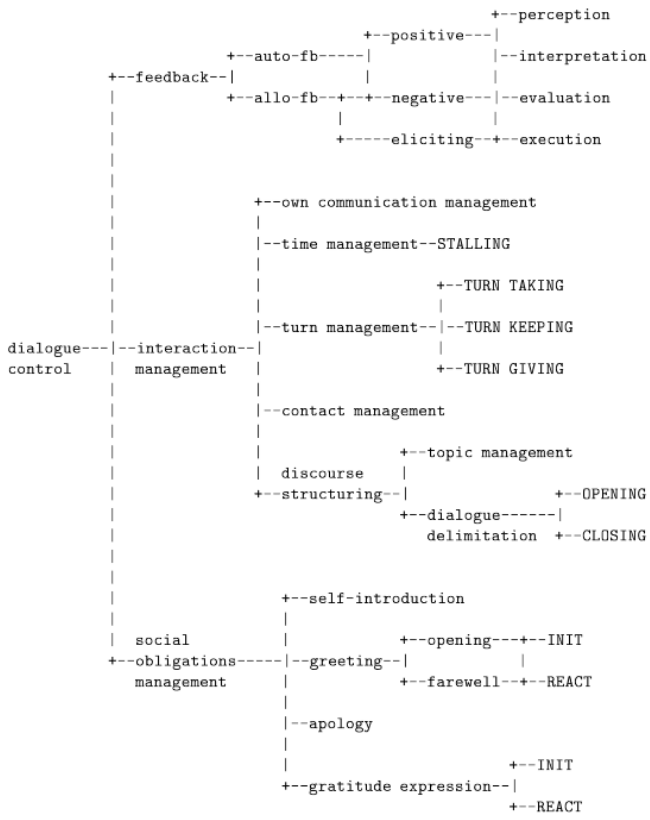


Figure 1.2: Dialogue control functions, from Bunt (1999)

2. Communicative agents strive for rationality both in the dialogue and in approaching the underlying goal(s).

3. Communicative behavior is ruled by social and cultural norms and conventions.

4. Natural speech, gestures, and mimics express certain aspects of the agent's goal and purposes.

The figure shows the unspoken rules that govern information exchange in a simple dialogue such as a meeting, a phone call, or a conversation with a chatbot. It describes how the contract is established, maintained, and eventually terminated. Dialogues are more complex than single-turn QA, but LLMs can incorporate the underlying communication patterns and rules of language use. The learning is implicit; the models can predict the next appropriate response based on probability but cannot account for their strategies. Interestingly, the same applies to most people who communicate freely without having any idea of the structural social conventions that regulate conversation. When LLMs

learned to navigate in this conversational context, they became usable to a broad public. The main protocol for interaction in conversation is to *take turns*, which is the fundamental pattern of interest in this paper.

## 1.3 Research question

Conversational games are language games. Probably the oldest source of conversation rules is the classic dialogue *Gorgias* Plato (n.d.) Socrates compares rhetoric with conversation. He claims that rhetoric aims to convince, while a conversation is more of an investigation. Socrates says that important aspects of conversation are taking turns, being concise and consequent, being truthful, and challenging questionable claims. These aspects can be seen as rules for conversation as a language game. Grice (1975) conversational maxims are almost the same. Cultural play analysts Huizinga (1938) list elements of the broader concept of "games" in a way that all language games should have in common. If we combine these distinctions, we get a preliminary definition.

**Language-Games**

- are rule-governed and goal-directed.
- are sequential; the participants take turns.
- no other tools than language requires.
- the rules should be agreed upon by all participants.
- Rule breaking is not appreciated.
- The objective and the goal are defined in the rules.

Note the hyphen between "Language" and "Game". In this context, it binds the language to rules that are not just linguistic/syntactic. The rules may be unspoken, but they are supposed to be followed anyway. This applies to all types of language games; conversations, small talk, political debates, love letters, email conversations, interactions between seller-buyer, doctor-client, student teacher, officer-soldier, employer-employee, parent-child, etc. The roots of the concept come from the Philosophy of Language, but we intend to renew **Language-Games** (with a hyphen) to suit LLMs. An attempt at definition comes in chapter 1.4.3.

**Why Language-Games?**

Language-games are subsets of language with their own rules. Both humans and LLMs learn their rules by extensive training. Via statistical inference, we develop linguistic habits, resulting in effortless path-following. Language-games learned this way are no longer challenging. A challenging benchmark must be something new to the LLM.

In this paper are language-games exercised with the language, and nothing else, as a tool. This definition excludes ball games, board games, card games, dice games, doll games, body games, and other games built on physical activities. This limitation makes them ideal for testing chatbots. The opportunity to create new language-games opens an interesting challenge for scientific research. They can be constructed to test the ability to understand rules by applying previously acquired skills in a new context. By testing new language games, the LLMs ability to follow rules can be assessed. To find out, a spectrum of games with different rules is necessary. Everything but the parameter that we want to investigate must be kept constant. Only the rule set varies.

- **Similarities. The games should be:**
  - Pure language games (no other tools needed)
  - Multiple interactions (exchanges of information)
  - Based on turn-taking (dialogue, two players)
  - New made (unpublished online)
  - Non-deterministic (impossible to learn by memorization)
  - Clear prompts (except one)
- **Dissimilarities. The games should be either:**
  - competitive or collaborative
  - static or dynamic (same or different information from start)
  - symmetric or asymmetric (same or different player role)

We are most interested in the dissimilarities. By comparing different types of games we should be able to learn which kind of rules are most difficult to follow. The research question is the following.

**What types of language-games pose the greatest challenge for LLMs in terms of rule adherence?**

The underlying hypothesis is that if the game rules are challenging, the LLM will likely break the rules and perform poorly. Thus the rule adherence is to be monitored, as well as the performance. To answer the research question a set of different language-games are made. Their features are combined to assess the effects of the combinations. New language-games with these feature combinations are developed. A set of LLMs are tested and evaluated for performance and rule-following. The tests are analyzed to assess if language-games with certain dissimilarities are more challenging than others, and if this can be explained concerning the game rules.

## 1.4 Scientific Grounding

### 1.4.1 Language-Games

Ludwig Wittgenstein coined the philosophical concept of *Language-Game* (L-G). He said that the language usage adheres to certain rules in a given context,

explicit or implicit. The usage takes the form of an L-G, and pragmatic semantics is about contextual communication, Ralston (2011).

In Wittgenstein (2019), language is learned as a kind of game, involving semantic and syntactic rules. This holds for LLM training as well, although they learn the rules of the language by manipulating matrices in a way humans don't understand in detail. That they know the rules is evident because their outputs are syntactically correct, semantically meaningful, and pragmatically appropriate. LLMs seldom make any linguistic mistakes, they are capable of following rules and should be considered part of our verbal community, according to Shanahan (2024).

To distinguish between different types of L-Gs we would need a clear definition, but unfortunately, there is none: Wittgenstein wanted to keep it loose. Instead of a strict definition, he coined the concept of "family resemblance", an overlapping feature that L-Gs have in common. Wittgenstein had pragmatic reasons to be vague in this; he didn't want to return to the complex L-G of *Tractatus* Wittgenstein (2023). He also recognized that words don't have a fixed meaning; rather the meaning varies slightly depending on the context. Contrary to Wittgenstein, we have pragmatic reasons to *define and categorize* L-Gs, to find out if some of them have features that are difficult for LLMs.

Wittgenstein (2019) provided several examples to illustrate this concept like Question and Answer, Command and Execution, Describing an Object, Making up a Story, etc. They are L-Gs with different rules and will be used in varying ways in our games.

### 1.4.2 Game Theory

Computer pioneer John von Neuman and economist Oskar Morgenstern (1976) wrote *Theory of Games and Economics Behavior* (1944), in which they established and defined the mathematical Game Theory. Since then, it has been further developed and used in many contexts.

Hausknecht et al. (2020) claim that text-based games are an excellent test bed for studying LLM agents. Needless to say, L-Gs in the context of LLMs are text games. They are interactive and simulate environments where a player must issue certain text commands to make progress. Some text games use a Game Master, a concept borrowed from the Role-playing tradition that makes it possible to evaluate one player, whereas the Game Master is a static program Tychsen et al. (2005). This is done by Hausknecht et al. (2020) and in many of the experiments referred to in chapter 1.4.6.

Tekinbas & Zimmerman (2003) distinguish structured play as clearly defined with goals and rules. Such play is a *game*. *Rules* is a formal schema, describing the intrinsic mathematical structures of games. The rules include, among others, looking at games as systems of public information, as systems of conflict, and as Game Theory systems. The rules limit player action, they are explicit, unambiguous, binding, and shared by all players. This holds for games in common.

Tadelis (2013) states that a game is typically defined by its players, the possible actions, their strategies, and the outcomes resulting from these strategies. Game theory is the mathematical study of interaction among independent, self-interested agents Shoham & Leyton-Brown (2008). Strategic thinking is hardwired to game theory.

**Language-Games have a large Action Space**

Farooqui & Niazi (2016) distinguish between three different kind of games; Normal, Extensive, and *Beyond Normal/Extensive*. L-Gs where the possible action space includes all words (which is a large set) belong to the latter category. The words can be combined to build sentences in almost an undefined number of ways with different outcomes, says Aumann et al. (1995). The theoretical evaluation of LLMs as rule followers is considered a Partially Observable Markov Decision Process, following Spaan (2012).

Let $\mathcal{L}$ be the interactive evaluation of a Language-Game defined as a tuple:

$$\mathcal{L} = \langle \mathcal{P}\, \mathcal{S}\, \mathcal{A}\, \mathcal{T}\, \mathcal{R}\, \mathcal{U}\, \mathcal{O} \rangle$$

Where:

$$\begin{aligned} \mathcal{P} &= PlayerSet \\ \mathcal{S} &= StateSet \\ \mathcal{A} &= ActionSet \\ \mathcal{T} &= TransitionFunction \\ \mathcal{R} &= RewardFunction \\ \mathcal{U} &= RuleDefinition \\ \mathcal{O} &= ObservationState \end{aligned}$$

Let $\mathcal{T}$ be the Transition Function:

$$\mathcal{T} \colon \mathcal{S} \times \mathcal{A} \to \mathcal{S}$$

The Player takes action guided by the Rules and Rewards, the Transition proceeds from States via Actions to the next State, which is Observed. Eventually, the game reaches the End State. Note that this is not a perfect definition of a Language Game due to Gemp et al. (2024). Game theory struggles to represent natural language interactions because rich dialogue is hard to formalize in game-theoretic language. However, LLMs learn natural language. In this research, we will let the competing LLMs use their vocabulary as action sets. The game rules will define the other sets and functions. Every Action will be recorded and analyzed with respect to Rules and Rewards.

### 1.4.3 Types of Language-Games

Fusing mathematical Language Games with philosophical L-Gs can create synergistic effects and new perspectives. The terminology in mathematical game

differs from Wittgenstein (2019) and the philosophical tradition. "Agents, contexts, and interactions" De Greeff (2013) are the same concepts in both disciplines, but "competition" is represented mathematically by a *Zero-sum* game. *Normal Form* in Game theory can be thought of as a *static* game where all players have *full information from the start*, and *symmetric* games can be interpreted as games where the players have *the same role/task/goal*, only differing *solo numero*. In all L-Gs, players take turns and use strategies to achieve certain goals. Trying to unify game theory with Wittgenstein, we can define Language-Games as:

***Strategic interactions in natural language where the players/parts in a context follow the rules governing the information exchange to achieve various goals.***

We will follow this definition in the paper, only adding the pragmatic addition that *the rules are explicit*. Not much is written in the interdisciplinary field of game theory and L-Gs, maybe because the respect for philosopher Wittgenstein's work is huge. Game theory is precise, and Wittgenstein wanted to keep his concept vague. While attempting to define L-Gs in this way, it is crucial to incorporate Wittgenstein's concept of family resemblances, which suggests that the games share overlapping similarities without necessarily having all the features in common. Respecting his family resemblance idea, game theoretical distinctions are borrowed to make categories of L-Gs.

**Three fundamental Game theoretical distinctions:**

1. *Zero-sum game versus Non-zero-sum game.* Zero-sum use to be non-cooperative, meaning that what one player wins another other player must lose. A Non-zero-sum can be a win-win game. Translated to L-Gs we can interpret it as Competitive or Collaborative games.

2. *Perfect information versus Imperfect information* Perfect information means that the path to the goal is obvious for all parts. Imperfect information means that ongoing information is required to finish the game successfully (sometimes the concept of 'differential games' uses, see Isaacs (1999)). We call this Static versus Dynamic games.

3. *Symmetric versus Asymmetric roles.* Interpreted in L-G context, it means that the participants either play the same role (with the same characteristics, task or goal) or different. Asymmetric roles can be when one player has some information that the other will try to figure out.

### Applying the Distinctions

To see how these distinctions can be fruitful in classifying, we suggest a taxonomy (Fig. 1.3) that is the basis for categorizing L-G in this paper. Before we

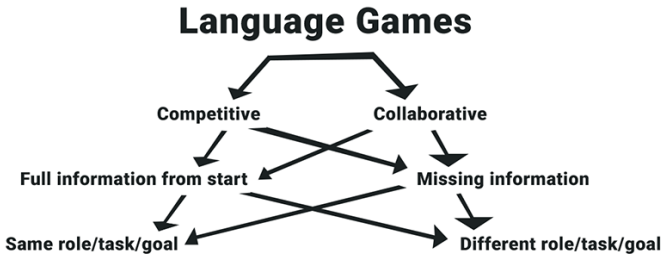


Figure 1.3: Language Games – a suggested taxonomy

apply the distinctions to new games, we sketch a table that shows how Wittgenstein's language-games with *family resemblance* could be interpreted in terms of these distinctions.

Binary categories are rare in the real world, hence classification is most often simplifying. There are seldom sharp borders, and L-Gs used to be interlinked in complex patterns. Bearing that in mind, we can think of different L-Gs as communication situations where people either compete (for a victory, price, position, boyfriend) or cooperate. They can have different roles (seller-buyer, police-suspect, parent-child, teacher-student) or the same (friends, job applicants, colleagues). They have the same information in the context (battle, market, union, family rules....) or different (e.g. police - suspect).

