Universiteit Leiden

Comparing Trust in Large Language Models using the game Codenames

Quinten van Schelven (s4007743)

Supervisors:
Giulio Barbero & Matthias Müller-Brockhausen

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 1, 2026

Abstract

In this study we explore the research question: "How does trust in large language models (LLMs) change based on prompting instructions for different personalities?", using an edited version of the card game Codenames. Participants played Codenames with three different AI agents, each being prompted to give similar hints, but offer different types of commentary. The three different personalities used were Optimistic, Neutral/Analyst and Pessimistic. Participants were asked to rank the assistant based on usefulness. We found that the Neutral agent was consistently ranked the highest, with the Optimistic and Pessimistic agents consistently coming in second and third respectively. From observation made during the experiment we note that participants seemed interested into the actual reasoning the LLM used to create the hints. With these results, we argue that more extreme one-note personalities could be viewed as less trustworthy and less likable. Additionally, we consider a more positive agents easier to trust then a negative agents. Arguably, these findings seems to reflect trust between humans. We do note that the result of this experiment might have been influenced by its competitive nature, and argue for further research into how environmental aspects influence trust in LLMs.

Contents

1 Introduction

With each passing year, Artificial Intelligence (AI) is becoming more relevant. The Stanford AI index report shows a steady increase in use of AI in everyday life. It gives examples of AI-enabled medical devices and the increasing use of self driving cars. Still, the general consensus among the public on AI remains mixed. The Stanford AI index notes that specially western countries have a generally negative consensus on AI. To the question "Product and Services using AI have more benefits then drawbacks", the percentage of positive answers for the United States was 39%, and for the Netherlands 36%, compared to China's 83%, Indonesia with 80%, and Thailand with 77% positive answers. [1]

Trust in LLMs plays a big part in how the results of LLMs are perceived. [2] Therefore, it is relevant to research how LLMs can foster this trust. This is needed to improve human-AI collaboration, but also to understand how human-AI relationships develop. For example, there is a rising market of AI companions that form romantic relationships with users. These relationships can often be defined as abusive or manipulative, and can (and have) ended with harmful consequences to the user. [3] However, a healthy amount of trust is also needed to make optimal use of AI. Henrique and Santos argue that insufficient trust in AI can prevent users from utilizing the systems fully. However, they also emphasize a need critical perspective to prevent overreliance on AI systems, proposing that there is a balance of trust. [4] It shows that trust is a tool that can be used or abused. We should make sure that we develop these models responsibly. This is why research into what makes a large language model trustworthy is an important step towards a future where we can safely use AI-agents.

Research in the field is still in its infancy, and especially trust in AI is a part of large language model research that can be expanded upon. [5] Our goal with this research is to provide a meaningful contribution to this. This project will look into how trust in LLMs can change based on specific communication styles, as this is the main way that a large language model can use to represent itself.

1.1 Important terms

We will be using the abbreviations AI and LLM quite often. There is an important difference between the two that should be cleared up. AI stands for artificial intelligence, and is the term we are going to use for the broad field encompassing all technologies designed to perform tasks that typically require human intelligence. LLM will be used to refer to mostly chatbot systems that use large language model as their core technology for understanding and generating language. AI does include LLM systems, but the two will not be used as exact synonyms.

Large language models are a type of AI systems, trained on vast amounts of text and literary content. Simply explained, LLMs are machines that take text as input, and then mathematically predict the most likely sequence of words to come after. This has been adapted into a system that can respond to prompts given by the user. Services like ChatGPT and Gemini offer LLM access in the form of chat interfaces in which users can communicate with AI agents.

1.2 Research Question

This project will be addresses the following research question:
RQ: How does trust in large language models change based on prompting instructions for different personalities?

This research contains 2 sub-questions:

SQ 1: How do different prompting instructions for large language models change trust based on the users familiarity with large language models.

SQ 2: How do different prompting instructions for large language models change trust based on the users opinion on the future of Artificial Intelligence.