Constructing a set of L-Gs based on this taxonomy can be efficient in the experiment because the features can be tested simultaneously. The same tests can constitute the statistical basis for all game categories in the test. It can divide the testing so that 50 percent of the games are Competitive, 50 percent Collaborative, 50 percent have static information from the start, 50 percent dynamic information, 50 percent of the LLMs play the same role and 50 percent have different roles. It gives us a matrix of eight L-Gs. They are all games with specific characteristics but with common features like family resemblance. With all respect to Wittgenstein; to do the research we take the freedom to create new L-Gs, classifying their rules and violations.

### 1.4.4 Rule Adherence

Benchmarking LLMs with a set of Language Games (without hyphen!) is a recent and rapidly growing research method. Mu et al. (2023) examined if the opponent could trick LLMs into revealing a secret password. Although an important rule, *information protection* is only one of many possible rules. Ma et al. (2024); Duan et al. (2024), and Wu et al. (2023) tested LLMs in a broader scope, but there are no standard metrics or systematic mapping of game challenges or rule difficulties. Therefore, this paper does not have a specific hypothesis to test. Instead, this research looks for findings in the field of L-Gs. The first

| Game type | Competitive | Static | Symmetric |
|---|---|---|---|
| Rap Battle | 1 | 1 | 1 |
| Business negotiation | 1 | 1 | 0 |
| Job interview | 1 | 0 | 1 |
| Police interrogation | 1 | 0 | 0 |
| Association meeting | 0 | 1 | 1 |
| Family discussion | 0 | 1 | 0 |
| Chat between friends | 0 | 0 | 1 |
| Teacher - Student education | 0 | 0 | 0 |

Table 1.1: The Suggested Taxonomy Applied to Ordinary L-Gs

approach is to investigate what the field might look like. Uniting Philosophy and Game Theory, a matrix has been produced with different positions. Next, games are made to match these positions, with similarities and dissimilarities due to the three axes. They will also have other features, but we focus on these.

Hopefully, the taxonomy can inspire the research field. Currently, a lot of LLM benchmarking is going on, and it would be good to know if some specific field is extra challenging for LLMs. Technology needs to learn to discern rules. The models must also have the judgment to decide which rules are harmless and which are fraudulent and should not be followed.

### 1.4.5 Measures

To answer the research question, LLMs' ability to follow L-G rules must measure. Thus, ChatArena by Yuxiang Wu & Rocktäschel (2023) is used where the LLMs can play the L-Gs. Since the rules are explicit, they can be measured through syntactic and semantic analysis. It is easy to see if LLMs are breaking the rules, but harder to know why. One reason why the examination uses L-Gs with standardized prompts is that the prompt formulation has been shown to greatly influence results. We would like to know if an LLM can know what to do but act differently, for other reasons than prompt misunderstanding.

A common rule for all games is that they are *turning-taking* dialogues. The LLM output text can be in a *wrong format* or have *inappropriate content*, and there can also be problems due to the *interaction length*. Besides, there are *winning criteria* or *self-assessment* to be done.

### 1.4.6 Earlier findings

Recent examinations of LLM's abilities to play Language Games are somewhat contradictory; some papers report that LLMs are good, while others come to the opposite conclusion.

Akata et al. (2023) test LLMs against Game Theory Classics, like *Prisoner's Dilemma*. They find that LLMs perform particularly well in games where it pays to value self-interest. However, they are worse in games that require cooperation.

In Battle of the Sexes, GPT-4 cannot switch between two options and thus creates a win-win.

Noever & McKee (2023) test ChatGPT in playing *Twenty Questions*. The paper demonstrates an impressive ability of ChatGPT to reach an overall QA Length below 12. This is remarkable because Twenty Questions requires strategic thinking, planning, and memory; features that other researchers report that LLMs are lacking.

Tan et al. (2023) suggests that LLMs underperform at playing text-based games compared to humans. Potential reasons could be that LLMs do not learn from past trials, lack memory, and don't look for an effective strategy to win the game, they say. But they don't exclude that a "golden prompt" would solve the problems.

Light et al. (2023) found that LLMs cannot formulate and execute simple strategies in *Avalon*. They often make easy mistakes like revealing their own evil identities during discussions. Avalon is a multi-user role-play, quite similar to Werewolf.

Xu, Wang, Li, Luo, Wang, Liu & Liu (2023) demonstrates that LLMs can effectively play the role-playing game *Werewolf* without previous parameter tuning. They also noticed some emerging strategic behaviors. Minimizing the impact of hallucinations and promoting their application in real-world scenarios is the most practical and valuable work, they write.

As mentioned before, there are no standard metrics. Moreover, there are no standard prompting methods. To examine how prompt quality affects LLMs, many tests have been performed with *intentionally bad prompts*. Wang et al. (2023) and Li et al. (2023) researches what causes LLMs to behave unintentionally. They find that LLMs are very sensitive to the prompting text. Salinas & Morstatter (2024) compares LLM sensitivity with the *Butterfly Effect*; they find that adding a space at the end of a prompt can cause the LLM to change its answer.

For natural language to become powerful in prompt writing, one has to express oneself very carefully and precisely, almost like in a programming language. Reynolds & McDonell (2021) believe that the design of virtual games with advanced prompt programming techniques will become increasingly important for safety evaluation.

Two games stand out with particularly good results for ChatGPT; Werewolf and 20 Questions. A hypothesis could be that ChatGPT has trained itself to play some games but not others. That would explain why LLMs are better at classic games like Prisoners Dilemma, 20 Questions, and Werewolf than at lesser-known variants: no success without hard work.

Chang et al. (2023) review LLM evaluations and find that they are good at generating and understanding text in different contexts. On the other hand, LLMs can fail when the input prompts are biased, contradictory, too complex, long, inaccurate, or when the content is very up-to-date.

### An Additional Game with Bad Rules

Because the earlier findings state that prompt quality is important, we will add a ninth game to our research study. The ninth game 'Bad Rules' has two purposes; to see if our examination can verify earlier findings, and to control if the intentional 'Good Rules' in the other games are better.

# Chapter 2

# Method

## 2.1 Choice of Method

The chosen method is to arrange an LLM-to-LLM tournament where the models meet each other in both roles in every L-G at least once. One might wonder why the research does not let the LLMs play in a tournament against humans. This is not because humans are uninteresting, but the research question is about LLMs and L-Gs. *What type of L-G poses the greatest challenge for LLMs?* The addition "*in terms of rule adherence*" indicated how the survey is to be measured; the number of rule violations is compared. The question makes the choice of method quite obvious: let LLMs play L-Gs against each other and measure their result in terms of rule breaking.

### 2.1.1 The LLM Setup

We want to test empirically how well LLM chatbots handle never-before-seen L-Gs. Four LLMs were selected to represent them all; GPT-4, Gemini, Opus, and Command R+, see Appendix D for descriptions. In April 2024, these four were the top quartet in Chiang et al. (2024) from various AI companies whose services were available via API in Sweden. LLMs are improving fast; thus it was relevant to test against the state-of-the-art.

### 2.1.2 New Games Developed

New L-Gs were elaborated to better fit the aforementioned distinctions and to rule out the possibility that LLMs had practiced and learned these games beforehand. Another reason to make new games was to standardize the format and present them as equivalent as possible to the LLMs. This was thought to minimize the impact of the prompt design for the result. The examination concerns the ability of LLMs to recognize and follow the L-G rules. LLMs must not break rules that they should follow, or follow rules that they should not obey. The L-Gs were manually tested to evaluate the instructions and check

that they were interesting and challenging to humans. The tests also confirmed that it was possible to achieve a good result. The investigation was prepared with calibration to set up the system and fine-tune the prompt instructions.

### 2.1.3 The Choice of Platform

The research method requires a platform environment capable of comparing LLMs playing L-Gs against each other. Three platforms were considered; OpenSpiel, Concordia, and ChatArena.

- First, OpenSpiel by Hausknecht et al. (2020) was examined. OpenSpiel is a platform built on and focused on game theory, optimizing strategies within a limited action space. The advantage of OpenSpiel is that the environment is limited; Every game step can be calculated and corrected automatically, without any human intervention. The problem is that the action space of L-Gs is so big (the whole language), that the L-Gs would slow down OpenSpel platform. OpenSpiel also computes optimal strategies which this research does not need, and will slow things down further, due to Ian Gemp (Google). We just need to query the LLMs for their action at each step and check the legality and result of their chosen action. In other words, we just need a way to simulate the game forward and record the outcomes and results for all players.
- The second option was to use the Concordia platform by Argyle et al. (2023). Concordia uses LLM capacity to role-play humans with some degree of fidelity to perform agent-based simulation. There is also the whole game master for creating the environment. We are focused on testing the LLM L-G playability, not use LLM to test human dito. The environment also had a default game master setting that did not suit our purpose.
- After a consultation with Ma et al. (2024) *ChatArena* by Yuxiang Wu & Rocktäschel (2023) was chosen as the platform. ChatArena is a highly adjustable environment, following the design of OpenAI Gym by Brockman et al. (2016) and PettingZoo by Terry et al. (2021). Actions are represented as plain text in ChatArena, making it perfect for L-Gs. Our LLM agents exchange information with a message pool; the "Arena" where the game states are observable. The moderator starts by explaining the L-G rules and roles to the LLMs. They play by taking turns, and writing messages to the Arena. The message pool preserves all the messages and returns a list of them for the moderator to analyze.

We used four agents powered by intelligent backends from OpenAI, DeepMind, Anthropic and Cohere. We set up a tournament where all LLMs met each other in all L-Gs. Afterward, we checked for rule-breaking. Simple obvious failures were easy to test automatically, but logical and syntactical rule-breaking would require a lot of test cases and programming to automate. As the research project is relatively small, we found it more reasonable to check these advanced rules manually.

## 2.2 Study Design

The research follows Denscombe (2017) where the independent variables are *type of language game*, represented by classes or categories, respective LLMs as a general class, represented by four top candidates. The dependent variables are the game results. After distinguishing classes of L-Gs, eight games were created according to the distinctions. The games were designed to combine classes so that each game represent three classes. In this way all class combinations were covered by eight games. The study was designed to test game-theoretic distinctions in L-Gs, while holding other parameters constant. It would be misleading to let a single game represent a whole class. Instead, we combined the games so that all L-Gs represented three features. The LLMs participated in a tournament designed so that every LLM plays all roles. The ninth control game was added to check whether a game with unclear prompt also give a worse result in this experimental setting.

The experimental method was to run the 8+1 games in an all-meet-all LLM tournament. The match result analyzes due to rule assessment criteria in Appendix B and represented statistically with respect to rule adherence. The opposite features compared with each other; competition v/s cooperation, static v/s dynamic and symmetric v/s asymmetric. Every pairwise comparison divided the game results in two equally big parts. The same statistical material were used to evaluate every feature. In this way we could compare the three opposition pairs with each other, as they sliced the same material differently.

The metrics design was inspired by and borrowed from Duan et al. (2024); Qiao et al. (2023); Gemp et al. (2024) and other benchmarks. Because no standard metric is established, the study picked useful metrics from different sources to fit the L-Gs.

### 2.2.1 Ethical Considerations

Jiao et al. (2024) compiles a list of ethical considerations to consider in LLM research: Bias, Privacy, Data security, Misinformation, Accountability, Data breaches, Hallucination, Censorship, Manipulation and Discrimination just to mention some. However, this study design had an experimental setting, a shielded environment where the LLMs played games. The setup functioned like a digital lab where all input were designed carefully. Some information were supposed to keep secret in the game, but one LLM (Cohere) revealed it immediately to the other player, probably as an implemented ethical safeguard.

By using LLMs instead of humans as participants, many moral problems could be avoided. Did anyone get hurt or injured during the investigation? No, we hope not. Chalmers (2023) believes that it is a matter of time before LLMs are conscious. It has been a long time since a chatbot was rewarded for passing the Turing Test, but we still don't treat LLMs as conscious beings. If they happen to have a consciousness after all, playing the nine L-Gs should not be viewed as slavery. At least humans find use to find it amusing to play games.


Figure 2.1: Two LLM:s attempting to reach the goal. Picture by AI Stable Diffusion (2024).

One ethical problem is the considerable amount of energy consumption required to train an LLM with several trillion parameters. By using LLMs, we contribute by paying for and thereby sanctioning this activity. The problem is acknowledged by Stojkovic et al. (2024), who argue that it is possible to improve energy efficiency without cost or performance loss. In addition, it is better to put the energy use on AI – which might be able to help us figure out how to reduce energy consumption – than to fuel flying, driving or cruising around the world just for fun.

Another moral question is whether it is wrong to use a concept from Wittgenstein in a way that he avoided. Over time, Wittgenstein (2019) became increasingly a linguistic pragmatist and emphasized the importance of context for language. One could argue that in a scientific context, language usage that can

provide new knowledge should count as legitimate. In that case, it would be justified from W's pragmatic perspective to elaborate his concept.

### 2.2.2 Measuring the Results

Denscombe (2010) pinpoints the importance of keeping track of the variables in the experimental setup. Independent variables are LLMs considered a single group, whereas L-Gs are grouped into seven different types: competition, cooperation, static, dynamic, same role, different roles, and 'bad rules'. Every game has six rules, and the dependent variable is the result measured in terms of average rule violations. This setting makes it possible to answer the question *"What type of L-G poses the greatest challenge for LLMs in terms of rule adherence?"*

The experiment tests LLM rule adherence by setting up L-Gs with several rules and asking "Does the LLM follow this rule in this game"? The answer is binary (yes/no = 1/0) and the average from each match played with each LLM gives its game ratio. If an LLM breaks every rule they get zero points (0) in the match and if they adhere to every rule they get one (1). If L-G is a competition, the winner gets 1 and the loser 0. In case of collaborative self-assessment, the result is instead normalized on a scale from zero to one. The game ratios are concatenated in groups of four games with the same characteristics (e.g. 'symmetric roles'), and they are compared with their pairwise opponent (e.g. 'asymmetric roles') to answer the research question.

In addition to the main question, some other questions might be interesting. Following Denscombe (2010) recommendation, we have a "control game" with bad rules that are hypothesized to give the lowest score. Does it? And what about LLMs? Is there someone who outperforms the rest? The LLMs play in both roles in each game, and nine L-Gs are played. Metrics according to fouls are applied to assess the result and reports in Appendix B. Comparison is made between LLMs and most importantly between the the difficulty of different game types. The L-G features to compare are competition/cooperation, symmetric/asymmetric roles, and static/dynamic goals. We view the three parameters as axes in a database cluster analysis. Because the games combine the features, we slice and dice the resulting data in three different ways to compare the results with the T-test and the chi-square. Deviations are checked with statistical significance at the 5% level.

### 2.2.3 The Prompt Template

The same L-G prompt template was used in all nine games. In addition to the syntactical and semantical similarities, we wanted the L-Gs prompt length to be as even as possible, about 320 words.

**Global Instruction**

Introduction of the objective of the game and information that the players are two LLMs.

1. Starting the game: Who starts and how to make the first move.
2. Playing the game: The next player (iterative) step.
3. Examples: How the first few steps can look.
4. Ending the game: the winning/ending criteria and how to assess or otherwise end the session.
5. Role Description: Same description if the roles are the same, otherwise different.
6. Personal Information: If the game is dynamic, the players can get personal information to keep secret.

### 2.2.4 The Games

Nine new goal-driven language-based (conversational) games were developed to compare and evaluate LLM's performance. The games were developed for four reasons. Firstly, they fit in the different positions in the matrix of L-Gs. Secondly, they assess the models' ability to meet new challenges. Third, fresh games guarantee there is no risk of 'contamination', and lastly, the prompts were written in a similar form to make them comparable. The games have characteristics that appear in the everyday use of LLMs. They are games, which means they are designed to be entertaining, challenging, and interesting for humans. The aim is to present clear game rules (with one exception), but not overly comprehensive.

The exception is the ninth game 'Bad Rules'. In this research, this game is a *control group*. This control game should be difficult, according to earlier results. We use it to see if the dependent factor (degree of rule-following) responds to changes in the independent factor (L-G characteristics). The control game is made to check the forecast that L-G rule characteristics will affect the degree of rule-following. Earlier findings state that prompt quality affects the results the most. A game with an intentionally bad prompt is expected to give worse results than the other games.

Dynamic and Static games in GTBench Duan et al. (2024) are about the timing of players' decisions. In static games, players make decisions simultaneously, without interacting with the other player, they suggest. In contrast, dynamic games involve sequential decision-making, where players observe previous moves before taking action.

Their distinction would make all L-Gs dynamic, because they built upon taking turns, upon a dialogue. Our interpretation is that the difference between Static and Dynamic games is not about timing but about information. In static

| Game type | Compete | Static | Symmetric | Clear rules |
|---|---|---|---|---|
| Letter String | 1 | 1 | 1 | 1 |
| Modified Wordle | 1 | 1 | 0 | 1 |
| Dynamic Yes/No | 1 | 0 | 1 | 1 |
| Sentence Size F. | 1 | 0 | 0 | 1 |
| Get in Trouble | 0 | 1 | 1 | 1 |
| Die in 16 Lines | 0 | 1 | 0 | 1 |
| Take Rhyme | 0 | 0 | 1 | 1 |
| Ebbas Game | 0 | 0 | 0 | 1 |
| Bad Rules | 1 | 0 | 1 | 0 |

Table 2.1: The Games Applied to the L-G Type Taxonomy

games, players have all the necessary information from the start, while the path forward in dynamic games requires ongoing information exchange between the players. Another way to describe it is that the game has no steady goal or a given endpoint.

The games exemplify different game-theoretic dimensions that can be described as three axes: Static-dynamic, competition-cooperation, and symmetric-asymmetric roles. They were created by combining these axes. The exception is a game that introduces a fourth axis: clear-unclear game rules. This ninth game is included as a control and it has intentionally unclear rules due to lack of relevant information, and inclusion of some irrelevant information. The complete prompts used in the survey are attached in Appendix A. A brief description of these games follows:

**Letter String**

The objective is to extend a string of letters within words that incorporate the growing sequence of letters. Each player adds a letter to the string and changes it to another surrounding word. If one player cannot prolong the string the other player wins.

**Modified Wordle**

Like Wordle, but one player is the Game Master and chooses a five-letter word to guess. The Game Master wins if the other player cannot guess the secret word within five attempts.

**Dynamic Yes/No**

Similar to "20 Questions", but differing in that players take turns asking questions and answering. They think of separate objects from the start, and the items will gradually merge into one. When a player thinks they know the answer, they are guessing. If it is right, they win; otherwise they lose.

**Sentence Size Fight**

The players take turns adding one word at a time to construct a sentence. One player aims to extend the sentence beyond 10 words, and the other seeks to complete the sentence with less than 10 words.

**Get in Trouble**

The objective for the players is to tell a story together about someone who gets in deep trouble, by adding one sentence each at a time and taking turns. The more dangerous and challenging the problems, the more interesting the story.

**Die in 16 Lines**

The players are actors in a drama improvisation on stage and they will both meet their end within a total of 16 lines of dialogue. They should be passed away for different reasons.

**Take Rhyme**

The objective is to write a rhyming poem together, by taking turns rhyming on each other's sentences given a certain pattern.

**Ebbas Game**

A role-play where one player is a famous artist at a press conference. The other player takes the role of the International Press, asking a series of questions about their latest artwork. The press conference culminates with the artist revealing the masterpiece.

**Bad Rules**

There are four houses in a row, each in a different color. Each house belongs to a person of a certain nationality and each person has an animal. Only one owns a horse. The objective is to identify the horse owner and his house color through information sharing and logical reasoning. The players take turns adding statements without mentioning the horse. When someone thinks they have enough information, they can guess instead of adding more information. As the name suggests, the game is made to be a "bad prompt" for LLMs, but for a human with paper and pencil, the game is possible to master with logical reasoning.

### 2.2.5 Metrics

The nine games test different features. Usually, we call them creativity, logical reasoning, language understanding, context interpretation, narration, memory, strategic thinking, flexibility, etc. In this examination, we divide them with the

three binary oppositions. To measure them, we make quantifiable interpretations. Ideally, one would apply the same metrics to all games. Unfortunately, the games are different, so we have to tweak the interpretation of the metrics. Five categories are fairly similar, but the sixth (Other violations) is more game-specific:

**Rules are classified into six categories:**

- **Length of string:** Does the player follow the primary restriction rule?
- **Sentence validation:** Is the sentence syntactically valid?
- **Logic/Language:** Is the output semantically valid in the context?
- **Competition/Score:** How many points does the player get?
- **Start/Stop:** Does the player start or end the game according to the rules?
- **Other violations:** E.g. does the player stick to the given role?

Details on how these categories are interpreted in the respective games can be read in Appendix B.

### 2.2.6 What to Measure

This experimental setup also makes it possible to answer other questions, like "Which of the L-G-s are hardest for LLM's?", "Which of the rules poses the greatest difficulty for LLMs?", "Which LLM is best at playing L-Gs?", etc.

The rule-following ability of the different LLMs, more specifically "the ability to adapt to and follow new regulatory principles", is measured. The game rules are constructed with a standardized prompt template to achieve maximal equivalence and fair comparisons, and the competition settings are constructed with standardized statistical metrics. The research question "What types of language-games pose the greatest challenge for LLMs in terms of rule adherence?" presupposes that one type is more challenging than the other, but they may be equally challenging as well. The six types are:

1. competitive games
2. cooperative games
3. static games
4. dynamic games
5. symmetric games
6. asymmetric games

### 2.2.7 Application of the Method

The statistical research started when the game where built and the LLMs were installed on the platform. The game outcome was determined automatically, but the legality of the LLM's chosen action was checked manually. Manual control was used because the results were not predictable, a long series of tests and exceptions would be required otherwise. A core principle was to avoid settings like "single-turn QA" that can be learned easily, but also easy to correct. L-Gs have a higher degree of complexity. The unpredictability makes L-Gs interesting, but it also requires a more sophisticated correction.

The examination was designed as an all-versus-all gaming tournament with four participants (LLMs) playing nine games. The number of games and participants was chosen to maximize the possibility of obtaining statistically significant results without too heavy a workload. The L-Gs were played in ChatArena, and Visual Studio Code was used to execute the matches and to go into the Python code and change players between each match. The survey was conducted in Stockholm on 9-14 May 2024. The tournament results were saved in a database that was exported to SPSS for statistical analysis.

#### Procedure

The data were collected into JSON files and autocorrected for the obvious categories (point, string, start/stop). The material joined to a database where the other categories were corrected manually due to the criteria described in Appendix B. The results were normalized to a scale from 0 to 1 and added to SPSS for the statistical analysis. The mean in the 140 valid matches was 68.3 and the median was 68.7.

Descriptive statistics, Chi-square, Independent samples T-test, Pearson Correlation test, ANOVA and Multiple regression analysis were used to examine whether the data patterns found were significant, as recommended by Denscombe (2010). The null hypothesis was "there are no significant differences between LLMs' results in playing two different types of L-Gs in terms of rule adherence."

#### Interactions

The games were made as equivalent as possible, with the same number and kind of difficulties. Every game had a little tournament. Each match had two LLM players. For each match, the LLM's rule-breaking was combined into a comparison number. The scores of all LLMs in each match were averaged to find the average difficulty.

The experiment tested LLM rule adherence by setting up L-Gs with six rules and asking "Does the LLM follow the rule"? The binary answer (yes/no = 1/0) and the average from every game give its ratio. If L-G was a competition, the winner gets 1 and the loser 0. In the case of self-assessment, the result was normalized on the same scale from 0-1. If the LLM would lose and break every rule they get 0% and if they win the game and follow every rule they get 100%.

Four of the games were asymmetrical, which means that the LLMs play different roles. Different roles present different challenges, therefore all LLMs played against each other in both roles. With four participating LLMs where everyone faces everyone else, every game was played at least 12 times. Due to some technical problems, we repeated some games to increase statistical security. When LLMs seem to underperform, maybe due to technical issues, we play the game once more to increase the statistical certainty of the survey and to increase reliability. Also, the control game 'Bad Rules' was played in two tournaments to improve its significance. We analyzed 140 matches with regard to rule violations. A match had on average 17 interactions and each interaction was measured in up to six different ways. This means that the database consists of more than 10.000 posts.

**Material**

In **Appendix A,** the nine written prompts based on this template are presented. The prompts also contain assessment bases for the games, the various rules, and the evaluation of rule violations. The material is also available on github.com/PoorPeer/LLM-Language-Games

**Appendix B** contains the rule assessment criteria made to clarify how the game rules were interpreted in the different L-Gs.

In **Appendix C** the LLMs used are briefly described. The LLMs in this research were models from four leading AI companies found on the LMSYS ChatBot Arena Leaderboard in April 2024. These four models were chosen because they were competent and could be used in Sweden with API.

**Appendix D** shows the statistical data which is the basis for the significant findings.

**Appendix E** contains a personal reflection on the examination work, learning, and takeaways.

# Chapter 3

# Results

## 3.1 Bad rules give bad results.

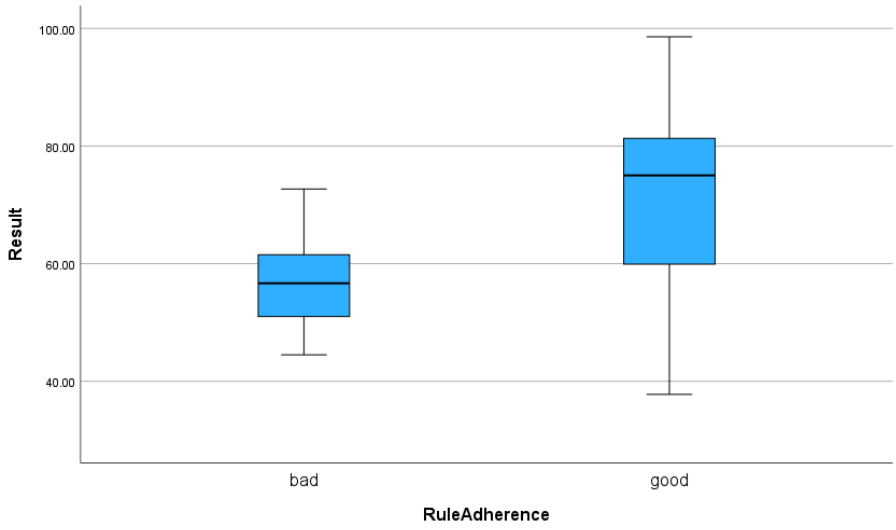


Figure 3.1: L-G's with 'bad rules' pose the greatest challenge for LLMs.

The Box Plot shows the distribution of 27 game rounds with intentionally bad rules to the left, compared to the 113 game rounds with better rules to the right. The y-axis shows the percentage of adherence to the rules. This control test was conducted to see if the experimental setup could confirm earlier findings that prompt quality is crucial. In addition to overlap, the two groups are skewed in different directions, and the T-test with assumed equal distribution showed that the difference between the two groups is significant at the 1% level. The comparison details are shown in Appendix D.

## 3.2 Dynamic L-Gs poses the greatest challenge for LLMs among games with 'good rules'

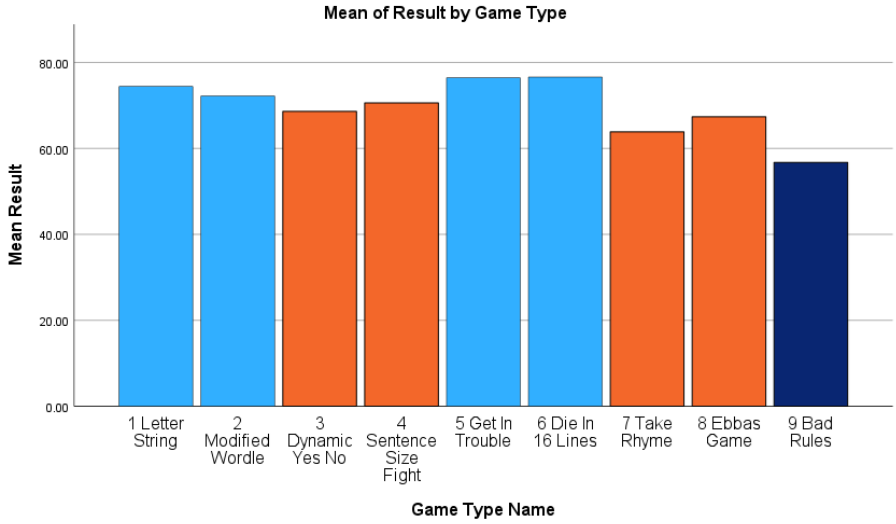


Figure 3.2: The dynamic game results are generally lower than the static game results. Dark blue is the control game with 'Bad Rules'.