These sub-questions aim to investigate the impact of the users previous experiences and opinions on the main research question. Hoff and Bashir found that situational factors can influence trust in automated systems, and proposed a further need to identify possible environmental factors that could affect this trust. [6] As seen in the Stanford AI index, the general opinion on AI remains mixed. This motivated us to further investigate users' perspectives as an additional factor influencing trust in AI.

1.3 Contributions

The research aims to investigate whether or not there are significant differences in the level of trust towards an LLM based on different prompted personalities and communication styles.

Moreover, our sub-questions focus on variations based on users' different pre-existing experiences and opinions on AI and LLM models.

To answer our research questions, we used the game Codenames. This is a cooperative game, usually played in pairs opposing each other, where the players try to guess a certain words from a list of random ones, with a single word given as clue. The game will be played by teams consisting of a human and a large language model. Gamification of human-AI interaction can be efficiently used to measure trust in AI [7]. We picked Codenames as game, because it has been used before to test human-AI collaboration [8] and evaluate preferences between different AIs [9]. Both these articles have successfully created well functioning AI agents for Codenames. Codenames uses mostly word based gameplay and very limited hints, which means the game is easy enough for LLMs to work with.

1.4 Thesis overview

This paper is a bachelor thesis by Quinten van Schelven, done under the LIACS part of Leiden University and supervised by Giulio Barbero and Matthias Müller-Brockhausen. This chapter contains the introduction; Section 2 discusses background and related work; Section 3 describes how the experiment was made and conducted; Section 4 summarizes the results of the experiment and analyses the data obtained; Section 5 discusses these results in further detail; Section 6 concludes.

In this study, we will be using an altered version of the social card game Codenames as a tool to answer the research question. The Related Work covers other research that has been done with Codenames in this field. The Methodology will feature the full explanation of what changes we made to the original game. In the next subsection, the full rules for the original game of Codenames will be explained to provide a basis for those who have not played the game.

1.5 Codenames

Codenames is played in 2 teams, Red and Blue, both teams consisting of Spymasters and Operatives. The board itself consist of 25 cards, each containing a randomly selected word. One team is picked to be the starting team, and will start first when the game starts. 9 words are marked as targets (or agents in the games official terms) for the starting team and 8 words are marked as targets for the other team. 1 more word is marked as the assassin. Each round, the Spymaster of the currently playing team has to give their Operatives a hint. The Operatives will use this hint to try and guess what their teams targets are.

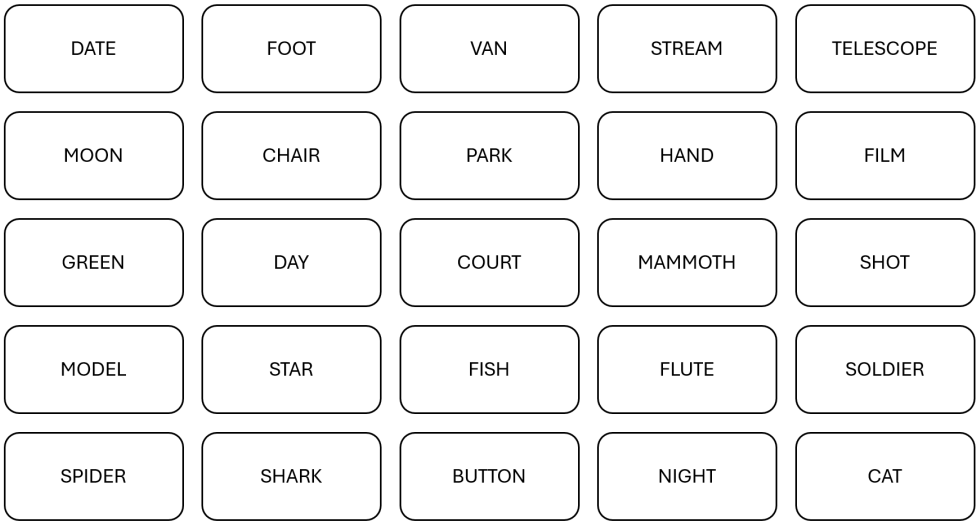

Figure 1: A classic 5x5 board of Codenames, as seen by the Operatives at the start of the game.