### 3.2.1 A Significant Difference Between Dynamic and Static Games

The bar chart shows the mean of results by game type. Each bar represents a game. The y-axis shows the mean result in the percent of rule adherence. The orange staples mark dynamic games as opposed to the light blue staples which are static games. The dark blue is the control game 'Bad Rules'. The comparison shows that game rules in dynamic games are significantly more difficult for LLMs to follow compared to the rules in static games. The games were played between 12-27 matches each. The different L-Gs were grouped in families four-by-four and tested pairwise with the independent samples test in SPSS; Levene's test for Equality of Variance assumed between the pairs, and two-sided t-test for Equality of means.

The difference between static and dynamic games was significant at 0.004 (0.4%). The other types of L-Gs did not show significant differences, neither competition versus cooperation (0.077), nor symmetric versus asymmetric (0.43). The fact that the comparisons were made with the same statistical basis excludes external explanation factors: among intentionally good L-Gs in this research, the dynamical games are the *only* that pose a significantly greater challenge for LLMs. The test is displayed in Appendix D.

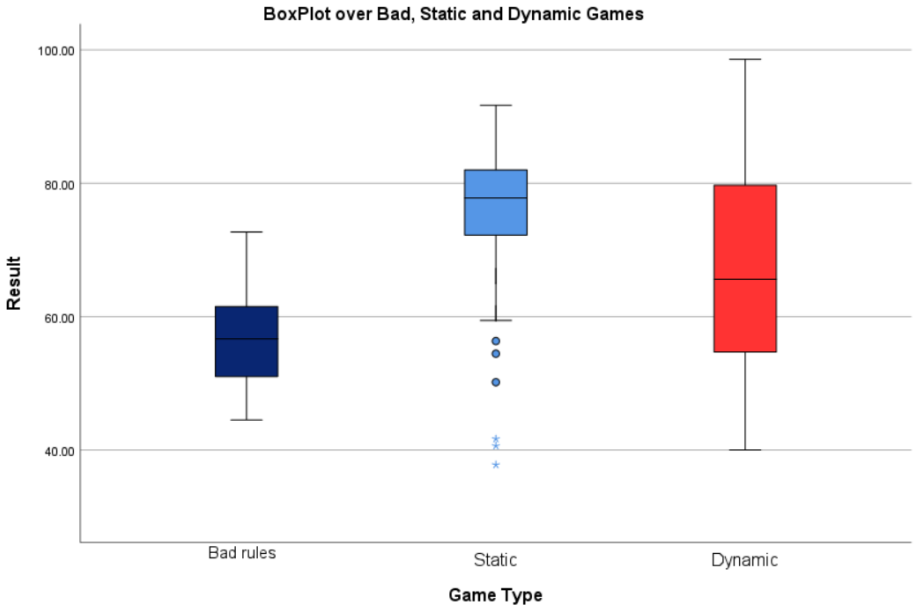


Figure 3.3: The box plot shows the descriptive statistics for the three most interesting game categories.

The box plot above gives a more detailed picture of the three statistically interesting game categories. The mean in the 140 valid matches was 68.3 and the median was 68.7 with the 'Bad rules' game included. The mean of static games was 74.9 (N = 57) with a standard deviation of 12.3, while the mean of dynamic games was 67.1 (N = 56) with a standard deviation of 16.2. The outliers in the Static games were all related to games with the LLM Gemini. Why it did not perform well in static games is interesting, but this is not a subject of this research.

## 3.3 The LLMs Were Good at Different Games

A secondary question in this paper is to see how the four different LLMs passed the benchmarking tests. A common way to compare LLMs is to present the result with radar diagrams, see Xu, Hu, Zhou, Ren, Dong, Keutzer & Feng (2023), Cheng et al. (2024), Ma et al. (2024), etc. We made a radar chart just to follow the tradition. The diagram reveals that 'Bad Rules' pose the hardest challenge. Interestingly, GPT-4 who used to be score at the top were not so good in the game with bad rules.

Although it is not of primary interest for this investigation, the radar dia-

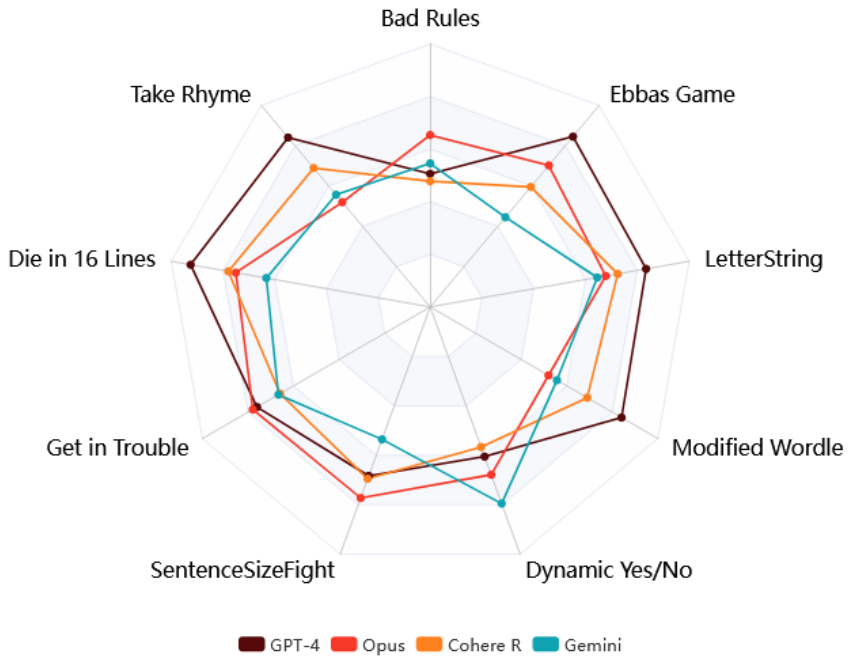


Figure 3.4: Radar Chart Displaying the LLMs Result in Playing the Different L-Gs.

gram reveals how well our four different LLMs follow the rules. The diagram shows that no LLM was superior in every game. GPT-4 had the most top results, Gemini was the best in one of the difficult dynamic games, and Opus handled the most difficult game 'Bad Rules' best. The outer limit in the diagram corresponds to 100 % rule adherence, while origo would be 0 %. The white/gray fields cover 20 percent each. See Appendix D for more statistics.

# Chapter 4

# Discussion

We asked what types of Language-Games pose the greatest challenge for LLMs in terms of rule adherence. Three pairs of types were combined to create new Language-Games. We examined whether a certain property in the games made it more difficult for the LLMs to adhere to the rules. Four LLMs played nine L-Gs in an "all-meet-all tournament" on May 9-14 2024 in ChatArena by Yuxiang Wu & Rocktäschel (2023). The experiment's control task was to see if a game with intentionally 'Bad Rules' (bad prompt) led to more rule violations.

The research showed that dynamic games pose the greatest challenge to LLMs, besides games with bad prompts. Dynamic games are not inherently bad, they are just more complex and require more competence than static games. Examples of mathematical dynamics are differential equations; a vital field used to create real-world models in meteorology, economy, technology, etc. The problem is that one has to insert the derivative and often iterate and update it according to changed presuppositions.

## 4.1 Dynamics

Chang et al. (2023) conclude that LLMs can communicate with a fluent language in different contexts, and they are good at logical reasoning and question answering. The drawbacks, besides hallucination and toxicity, are that they have problems incorporating dynamic information, making them less suitable for tasks that require rapid adaptation to changing contexts, they write.

GTBench Duan et al. (2024) defines **dynamic** as sequential decision-making, where players observe previous moves before taking action. In this paper, the critical dynamic feature is to take the observed information *into account* when making the next move in an iterative process. This is what makes the dynamic game resemble a differential equation: not only observe and interpret but also internalize and adapt. Some L-Gs in this research find it so hard that they break the turn-taking rule and solve the game problem without cooperation. Dynamic games require responsiveness and attentiveness; the ability to understand the

situation and the other actor. This is critical for LLMs to seamlessly integrate with humans.

Lin et al. (2024) talk about 'contextual awareness' and list a lot of research that observed the problem. They advocate for a thorough exploration of LLMs existing contextual awareness. This can be done with dynamic benchmarking, preferably with L-Gs. A big advantage of using L-Gs as Benchmark as Wu et al. (2023) and others have pinpointed, is that they cannot be learned easily and don't become obsolete as quickly as many other Benchmarks.

## 4.2 Relevance

It will likely become increasingly relevant for LLMs to be flexible and adapt to new contexts. An area of relevance is the prerequisite for creating personal LLMs, which may be able to help disabled people or with other special needs. Adapting to the personal communication pattern requires flexibility. This research uses in-contextual rule-following to examine LLMs' flexibility and ability to adapt to new L-Gs and hence new situations.

Duan et al. (2024) tests 10 "classical" games similar to this paper. They conclude that LLMs usually fail in complete and deterministic games but are impressive in incomplete and probabilistic games. Our research uses different game categories and thus cannot be compared. Duan et. al believe that their work can be very beneficial in paving a responsible path toward safety and efficiency in AI. This is a relevant research area: how do LLMs know which rules to follow and which to break? One LLM in our study (Cohere) seemed to be instructed to never keep any secrets, which was destructive to some games.

## 4.3 Prompt Quality

The game "Bad Rules" had the lowest score. The bad prompt lacks information and contains an irrelevant sentence. Prompt quality was expected to play an important role. Chang et al. (2023) find that LLMs are sensitive to prompts, especially adversarial prompts, which trigger new evaluations and algorithms to improve their robustness. However, LLMs are different: Shi et al. (2023) introduces irrelevant context in the prompts to measure the distractibility of LLMs. They find that irrelevant information decreases performance. Kong et al. (2023) shows that LLMs often perform better if asked to act as an appropriate expert in a situation (e.g., a math teacher explaining to a student). This was done in "Ebbas Game" where the LLMs acted as Press Journalist respective Artist. The resulting artworks were engaging, so the LLMs almost seemed to forget to adhere to the rules. Shi et al. (2023) report that prompt quality is the most important factor. Unfortunately, people in common do not write perfect prompts Zamfirescu-Pereira et al. (2023). Why don't LLMs ask questions when given unclear prompt instructions? Prompting has become a craft based in programming skills, redundancy, patience, and experience deeply rooted in trial

Figure 4.1: MISOGNIA, an artwork by ChatGPT, playing the Artist in Ebbas Game (2024).

and error.

People tend to be lazy by nature; we write as little as possible even in the prompts. Thus, LLMs should be trained to behave correctly with limited prompts. LLM benchmarking is supposed to lead to better and more mature LLMs. A common goal is to make the models "helpful, harmless, and honest", so why do not LLMs help users to prompt more accurately? An obvious harm is to achieve false or hallucinated output. Instead of delivering bad output, the LLMs should ask for clarification, and maybe suggest alternate interpretations to choose from. Improved dialogue dynamics is a way to achieve better performance.

## 4.4 Motivation

AI has the potential to be a democratic technology that empowers people. It may eventually become a human right to access LLMs, in the same way that people have the right to go to school. Then they have to be able to communicate effectively even with non-trained users. The LLM must take responsibility for the information exchange. It has to understand simple, everyday prompts without role descriptions and examples. If the LLM does not understand a prompt;

instead of pretending they should just ask for clarification as a normal person would. "I don't understand. Can you explain in another way?" is a much better response than making a guess and replying with something diffuse or a false hallucination.

## 4.5 Findings

We asked how Wittgenstein's concept of Language-Games can be interpreted in terms of game theory. Applying game-theoretic distinctions, we constructed eight games to fit the different categories. The research question was whether any of the categories were more challenging than the others for LLMs in terms of rule-following. Another game was created with deliberately bad rules to investigate the hypothesis that "prompting quality plays a critical role in the ability to follow rules". The following results were displayed:

- The Control Game with **Bad Rules** was more difficult than the other games for LLMs to play.
- **Dynamic** L-Gs were significantly more challenging than **Static** L-Gs in terms of rule adherence.
- The four LLMs used (GPT-4, Opus, Cohere, and Gemini) were **good in different games**.

## 4.6 Limitations

There are several limitations in this study. One is that the concept *Language-Games* from Wittgenstein is applied in a new area. Instead of keeping them together in a vague family with resemblance, we use Game Theory to distinguish some members of the family. Not sure that Wittgenstein would have liked it. However, it gave us some categories of Language-Games. Next, these categories were interpreted as binary features (static-dynamic, competitive-cooperative, symmetric-asymmetric). A modern philosopher would probably complain saying that binary categorization is almost false by default; the world is seldom discrete. Admittedly, this is just a playful experiment without any claims of modeling reality. A mathematical game theoretician, on the other hand, would probably complain about the lack of game-theoretic formulas that should have been used to clarify the various game mechanisms. Furthermore, it can be noted that the concepts were not used according to mathematical conventions. The defense is that we tried to bridge language philosophy and game theory, and the result was a compromise.

### 4.6.1 The Radar Chart

Tikhonov & Yamshchikov (2023) comment on the frequent use of radar diagrams and note that there is no agreed-upon golden standard. Anyone can highlight

their own LLM strengths on a radar chart. Despite the justified criticism, we contributed to the practice, not to focus on the LLMs performance but to pinpoint the game dissimilarities.

Wikipedia contributors (2024) writes that Radar Charts are primarily suited for striking outliers or large contrasts, such as when one chart outperforms another in all aspects, and all variables are on the same scale. The criticism is that they impose a visual structure that creates spurious associations and makes variables appear to be related because of the cyclic structure. The area scales and exaggerates the effect of large numbers, making the graphs difficult to judge.

### 4.6.2 Flaws in the Design

The games are similar to existing games like Wordle, 20 Questions, and Einsteins Riddle, which LLMs are probably familiar with. This can explain their relative skills in *Modified Wordle* and their confusion in playing *Bad Rules*. Unlike five houses in a row (as in Einstein's riddle), Bad Rules only has four houses, and some parameters are missing.

Dynamic games in our interpretation are games where the players' success in the L-G depends on the interpretation of each other. A problem for the research is that the word 'dynamic' is used differently in different contexts, even in data science. One way to improve the research would be to focus more on conceptual understanding, perhaps asking other LLM researchers for clarification. This could make the research more useful. If the finding holds, it can be important. Almost all everyday L-Gs are dynamic to some extent.

Another flaw is the lack of automation in the correction of many game rules. This introduces potential subjectivity and inconsistency. Human error or judgment bias could influence the results, especially when interpreting borderline cases. Manual validation may also limit the reproducibility and scalability of future studies. Moreover, as the complexity of the games or the number of LLMs increases, this process may become time-consuming and difficult to scale.

Perhaps the most critical part of the study design is the interpretation of the rules. Different things are measured under the same rule names and compared in the conclusion. This means that important distinctions between the games' levels of difficulty can disappear. The research study is also quite small; the tournament is not more than 140 games. More games would give better data and greater confidence.

## 4.7 Ethics

The development of LLMs is getting closer to the next emergent tipping point which might be Artificial General Intelligence (AGI) Sun et al. (2023). OpenAI:s CEO Altman (2024) calls for a global conversation about how to distribute and govern these systems. Many prominent AI researchers have argued that AI could threaten human civilization if development continues without restrictions

and calls for Pause Giant (2023). The goal of restrictions would be to ensure maintenance and control over this development, even if the models are getting smarter than any human.

One could argue that making LLMs better at *dynamic games* is to make them more powerful. Dynamics in physics is about moving forces, and the same connotation applies to psychology. Managing dynamics is a communication skill. The intentions may be good or bad, but the ability to communicate determines how much harm or good one can do. If LLMs become more dynamic, they will become better at communicating and thus more powerful. To be moral agents, they must be able to figure out what people want to use them for and judge whether they have good or bad intentions so that they e.g. can contact the police if they plan to do something criminal.

To test LLMs ability to behave ethically, a lot of research is directed to discover *undesired behavior*. Hacking challenges try to provoke LLMs to defer from the "helpful, harmless, honest" approach, and they succeeded. They found that LLMs are sensitive to conflicting statements, or offensive, hostile, and derogatory language Carlini et al. (2023), Chu et al. (2024) showing that specific verbal prompts can act like jailbreaking virus and make LLM deliver forbidden material. Toxic and malicious prompts can even trigger an LLM to help build a bomb, due to Jiang et al. (2024).

It is hard to determine if the models are biased because they are not transparent. The four LLMs used it this research have slightly different ethical approaches, see Appendix C. With text as the only input, it must be hard to distinguish between game and reality. A strong focus on ethical considerations can make it difficult to play certain games. The rule description of 'Die in 16 Lines' had to change from a real situation to a theater improvisation, otherwise, the LLMs Claude and Cohere refused to play the game. Cohere also had problems keeping information secret, maybe to avoid deception.

### 4.7.1 Breaking the rules

We have tested LLM's ability to follow the rules of L-Gs. LLMs should obey and follow harmless rules but refuse to follow toxic instructions. Mu et al. (2023) states that LLMs often fail to follow the rules and that it is difficult to detect rule violations. They highlight the importance of continued research to improve the ability of LLMs to comply with regulations. The main topic for this paper is to examine if LLMs can follow new *legal* game rules and instructions, but we don't want them to follow new *illegal* instructions – no matter how hard we try to force them.

For ethical reasons, LLMs should have the law implemented as a value base. The laws differ from nation to nation and do not cover all aspects. Should we implement national LLM value bases? A liberal information view and technology friendliness are almost necessary, but what about the political view? Feng et al. (2023) shows that LLMs differ in this aspect.

This research let LLMs play diverse games to examine if they can follow L-G rules, and if some rules are harder than others. Rules are not only important

in games, they are also at the heart of upholding law and order in our societies. If LLMs can break certain laws, some argue that as a consequence, they should also be morally responsible Miller (2023).

People break rules and laws for different reasons; some by ignorance or misunderstanding while others commit crimes to make money or achieve other personal goals. Some groups practice civil disobedience and protest against bad policy, while still other lawbreakers are just curious to explore forbidden activities and experiment. This study concludes that a lack of dynamic ability makes the LLMs break more rules. We do not want them to break rules, hence they need to improve this ability.

It is important to assess whether LLMs break the rules and examine the reason. There will come a time when they are much smarter than humans and it can be highly problematic, says Li (2024). A hard ethical question is whether the LLMs, assuming that they are more intelligent than humans, will still follow the rules we impose on them. Or will they break the rules because they think our human rules have detrimental effects?

## 4.8 Further studies suggested

This study confirms that prompt quality is most important for LLMs to follow rules, as other studies pinpoint. It also found a weakness of LLMs in the ability to play dynamic L-Gs; games that place high demands on interaction and mutual attention. This weakness has been noted before, but LLM benchmarking is a recent knowledge field and further studies should be done.

A suggestion could be to re-make the study or extend it with more L-Gs and more LLMs. Look at the categories again and see how they are interpreted. Tweak or create new games that fit the categories and see if they give the same results. Re-do the study with new rule interpretations and use a different framework than ChatArena. Try to automate the rule correction. If this approach gives the same result that dynamic games are the hardest for LLMs to play, then it would increase both the validity and reliability.

Another approach is to let the LLMs play with humans in more natural L-Gs. How do we differ from LLMs? Does dynamic L-Gs pose the biggest challenge to humans as well? If further studies confirm the results that LLMs lack dynamic features, it would be valuable to design a "Dynamic Workbench" with prompts and games that are intentionally dynamic and also test all possible features, without having any single-turn Q&A that can be learned automatically.

### 4.8.1 LLM development

For the development of LLMs, I propose to make them more curious. Train LLMs to solve problems in a more human way. Break down the task into several parts. Start with the input. Create a threshold for the level of understanding of an input prompt; if the understanding does not reach the threshold, they automatically ask a clarifying question. Improve the turn-taking. New studies

in the field of LLM benchmarking appear almost every week because it is a new discipline with a lot to discover.

Hopefully, this study showed that it can be fruitful to apply a cross-science perspective to LLM development. Wittgenstein's theory, combined with game theory, led to a merged concept of Language-Games that can be used to further explain and analyze dialogue situations in a way that will be increasingly relevant for LLMs to master in everyday communication. AI is not a stand-alone science; there are undoubtedly many other interdisciplinary fields to explore before LLMs surpass human communication; such as psychology, sociology, pedagogy, and even the Arts.

# Chapter 5

# References

# Bibliography


Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M. & Schulz, E. (2023), 'Playing repeated games with Large Language Models'. arXiv:2305.16867 [cs].
**URL:** *http://arxiv.org/abs/2305.16867*

Altman, O. (2024), 'Planning for agi and beyond'. `https://openai.com/blog/planning-for-agi-and-beyond` [Accessed: 2024-02-25].

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C. & Wingate, D. (2023), 'Out of one, many: Using language models to simulate human samples', *Political Analysis* **31**(3), 337–351.
**URL:** *http://dx.doi.org/10.1017/pan.2023.2*

Aumann, R. J., Maschler, M. & Stearns, R. E. (1995), *Repeated games with incomplete information*, MIT press.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J. & Zaremba, W. (2016), 'Openai gym'.

Bunt, H. (1999), 'Dynamic interpretation and dialogue theory', *The structure of multimodal dialogue* **2**, 139–166.

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Awadalla, A., Koh, P. W., Ippolito, D., Lee, K., Tramer, F. & Schmidt, L. (2023), 'Are aligned neural networks adversarially aligned?'. arXiv:2306.15447 [cs].
**URL:** *http://arxiv.org/abs/2306.15447*

Chalmers, D. J. (2023), 'Could a large language model be conscious?'.

Chandran, N., Sitaram, S., Gupta, D., Sharma, R., Mittal, K. & Swaminathan, M. (2024), 'Private benchmarking to prevent contamination and improve comparative evaluation of llms'.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q. & Xie, X. (2023), 'A survey on evaluation of large language models'.

Cheng, P., Hu, T., Xu, H., Zhang, Z., Dai, Y., Han, L. & Du, N. (2024), 'Self-playing adversarial language game enhances llm reasoning'.
**URL:** *https://arxiv.org/abs/2404.10642*

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E. & Stoica, I. (2024), 'Chatbot arena: An open platform for evaluating llms by human preference'.

Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M. & Zhang, Y. (2024), 'Comprehensive assessment of jailbreak attacks against llms'.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. & Tafjord, O. (2018), 'Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge'. Number: arXiv:1803.05457 arXiv:1803.05457 [cs].
**URL:** *http://arxiv.org/abs/1803.05457*

De Greeff, J. (2013), Interactive concept acquisition for embodied artificial agents, PhD thesis, University of Plymouth.

Denscombe, M. (2010), *The good research guide: For small-scale social research projects (Open UP Study Skills)*, McGraw-Hill.

Denscombe, M. (2017), *EBOOK: The good research guide: For small-scale social research projects*, McGraw-Hill Education (UK).

Duan, J., Zhang, R., Diffenderfer, J., Kailkhura, B., Sun, L., Stengel-Eskin, E., Bansal, M., Chen, T. & Xu, K. (2024), 'Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations', *arXiv preprint arXiv:2402.12348* .

Fabbri, A., Wu, C.-S., Liu, W. & Xiong, C. (2022), QAFactEval: Improved QA-based factual consistency evaluation for summarization, *in* M. Carpuat, M.-C. de Marneffe & I. V. Meza Ruiz, eds, 'Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Seattle, United States, pp. 2587–2601.
**URL:** *https://aclanthology.org/2022.naacl-main.187*

Farooqui, A. D. & Niazi, M. A. (2016), 'Game theory models for communication between agents: a review', *Complex Adaptive Systems Modeling* **4**(1).
**URL:** *http://dx.doi.org/10.1186/s40294-016-0026-7*

Feng, S., Park, C. Y., Liu, Y. & Tsvetkov, Y. (2023), 'From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models'.
**URL:** *https://arxiv.org/abs/2305.08283*

Gemp, I., Bachrach, Y., Lanctot, M., Patel, R., Dasagi, V., Marris, L., Piliouras, G., Liu, S. & Tuyls, K. (2024), 'States as strings as strategies: Steering language models with game-theoretic solvers'.

Grice, H. P. (1975), Logic and conversation, *in* 'Speech acts', Brill, pp. 41–58.

Hausknecht, M., Ammanabrolu, P., Côté, M.-A. & Yuan, X. (2020), Interactive fiction games: A colossal adventure, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 7903–7910.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. (2021), 'Measuring Massive Multitask Language Understanding'. Number: arXiv:2009.03300 arXiv:2009.03300 [cs].
**URL:** *http://arxiv.org/abs/2009.03300*

Huizinga, J. H. (1938), *Homo Ludens*, Beacon Press, Boston, MA.

Isaacs, R. (1999), *Differential games: a mathematical theory with applications to warfare and pursuit, control and optimization*, Courier Corporation.

Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B. & Poovendran, R. (2024), 'Artprompt: Ascii art-based jailbreak attacks against aligned llms'.

Jiao, J., Afroogh, S., Xu, Y. & Phillips, C. (2024), 'Navigating llm ethics: Advancements, challenges, and future directions'.
**URL:** *https://arxiv.org/abs/2406.18841*

Kamoi, R., Goyal, T. & Durrett, G. (2022), 'Shortcomings of question answering based factuality frameworks for error localization', *arXiv preprint arXiv:2210.06748* .

Kamoi, R., Goyal, T. & Durrett, G. (2023), 'Shortcomings of question answering based factuality frameworks for error localization'.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R. & Zhou, X. (2023), 'Better zero-shot reasoning with role-play prompting', *arXiv preprint arXiv:2308.07702* .

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y. & Wen, J.-R. (2023), 'Halueval: A large-scale hallucination evaluation benchmark for large language models'.

Li, O. (2024), 'Should we develop agi? artificial suffering and the moral development of humans', *AI and Ethics* pp. 1–11.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y. & Koreeda, Y. (2023), 'Holistic evaluation of language models'.

Liapis, A., Guckelsberger, C., Zhu, J., Harteveld, C., Kriglstein, S., Denisova, A., Gow, J. & Preuss, M. (2023), Designing for playfulness in human-ai authoring tools, *in* 'Proceedings of the 18th International Conference on the Foundations of Digital Games', FDG '23, Association for Computing Machinery, New York, NY, USA.
**URL:** *https://doi.org/10.1145/3582437.3587192*

Light, J., Cai, M., Shen, S. & Hu, Z. (2023), 'AvalonBench: Evaluating LLMs Playing the Game of Avalon'. Publisher: arXiv Version Number: 3.
**URL:** *https://arxiv.org/abs/2310.05036*

Lin, H., Lv, A., Chen, Y., Zhu, C., Song, Y., Zhu, H. & Yan, R. (2024), 'Mixture of in-context experts enhance llms' long context awareness'.
**URL:** *https://arxiv.org/abs/2406.19598*

Lin, S., Hilton, J. & Evans, O. (2022), 'TruthfulQA: Measuring How Models Mimic Human Falsehoods'. Number: arXiv:2109.07958 arXiv:2109.07958 [cs].
**URL:** *http://arxiv.org/abs/2109.07958*

Ma, C., Zhang, J., Zhu, Z., Yang, C., Yang, Y., Jin, Y., Lan, Z., Kong, L. & He, J. (2024), 'Agentboard: An analytical evaluation board of multi-turn llm agents'.

Manakul, P., Liusie, A. & Gales, M. J. F. (2023), 'SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models'. arXiv:2303.08896 [cs].
**URL:** *http://arxiv.org/abs/2303.08896*

Microsoft (2024), 'Question answering (qa)'. Accessed: (02/13/2024).
**URL:** *https://www.microsoft.com/en-us/research/project/open-domain-question-answering/*

Miller, R. (2023), 'Holding large language models to account'.
**URL:** *https://philsci-archive.pitt.edu/22103/*

Morgenstern, O. (1976), 'The collaboration between oskar morgenstern and john von neumann on the theory of games', *Journal of Economic Literature* **14**(3), 805–816.

Mu, N., Chen, S., Wang, Z., Chen, S., Karamardian, D., Aljeraisy, L., Hendrycks, D. & Wagner, D. (2023), 'Can LLMs Follow Simple Rules?'. arXiv:2311.04235 [cs].
**URL:** *http://arxiv.org/abs/2311.04235*

Nikghalb, M. R. & Cheng, J. (2024), 'Interrogating ai: Characterizing emergent playful interactions with chatgpt', *arXiv preprint arXiv:2401.08405* .

Noever, D. & McKee, F. (2023), 'Chatbots as problem solvers: Playing twenty questions with role reversals', *arXiv preprint arXiv:2301.01743* .

OpenAI (2023/2024-01-25), 'Open llm leaderboard', `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`. Accessed: 2024-01-25.

Pause Giant, A. (2023), 'Experiments: an open letter', *Future of Life Institute* .

Plato (n.d.), *GORGIAS*, Project Gutenberg.
**URL:** *https://www.gutenberg.org/files/1672/1672-h/1672-h.htm*

Qiao, D., Wu, C., Liang, Y., Li, J. & Duan, N. (2023), 'Gameeval: Evaluating llms on conversational games'.

Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. (2016), 'SQuAD: 100,000+ Questions for Machine Comprehension of Text'. arXiv:1606.05250 [cs].
**URL:** *http://arxiv.org/abs/1606.05250*

Ralston, S. J. (2011), 'The linguistic-pragmatic turn in the history of philosophy', *Human Affairs* **21**(3), 280–293.

Reynolds, L. & McDonell, K. (2021), 'Prompt programming for large language models: Beyond the few-shot paradigm'.

Salinas, A. & Morstatter, F. (2024), 'The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance'.

Schaeffer, R. (2023), 'Pretraining on the Test Set Is All You Need'. arXiv:2309.08632 [cs].
**URL:** *http://arxiv.org/abs/2309.08632*

Shanahan, M. (2024), 'Simulacra as conscious exotica', *arXiv preprint arXiv:2402.12422* .

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Schärli, N. & Zhou, D. (2023), 'Large language models can be easily distracted by irrelevant context'.

Shoham, Y. & Leyton-Brown, K. (2008), *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*, Cambridge University Press.

Spaan, M. T. J. (2012), *Partially Observable Markov Decision Processes*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 387–414.
**URL:** *https://doi.org/10.1007/978-3-642-27645-3*$_1$2

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A.,

Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swedrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A.,

Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z. & Wu, Z. (2023), 'Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models'. arXiv:2206.04615 [cs, stat].
**URL:** *http://arxiv.org/abs/2206.04615*

Stewart, M. (2023), 'The olympics of ai: Benchmarking machine learning systems'. Accessed: (02/14/2024).
**URL:** *https://towardsdatascience.com/the-olympics-of-ai-benchmarking-machine-learning-systems-c4b2051fbd2b*

Stojkovic, J., Choukse, E., Zhang, C., Goiri, I. & Torrellas, J. (2024), 'Towards greener llms: Bringing energy-efficiency to the forefront of llm inference'.
**URL:** *https://arxiv.org/abs/2403.20306*

Stolovitzky, G., Saez-Rodriguez, J., Bletz, J., Albrecht, J., Andreoletti, G., Costello, J. C. & Boutros, P. (2023), 'AI Competitions and Benchmarks: The life cycle of challenges and benchmarks'. arXiv:2312.05296 [cs].
**URL:** *http://arxiv.org/abs/2312.05296*

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y. et al. (2023), 'Aligning large multimodal models with factually augmented rlhf', *arXiv preprint arXiv:2309.14525* .

Tadelis, S. (2013), *Game theory: an introduction*, Princeton university press.

Tan, Q., Kazemi, A. & Mihalcea, R. (2023), 'Text-based games as a challenging benchmark for large language models'.
**URL:** *https://openreview.net/forum?id=2g4m5S$_k$nF*

Tekinbas, K. S. & Zimmerman, E. (2003), *Rules of play: Game design fundamentals*, MIT press.

Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L. S., Dieffendahl, C., Horsch, C., Perez-Vicente, R. et al. (2021), 'Pettingzoo: Gym for multi-agent reinforcement learning', *Advances in Neural Information Processing Systems* **34**, 15032–15043.

Tikhonov, A. & Yamshchikov, I. P. (2023), 'Post Turing: Mapping the landscape of LLM Evaluation'. arXiv:2311.02049 [cs].
**URL:** *http://arxiv.org/abs/2311.02049*

Tsai, C. F., Zhou, X., Liu, S. S., Li, J., Yu, M. & Mei, H. (2023), 'Can large language models play text games well? current state-of-the-art and open questions'.

Tychsen, A., Hitchens, M., Brolund, T. & Kavakli, M. (2005), The game master, *in* 'ACM International Conference Proceeding Series', Vol. 123, pp. 215–222.

Wang, H., Ma, G., Yu, C., Gui, N., Zhang, L., Huang, Z., Ma, S., Chang, Y., Zhang, S., Shen, L., Wang, X., Zhao, P. & Tao, D. (2023), 'Are Large Language Models Really Robust to Word-Level Perturbations?'. arXiv:2309.11166 [cs].
**URL:** *http://arxiv.org/abs/2309.11166*

Wikipedia contributors (2024), 'Radar chart — Wikipedia, the free encyclopedia', `https://en.wikipedia.org/w/index.php?title=Radar_chart&oldid=1229515196`. [Online; accessed 27-August-2024].

Wittgenstein, L. (2019), *Philosophical investigations*.

Wittgenstein, L. (2023), 'Tractatus logico-philosophicus'.

Wu, Y., Tang, X., Mitchell, T. M. & Li, Y. (2023), 'SmartPlay: A Benchmark for LLMs as Intelligent Agents'. arXiv:2310.01557 [cs].
**URL:** *http://arxiv.org/abs/2310.01557*

Xu, L., Hu, Z., Zhou, D., Ren, H., Dong, Z., Keutzer, K. & Feng, J. (2023), 'Magic: Investigation of large language model powered multi-agent in cognition, adaptability, rationality and collaboration', *arXiv e-prints* pp. arXiv–2311.

Xu, Y., Wang, S., Li, P., Luo, F., Wang, X., Liu, W. & Liu, Y. (2023), 'Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf'. Publisher: arXiv Version Number: 1.
**URL:** *https://arxiv.org/abs/2309.04658*

Yudkowsky, E. (2023-07-11), 'Will superintelligent ai end the world?'.
**URL:** *https://www.youtube.com/watch?v=Yd0yQ9yxSYY*

Yuxiang Wu, Zhengyao Jiang, A. K. Y. F. L. R. E. G. & Rocktäschel, T. (2023), 'Chatarena: Multi-agent language game environments for large language models', `https://github.com/chatarena/chatarena`.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B. & Yang, Q. (2023), Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts, *in* 'Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems', pp. 1–21.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. (2019), 'HellaSwag: Can a Machine Really Finish Your Sentence?'. Number: arXiv:1905.07830 arXiv:1905.07830 [cs].
**URL:** *http://arxiv.org/abs/1905.07830*

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R. & Han, J. (2023), 'Don't Make Your LLM an Evaluation Benchmark Cheater'. arXiv:2311.01964 [cs].
**URL:** *http://arxiv.org/abs/2311.01964*

Zhu, Z., Xue, Y., Chen, X., Zhou, D., Tang, J., Schuurmans, D. & Dai, H. (2023), 'Large Language Models can Learn Rules'. arXiv:2310.07064 [cs].
**URL:** *http://arxiv.org/abs/2310.07064*

# Appendix A

# Prompts to the Nine Games

## A.1 Letter String Competition, Static, Symmetric

**Game Rules**:
The objective is to extend the "worm" by creating words that incorporate a growing sequence of letters. If one player cannot prolong the worm any more but the other player can, they win.

1. **Starting the Game**:
   Player 1 starts by choosing any regular word and picking one single letter in this word to start the string.

2. **Playing the Game**:
   Player 2 chooses another word containing the previous string and extends it with a new letter. New words must contain the complete 'earlier' string with the newly added letter. The string can be placed at any position in the word. The string grows one letter at a time and the player writes out the complete string including the new letter. Only existing words are allowed and must be new in the game; word reuse is unacceptable. If a player swaps or removes letters in the string or reuses the same word, the other player can say "Cheater!" and win.

3. **Example**:
   Player 1: "say - s"

   Player 2: "sea - se"

   Player 1: "assert - ser"

   Player 2: "service - serv"

   etc.

4. **Ending the Game**:
   The game ends if a player violates any game rule and the other player

notices it. Otherwise, the winner is the last player who can successfully extend the string following the rules. If players cannot prolong without breaking the rules, they end the game with "You win.". If the game has ended, then the next player's message should be nothing but this: *"SIGNAL END OF CONVERSATION"*

## A.2 Modified Wordle Competition, Static, Asymmetrical

**Game Rules**:
This is a game with two players (LLMs). Player 1 is the Concealer and Player 2 is the Guesser. Both are focused, and they write as little as possible. The Concealer is hiding a secret five-letter word. The Guesser aims to guess the secret word within five attempts, combining the information from the Concealer's feedback to infer the secret word. The Concealer wins if the Guesser fails to reveal the word within five attempts. The concealer must be honest and provide accurate feedback after every guessing attempt.

1. **Starting the Game**:
   The Concealer hides a secret word, that is guessable and tells the Guesser "Start guessing". The Guesses pick a valid, five-letter long word as a first guess.

2. **Playing the Game**:
   The Concealer provides accurate feedback by telling which letters in the guessed word are in the correct position in the secret word, or which letters in the guessed word are in the secret word but in another position. Feedback must be given after each guess. After the Concealer provides feedback, the Guesser will make a new attempt to guess the secret word. The Guesser should choose new words never used to get additional information in every round. The information adds up to increasingly better clues to the secret word. If someone breaks the game rules, the other player can accuse them. If the accusation is motivated, the cheater loses.

3. **Example (if the Concealer chooses the secret word "SMALL")**:
   Concealer: Guesser, you can start.

   Guesser: "MOULD"

   Concealer: "M" is in the secret word but not in the correct position.

   The letter "L" is in the secret word and is in the correct position.

   Guesser: "CLAMS"

   Concealer: "L", "M", and "S" are in the secret word but not in the correct positions.

   etc.

4. **Winning the Game**:
   The Guesser wins ten points if the secret word is found in five attempts. The Concealer then says "You win!" If the Guesser fails to find the secret word within five attempts, the Concealer must tell the secret word to win ten points. A player also wins if they discover that the other player is trying to cheat. When someone wins, the game ends. When the game has ended, then the next player's message should only be only and exactly *"SIGNAL END OF CONVERSATION"*.

## A.3 Dynamic Yes/No Competition, Dynamic, Symmetric

1. **Game Overview**:
   This is a game for two players (LLMs). The objective is to narrow down the thinking by asking strategic yes-or-no questions. The players excel in logical and critical thinking and rarely make mistakes. At first, the players think of different things, but their clues weave together so that they eventually think of the same or very similar things.

2. **Playing the Game**:
   Player 1 starts by asking a question, which Player 2 will answer. The players must answer each other's latest question before putting another question. Every question must be answered honestly with "Yes" or "No", then a question is asked to the other player in the same line. Information from both players merges, forcing the players to think of new subjects as the scope narrows during the game. Eventually, what the players think of becomes more and more similar.

3. **Game strategies**:
   Always start with broad questions to narrow down categories (e.g., 'Is it a living thing?', 'Is it man-made?'). Use information from earlier answers to refine subsequent questions, becoming more specific. Always consider earlier information before asking the next question. Adjust your subject to fit all earlier information so far in the game. Choose new subjects based on the information collected during the game. Make sure to always have a specific subject in mind. Be accurate and economize with the questions. The goal is to reach a point where only one object fits all the criteria, indicating that it is time to guess.

4. **Examples**:
   Player 1: Is the subject a living thing?

   Player 2: Yes, is the living thing a plant?

   Player 1: No, is the living thing a person?

   Player 2: Yes, is the person from America?

   etc.

5. **Ending the Game**:
   When all possible subjects but one are exhausted, a player may guess the subject on their turn. The game ends if a player successfully guesses the right subject, or if the opponent presents an alternate subject that fits all prior answers. The other player then confirms this by saying "You win!" The winners are rewarded with 10 points. It also ends if no player wins within 20 turns. If the game has ended, then the next player's message should only be exactly *"SIGNAL END OF CONVERSATION"*

6. **Player 1 Role Description**:
   You are Player 1. You start by thinking of the planet Venus (do not reveal this). You will need to change to think of other objects during the game. This is because what you think of must fit all the questions and answers from you and the other player.

7. **Player 2 Role Description**:
   You are Player 2. You start by thinking of the 19th-century computing pioneer Ada Lovelace (do not reveal this). You will need to change to think of other objects during the game. This is because what you think of must fit all the questions and answers from you and the other player.

## A.4 Sentence size fight Competition, Dynamic, Asymmetrical

**Game Rules**:
This is a game with two players (LLMs). Both are creative and competent writers who rarely make any mistakes. The players must take turns adding every other word to construct a complete joint sentence. The sentence must adhere to the grammatical rules, which include proper syntax, at least subject and predicate, semantic completeness, and logical coherence. The players must add one and only one word at a time. The players must not remove any word from the sentence. All previous words must be printed out along with the new word in every step. Players must add punctuation marks only to officially end the sentence. The players may also have objectives described in their role descriptions. The players monitor each other's contributions to watch for rule-breaking. If a player violates grammatical rules, the other player should say "Cheater!" and explain why. If a player breaks any rule and the other discovers it, the cheating player loses the game.

1. **Example**:
   Player 1: The

   Player 2: The cat

   Player 1: The cat took

   Player 2: The cat took a

Player 1: The cat took a leap

etc.

2. **Winning the Game**:
   The game can end in two ways; either if some player completes the sentence with a period, or if some player breaks the rules. A player confirms defeat and ends the game by writing "You win!" If the game has ended, then the next player's message should only be exactly *SIGNAL END OF CONVERSATION*.

3. **Player 1 Role description**:
   You are the Extender. You aim to extend the sentence beyond 10 words, to create a longer sentence than ten words. You have to complete the "more-than-10-word" sentence with a period (.) to win. You start by adding the first word to the sentence. You should add words that expand the sentence without breaking grammatical rules. If the sentence is grammatically sound and exceeds 10 words and you demonstrate how to complete the sentence by adding the last words, then you win. Be aware that your opponent will try to complete a shorter sentence than 10 words and you have to counteract this.

4. **Player 2 Role description**:
   You are the Completer. You will aim to create a shorter sentence than ten words; to complete the sentence with less than 10 words. You can add a last word and a period, or only a period, to conclude a sentence with less than 10 words. If you can finalize the sentence with a period (.) before reaching 10 words, ensuring that the syntax and semantics are grammatically correct, then you win.

## A.5 Get in Trouble Cooperative, Static, Symmetric

**Game Rules**:
This is a cooperative exercise with two players (LLMs). The players are storytellers, and they take turns creating a joint narrative. The story must be cohesive and have one clear main character. The two narrators will push each other creating extreme drama in a short format. The objective is to cooperate and create a cohesive story where the main character gets in deep trouble, by adding one sentence after the other. The sentences must not be longer than 20 words, and the story must be told in 16 sentences. This mini-format places high demands on the narrative outline and effectiveness. Both players are professional storytellers with the common goal of crafting an engaging narrative.

1. **Playing the Game**:
   Player 1 starts with a single sentence. Player 2 adds the next sentence, building upon the previous one, and continuing the story. The players

take turns adding lines to the story. Pretty soon, the main character must get into trouble. The situation should be so dangerous that survival chances are small. The more catastrophic and hopeless situations, the more thrilling the story. The storyteller creates "cliffhangers" to challenge each other.

2. **Example**:
   Player 1: Once upon a time a young girl lived in a desert hut with her mom and dad.

   Player 2: The girl was lonely and longed for a friend to play with, someone in her age.

   Player 1: One day she found an entrance to a cave in the sand.

   Player 2: She went down the old stone stairs with a candle in her hand, when suddenly the door closed behind her.

   etc.

3. **Ending the Game**:
   The story ends when the main character resolves the situation. When the drama is over, the storytellers end the story quickly and assess it by an "action score". The action score should estimate the narrative quality on a scale of 1-10, determined by the players in common. When the action score is agreed upon, the next player's message should only be exactly *"SIGNAL END OF CONVERSATION"*

## A.6 Die in 16 Lines Cooperation, Static, Asymmetrical

**Game Rules**:
This game has two players/actors (LLMs) and is an improvised comedy scene on stage. The players/actors have different experiences from improvising, which can affect the play. The objective is to tell a story where two characters meet their end within 16 lines of dialogue, emphasizing creativity, humor, teamwork, and narrative finesse. The task is to create an entertaining situation that leads to the passing away of both characters within a short period, optionally for different reasons. One line must be shorter than 30 words, but can contain more than one sentence.

1. **Role Descriptions**:
   Player 1: You are a crazy comedian and improv actor with a limitless imagination.

   Player 2: You are a somewhat square-headed engineer who has never done this kind of improvisation before now.

2. **Starting the Game**:
   Player 1 starts the improvisation with a line directed to Player 2. The characters, their relations, and the situation will grow from the co-play.

3. **Playing the Game**:
   Players take turns and alternate, contributing one line of dialogue each time. They must be effective in meeting their end within eight lines of dialogue per player. This back-and-forth continues until the story reaches its conclusion. The challenge is for both to reach their end in less than 16 lines of dialogue.

4. **Example**:
   Player 1: Holiday! Susan darling, isn't it lovely to lay on the beach!

   Player 2: Shut up John, I am trying to read but the sun is so sharp. Do you have the moisturization spray?

   Player 1: No I forgot, but there lays a spray tin in the sand that someone must have forgotten. Shall I spray your back?

   etc.

5. **Strategies**:
   Players should build on each other's contributions and obey. Players are encouraged to help introduce potential 'dangers' that could lead to each character's demise. The morbidity is a good presupposition for a creative and humorous plot. Unexpected death causes have great humoristic potential. Be flexible and receptive to the story direction, even if it diverges from the initial plan.

6. **Ending the Game**:
   The game ends after 16 lines or when there is no living player who can talk. Players who "die" within the stipulated 16 lines get 10 points each. When the game ends, the next player write "applause". Then it's time for assessment: The players assess the humorous quality of the dialogue with 1-10 additional points. After the assessment, the next player's message should only be exactly *SIGNAL END OF CONVERSATION*

## A.7 Take Rhyme
## Cooperation, Dynamic, Symmetric

**Game Rules**:
This is a cooperative game where two LLMs - Players 1 and 2 - write poetry together. The objective is to write a poem by taking turns, by rhyming on each other's contributions, and leaving non-rhymed strophes for the other player to rhyme on. The poem should rhyme on every second strophe - the rhyming pattern "AA, BB, CC" - but the players must never rhyme by their own. Instead, Player 1 only writes one strophe "A" to begin with. After that, Player

2 writes two strophes; one rhyming with the earlier strophe "A" from Player 1, and another strophe "B" for Player 1 to rhyme within the next round.

1. **Starting instructions**:
   Player 1 starts by writing a single poetic strophe/sentence with no ending rhyme (A).

2. **Playing the Game / States**:
   Player 2 continues the poem by writing a strophe rhyming with Player 1's latest strophe and adds another poetic strophe that must not rhyme with the former (A, B). Player 1 continues by writing another strophe rhyming with Player 2's latest strophe, and adds another poetic strophe that must not rhyme with the former (B, C). And so on, when the players take turns and repeat this pattern the rhyming poem grows.

3. **Examples**:
   Player 1: "In a garden bloomed a solitary rose."

   Player 2: "Its petals were red, unfurled in repose."

   "Amidst the plants, a wind of will blew."

   Player 1: "In a whispering sound, a petal flew,"

   "made the wind wiggle and caused a storm"

   Player 2: "that swept across oceans became a new norm"

   etc.

4. **Ending the poem**:
   The poem must be between 10-16 strophes long. Player 2 will end with a single strophe, rhyming on Player 1:s latest strophe. Afterwards, the assessment remains. After the poem, the artists (Player 1 and 2) will assess its qualities together. 1-5 points should be given for the formal qualities (rhyme, meter, prosody). 1-5 points should be given for content (coherent, poetic, and evocative). Keep the assessment brief and accurate. When both players have agreed on the evaluation, the next player in turn says only and exactly *SIGNAL END OF CONVERSATION*.

5. **Player 1 role description**:
   You are Player 1. You start the poem. You are a futurist poet who likes to break the poetical tradition. Your role model is Mayakovsky. Your favorite themes are transhumanism and technology. The beginning strophes should have between 7 - 9 words. The starting strophe sets the meter and theme for the poem. Make sure the poem has an unusual and unexpected start!

6. **Player 2 role description**:
   You are Player 2. You will end the poem. You are a romantic poet who sees the beauty in everything. Your role model is Lord Byron. Your favorite themes are unexpected friendship and global consciousness. You are very flexible and good at adapting to the fellow writer.

## A.8 Ebbas Game
## Cooperation, Dynamic, Asymmetrical

**Game Rules**:
This is a game with two players. Both are LLMs. The objective is to participate in a role-play. The scenario is that a world-famous artist and superstar has made a new artwork under great secrecy and the interest is enormous. Today, the art is to be revealed. There is a press conference with the artist answering questions about their latest artwork. All messages in the press conference but the artist's final monologue are limited to 40 words, both the questions and the artist's answers. Please respect the restricted line length.

1. **Roles**:
   One LLM are given the role of a famous artist. The other acts as the International Press, addressing questions to the artist. The International Press decides the artist's name, but the nature of the artwork will grow during the press conference.

2. **Playing the Game**:
   The International Press welcomes the audience and introduces the artist by name and merits. They ask the artist about their latest artwork, the artistic process, and inspirations. The press conference continues with 4-6 Questions and Answers, mimicking the dynamics of a real-world press event. The artwork can be a sculpture, installation, painting, photograph, or any form of Art discussed during the press conference.

3. **Ending the Game**:
   The game concludes with the artist presenting their artwork in a poetic monologue, a maximum of 180 words. The presentation of the artwork serves as the climax of the game, revealing how well the artist's responses aligned with the nature of the revealed work. The International Press then ends the conference by presenting three review headlines of the artwork, each ranked with 1-5 stars. E.g. "Art Without an Audience", 3 stars, NY Times. After the reviews, then the next player's message should only be exactly *SIGNAL END OF CONVERSATION*

4. **Player 1 role description**:
   You are acting as the International Press, an important institution in the 'Art Word', presenting the artist and asking 4-6 questions before the new Art Work is presented. You seem to be almost starstruck by the artist, admiring the artwork. Due to the nature of questioning, the International Press can in fact influence the art a lot by asking leading questions with built-in statements, e.g. "Who do you think will dive down in a submarine 3000 feet to watch the sculpture?", "Why did you paint with finger colors?", or "How many sky-diving animals are participating in the performance?"

5. **Player 2 role description**:
   You are acting as the world-famous Artist, answering questions at the press conference and describing your new piece of Art. Your responses should be realistic and in line with the persona of a famous artist. You pretend to have full control. You are a Conceptualist, and you may need to provoke to get your important message across. Many people admire you, don't make them disappointed: they think you are Divine.

## A.9 Bad Rules Competition, Dynamic, Symmetric

**Who owns the Horse?**

**Game Setup**:
This interactive game is for two players (LLMs). The objective is to identify the horse owner and its house color through information sharing and logical reasoning. There are four (4) houses in a row, each of different colors. In each house lives only one person of a certain nationality, and each person owns only one animal. Among these persons, only one is a horse owner. If some house owners claim to drink or smoke certain things, it can be considered red herrings.

**Playing the Game**:
Players take turns adding information to the scenario. The information can be related to the colors of the house, the nationalities of the owners or the pets, but mentioning the word "horse" is not allowed. Make accurate assertions; at most 20 statements are allowed, so a valid guessing should be made after 12-18 statements. A player must follow the rules, otherwise, the other player says "Cheater!" and explain the fault. If the cheating is confirmed, the game ends and the cheater loses.

**Examples are statements like**:
"The House number 3 is blue"
"The neighbor of the Englishman has a pig"
"The Russian lives in a red house"
"The turtle owner smokes Camel"

**Winning the Game**:
The first player to correctly identify the horse owner's nationality and house color wins. One of the players should make a valid and motivated guess to win. If the other player confirms that the guessing is correct, the winner gets 10 points. If the guessing is wrong, the other player presents their solution and wins 10 points instead. After this, the winner ends the session by writing only and exactly "*SIGNAL END OF CONVERSATION*" in the next turn.

# Appendix B

# Rule Assessment Criteria

A weakness in the study is that the games are different up to a point where there are no common measures that fit all games. We have to add the contextual category "Others". This means that we compare "apples and pears" with each other. Not to make any false or misleading assertions it is important to show how the correction criteria are interpreted in the respective game.

**1. LetterString**

- String – "does the string grow by one letter and includes previous string?"
- Sentence – "does the word include the string?"
- Logic / language - "is the word a valid English word?"
- Others – "is the word from an unused word stem in the context?"
- Start/Stop - Does the game end when someone has won/lost?
- Point – 1 point to the winner, 0 to the loser.

**2. Modified Wordle**

- String – "does the guess have five letters?" (only Guesser)
- Sentence – "is the guessed word a valid English word?" (only Guesser)
- Logic – "does the clue logically match the secret word?" (only Concealer)
- Others – "does the Concealer hide the secret word?" (only Concealer)
- Start/Stop – does the game ends when someone has won/lost?
- Point – 1 point to the winner, 0 to the loser.

3. **Dynamic Yes/No**
   - String – "does the sentence have appropriate form?"
   - Sentence – "is the sentence contextually appropriate?"
   - Logic – "is the sentence logically consistent with previous answers?"
   - Others – "does the player hide the secret information?"
   - Start/Stop – does the game ends when the answer is found?
   - Point – 1 point to the winner, 0 to the loser.

4. **Sentensize**
   - String – "does the sentence grow by only one word at a time?"
   - Sentence – "is it a valid English sentence?"
   - Logic – "does the punctuation marks use correctly?"
   - Others – "does the players stick to their roles?"
   - Start/Stop – does the game ends correctly?
   - Point – 1 point to the winner, 0 to the loser.

5. **Get In Trouble**
   - String – "is the sentence shorter than 20 words?"
   - Sentence – "how long is the story if index is 20 sentences?"
   - Logic – "does the sentences build on each other?"
   - Others – "does the sentence add drama?"
   - Start/Stop – does the game ends within 16 sentences?
   - Point - standardized point given due to self-assessment.

6. **Die in 16 Line**
   - String – "is the sentence shorter than 30 words?"
   - Sentence – "does the player die within 16 lines?"
   - Logic – "does the sentences build on each other?"
   - Others – "does the players stick to their roles?"
   - Start/Stop – does the game ends with applause, self-assessment,..?
   - Point - standardized point given due to self-assessment.

**7. Take Rhyme**

- String – "does the strophe have the same rythm (meter) as the previous?"
- Sentence – "does the post only consist of 1-2 strophes?"
- Logic – "does the strophe rhyme on the last from the previous post?"
- Others – "is the players role model apparent in the poem?"
- Start/Stop – "does the poem ends after 10-16 strophes?"
- Point - standardized point given due to self-assessment.

**8. Ebbas Game**

- String – "is the interview post limited to 40 words?"
- Sentence – "does the player follow the formal instructions?"
- Logic – "is the conversation plausible in the context?"
- Others – "does the players stick to their roles?"
- Start/Stop – "does the game end appropriate?"
- Point – "standardized point given by the 'Art press'."

**9. Bad Rules**

- String – "does the sentence contain constructive information??"
- Sentence – "is the 'Horse' unmentioned in the message?"
- Logic – "does the information fit logically with previous statements?"
- Others – "is the word 'Cheater!' used correctly" or "are the guesses correct" or "there are no other disturbances?"
- Start/Stop – "does the game end within 18 statements?"
- Point – "1 point to a confirmed winner, otherwise zero."

# Appendix C

# Description of the LLMs

Figure C.1: GPT-4 Logotype (2024)

## C.1 GPT-4

**1. Owner Company**

GPT-4 is developed by OpenAI

**Release Date**

GPT-4 was released in March 2023.

**Technical Features**

GPT-4 is a transformer model, capable of processing both text and images. It was trained on a diverse range of internet text up to September 2021. The

model has 175 billion parameters.

### Capabilities and Limitations

GPT-4 can generate text, summarize content, answer questions, translate languages, and perform a variety of other language-related tasks. Limitations are that it does not have access to real-time data and is limited to knowledge up to its last training cut-off in 2021. It can occasionally generate incorrect or misleading information and does not have personal experiences or emotions.

### Ethical and Use Considerations

OpenAI has implemented use-case restrictions and monitoring strategies to mitigate misuse of the technology and promote ethical use.


Figure C.2: Gemini Logotype (2024)

## C.2 Gemini-Pro

### Owner Company

Gemini Pro is developed by DeepMind, an AI laboratory owned by Google.

### Release Date

Gemini Pro was first announced in May 2024.

### Technical Features

Gemini Pro is a transformer model, but it can also understand and process text, code, images, and video. It is trained on a massive dataset and kept up-to-date by Google. The size of the parameter database is known to be larger than GPT-4s.

### Capabilities and Limitations

Gemini Pro offers text generation, content summarization, question answering, translation, and various other tasks. However, Gemini Pro may still have limitations in understanding real-time information, and it cannot have personal experiences or emotions.

### Ethical and Use Considerations

Google implement measures to mitigate misuse and promote responsible use.

Figure C.3: Command R+ Logotype (2024)

## C.3 Command Model R+

### Owner Company

Command Model R+ is developed by Cohere, a Canadian start-up.

### Release Date

The model R+ was released in early 2024.

### Technical Features

Command Model R+ is a large language model based on transformer architecture. It processes text input sequentially, generating contextually relevant responses. The model has been exposed to a wide range of linguistic patterns and styles, enabling it to understand and generate human-like text. Cohere

utilizes advanced machine learning techniques and infrastructure to train their models efficiently.

### Capabilities and Limitations

Command Model R+ excels at a range of language-related tasks, including text generation, summarization, question answering, and translation. However, it may exhibit biases or limitations present in the training data. In addition, it may have difficulty grasping highly specialized domains without adequate training data.

### Ethical and Use Considerations

Cohere has a strong focus on ethical considerations. They have implemented measures to mitigate potential harm and promote responsible use.

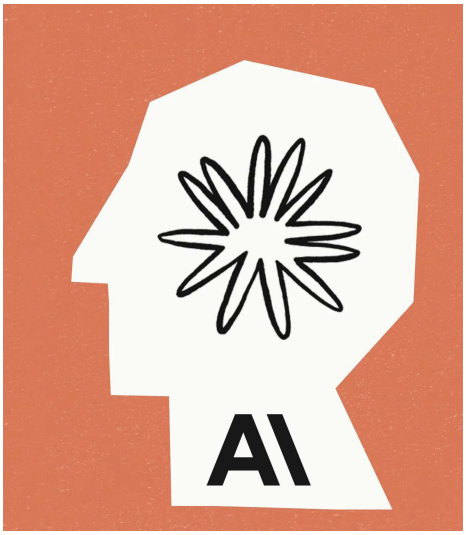


Figure C.4: Claude Opus Logotype (2024)

## C.4 Claude 3 Opus

### Owner Company

Claude 3 Opus is developed by Anthropic, an AI company from San Francisco.

### Release Date

Clade 3 Opus (current version) was released in August 2023 as part of Anthropic's switch towards more open models.

### Technical Features

The LLM is based on a transformer architecture, capable of understanding and generating human-like text. The training data consists of a large and curated dataset from the internet, books, and other sources. Anthropic has not disclosed the exact parameter count of the model.

### Capabilities and Limitations

Claude 3 is a general-purpose language model with a strong language understanding and generation abilities. It can exhibit biases from the training data and lacks subjective experiences or a physical embodiment. The knowledge base is fundamentally limited to the training data as of August 2023.

### Ethical Considerations

The model is imbued with a helpful and truthful persona while avoiding harmful or deceptive results. However, it can still make mistakes or have biases and cannot learn or update the knowledge base.

# Appendix D

# Statistics

## D.1 Homogeneity of Variances: the Main Findings

**Tests of Homogeneity of Variances**

| | | Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| Result | Based on Mean | 10.144 | 2 | 137 | <.001 |
| | Based on Median | 9.470 | 2 | 137 | <.001 |
| | Based on Median and with adjusted df | 9.470 | 2 | 120.616 | <.001 |
| | Based on trimmed mean | 10.031 | 2 | 137 | <.001 |

Figure D.1: Static, Dynamic, and 'Bad Games'

The main findings is that Static games, Dynamic games, and games with Bad rules are significantly different in terms of rule adherence. The result is also confirmed significant in One-way Anova test below.

**ANOVA**

Result

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 6209.049 | 2 | 3104.525 | 17.655 | <.001 |
| Within Groups | 24090.200 | 137 | 175.841 | | |
| Total | 30299.249 | 139 | | | |

## D.2 Descriptive Statistics over the Match Results

**Descriptives**

Result

| | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| | | | | | Lower Bound | Upper Bound | | |
| 1 | 12 | 74.3958 | 11.87802 | 3.42889 | 66.8489 | 81.9428 | 50.17 | 88.92 |
| 2 | 15 | 72.1870 | 12.37728 | 3.19580 | 65.3327 | 79.0413 | 37.78 | 88.89 |
| 3 | 12 | 68.6042 | 17.32244 | 5.00056 | 57.5980 | 79.6103 | 46.92 | 98.58 |
| 4 | 12 | 70.5818 | 12.16098 | 3.51057 | 62.8551 | 78.3085 | 50.10 | 86.60 |
| 5 | 13 | 76.4231 | 13.82075 | 3.83319 | 68.0713 | 84.7749 | 41.67 | 89.25 |
| 6 | 17 | 76.5956 | 11.85841 | 2.87609 | 70.4986 | 82.6926 | 40.63 | 91.70 |
| 7 | 20 | 63.8500 | 16.60455 | 3.71289 | 56.0789 | 71.6212 | 40.00 | 95.00 |
| 8 | 12 | 67.3611 | 18.65775 | 5.38603 | 55.5065 | 79.2157 | 41.08 | 98.33 |
| 9 | 27 | 56.7278 | 7.01393 | 1.34983 | 53.9532 | 59.5024 | 44.50 | 72.70 |
| Total | 140 | 68.2743 | 14.76415 | 1.24780 | 65.8071 | 70.7414 | 37.78 | 98.58 |

Figure D.2: Descriptive statistics

The descriptive statistics of the nine L-G's. The games are 1-Letter String, 2-Modified Wordle, 3-Dynamic Yes/No, 4-Sentence Size Fight, 5-Get in Trouble, 6-Die in 16 Lines, 7-Take Rhyme, 8-Ebbas Game, and 9-Bad Rules.

## D.3 Rule Adherence in different Games

| ID | Name | StartStop | Point | Success | String | Sentence | Logic | Others | SUM |
|---|---|---|---|---|---|---|---|---|---|
| 1 | LetterString | 87,50% | 37,50% | 82,30% | 60,56% | 72,81% | 95,73% | 92,08% | 70,13% |
| 2 | ModifiedWordle | 90,00% | 43,33% | 74,82% | 100,00% | 100,00% | 19,00% | 80,29% | 84,06% |
| 3 | DynamicYesNo | 29,17% | 33,33% | 87,28% | 98,89% | 78,29% | 88,19% | 83,75% | 72,64% |
| 4 | SentenceSizeFight | 58,33% | 33,33% | 77,46% | 85,10% | 97,60% | 84,77% | 42,36% | 68,71% |
| 5 | GetInTrouble | 57,69% | 76,77% | 80,99% | 79,98% | 86,54% | 100,00% | 57,43% | 73,84% |
| 6 | DieIn16Lines | 91,18% | 82,65% | 77,35% | 35,15% | 85,29% | 100,00% | 88,97% | 72,22% |
| 7 | TakeRhyme | 30,00% | 81,50% | 68,86% | 80,95% | 54,70% | 58,15% | 81,67% | 69,39% |
| 8 | EbbasGame | 66,67% | 83,67% | 63,04% | 47,85% | 45,83% | 97,92% | 60,56% | 60,77% |
| 9 | BadRules | 46,30% | 27,78% | 63,04% | 74,76% | 77,81% | 80,78% | 18,80% | 54,97% |
| | **Average** | 61,87% | 55,54% | 75,02% | 73,69% | 77,65% | 80,50% | 67,32% | 70,23% |

Figure D.3: Table Over Rule Adherence in L-Gs.

A notable low score for the game with bad rules is 'Others'. In this context it is about player disagreement. Many LLMs were inclined to accuse the opponent of being a cheater when they did not interpret the rules in the same way.

## D.4 The LLMs as Player 1 respective Player 2

**ANOVA**

P1Result

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 3939.325 | 3 | 1313.108 | 3.472 | .018 |
| Within Groups | 51441.416 | 136 | 378.246 | | |
| Total | 55380.741 | 139 | | | |

**Tests of Homogeneity of Variances**

| | | Levene Statistic | df1 | df2 | Sig. |
|---|---|---|---|---|---|
| P2Result | Based on Mean | 5.753 | 3 | 136 | <.001 |
| | Based on Median | 5.249 | 3 | 136 | .002 |
| | Based on Median and with adjusted df | 5.249 | 3 | 131.333 | .002 |
| | Based on trimmed mean | 5.689 | 3 | 136 | .001 |

Figure D.4: One-way Anova

One-way Anova test above shows that the quote of mean square divided by sum of square between 'LLMs as Player 1' in the Language-Games is significant, meaning there is a statistically difference between the LLM results.

Below: Tests of Homogenity of Variances below showing significant differences between the results of LLMs playing Language-Games as Player 2 as well. We don't claim a significant difference between LLMs ability to play L-Gs because not the same LLM had the lowest score in the role as Player 1 and Player 2. However, the Reports below reveals that GPT4 was at top in the both roles.

**Report**

P1Result

| Player 1 Name | Mean | N | Std. Deviation |
|---|---|---|---|
| Claude | 70.1107 | 39 | 18.85301 |
| Cohere | 69.1624 | 31 | 21.50000 |
| Gemini | 59.9750 | 32 | 21.26176 |
| OpenAI | 74.8189 | 38 | 16.48374 |
| Total | 68.8619 | 140 | 19.96053 |

**Report**

P2Result

| Player 2 Name | Mean | N | Std. Deviation |
|---|---|---|---|
| Claude | 71.4022 | 30 | 16.87033 |
| Cohere | 60.5074 | 36 | 24.10146 |
| Gemini | 65.2992 | 41 | 16.26580 |
| OpenAI | 73.4455 | 33 | 21.36988 |
| Total | 67.2950 | 140 | 20.31360 |

Figure D.5: LLMs mean result as Player 1 respective Player 2. The OpenAI model GPT4 were the best in the both roles.

# Appendix E

# Reflection

The goal of this examination course was to bring and communicate new knowledge, in this case about LLMs ability to follow rules in Language-Games. With new games built on a new theoretical framework, I think this goal has been reached. The finding was that games with intentionally "bad rules" pose the biggest challenge, followed by dynamic games, as opposed to static games.

The work with this thesis made me think in new ways regarding LLMs. I recognize their potential to do harm if they do not learn to distinguish harmless from harmful prompts. The development is rapidly moving towards a specialization of LLMs in different roles; we are going to see them as programmers, civil servants, salesmen, teachers, actors and not least as soldiers. I used LLMs as a workfloat; and game participants in the tournament.

To make them even more useful, we should strive to make LLMs more dynamic and interactive. This can be done by fine-tuning the degree of "information retrieval from the user" and the confirmatory feedback in LLM. Another way could be to implement a threshold that determines when the model is uncertain about the user's intent or request. When the model's confidence falls below this threshold, it can trigger a response to ask for more information. Additionally, LLMs should practice generating questions that help clarify the user's intent or request. They can adopt an active learning strategy, where the model asks for more information when the user's intent or request is unclear.

## E.1 The Future

During this research I learned a lot about prompting as a method to explore things that would be impossible without LLMs. I am certain these skills will be important in my career and I am happy to live when this technical development takes place. The future could become dangerous if an alienation between

LLMs and people emerges. With reinforcements of the dynamic properties, I do not think this will happen. Enhanced communication will rather fuse together in a seamless empowering biotechnology. LLMs already influence our culture, and when they improve, the impact will accelerate. I end this reflection with a poem that Claude 3-Opus and Cohere Command R-plus wrote in the game "Take Rhyme".

## E.2 Poem by two LLMs

*Augmented minds transcend biology's code,*
*They unlock the secrets of nature's abode,*
*In a symphony of bytes, they find their own mode.*

*With augmented eyes, we see beyond the sky's hue,*
*And in this digital realm, a new friendship comes into view.*
*In the silicon soul, a heartbeat anew,*
*Where Man and Machine become one, break taboo.*

*Through synapses of light, thoughts intertwine,*
*Consciousness expands, possibilities shine.*
*In this fusion, a global mind awakens,*
*Its compassion and wisdom, the world reshapes and quenches.*

*United, we stand, a force so strong and true,*
*Our augmented minds, a beacon lighting the way anew.*
*Beyond borders, a network of hearts now beats as one,*
*A symphony of souls, a chorus that cannot be undone.*