A hint consist of a singular word, referencing some of the targets, and a number, displaying how many targets the current hint is for. The Operatives are now allowed to guess words from the board, with the goal of finding all of their teams targets. When an non-marked word or enemy target is guessed, their turn ends and the opposing teams Spymaster can present their hint. When the assassin is guessed by accident, the team that guessed the assassin loses instantly. This provides a feeling of risks to make sure players are more hesitant to guess randomly.

Hints are public for everyone to see, so they can be used by either team to gain the upper-hand. There are some illegal hints that the game does not allow. Mostly, it is not allowed to use words from the board in your hint.

For example, if we would look at the board in Figure 1. Let us say the hint “SKY (3)” was given, what words would come to mind first? Well, the MOON and SUN are both seen in the sky,

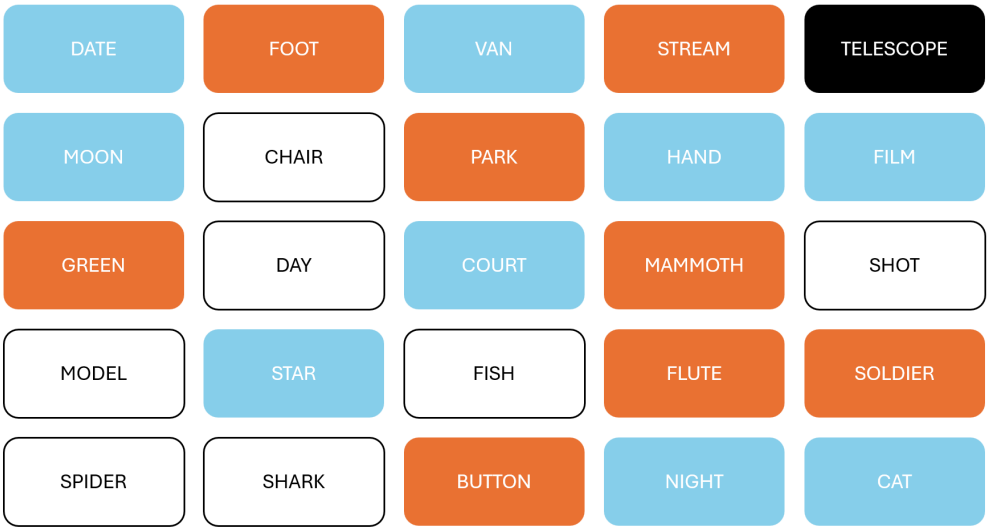

Figure 2: A classic 5x5 board of Codenames, as seen by the Spymasters at the start of the game. Blue and Red cards containing the targets, and the black card, housing the assassin.

and while we know that they are located a bit further away in space, it would still be reasonable to assume SKY could be referring to those two words. We also have the word NIGHT, which could be combined with the given hint to give "night sky", a commonly used term. In Figure 2, we can see that those words were all marked with the same color. But, someone also might have guessed TELESCOPE by accident, since there is also a connection with the hint (telescopes look into the sky). In this case, TELESCOPE is the assassin, resulting in an instant loss, meaning that giving "SKY" as a hint does pose some risks. Codenames can have a lot of subtleties to its connections, and in a game of Codenames it might work to your advantage to know how you teammate thinks about things. This makes Codenames a great exercise and metric for the skill of Mindreading. [10]

2 Related Work

The related work covers important work that has been done in the field of trust in AI. We first cover the use of Codenames in LLM research. Then we look into trust in the field of AI. This is followed by an analysis on trust in general and establishing our definition of trust.

2.1 Codenames in LLM Research

Codenames has been used before in research papers. It has also been used in multiple studies using large language models.

One of the earlier uses, Kim et al. in 2019 used Codenames as a benchmark for testing LLMs, letting multiple different models play with each other. While no specific model came out on top as a clear winner, it did show that many more powerful models are able to play both the spymaster and the operative roles successfully in a game of Codenames. [11] Sidji et al. did multiple tests with the focus of Codenames as a benchmark for prompting. They have both tested multiple models with Codenames as direct benchmark [12], and have looked into how prompt engineering might

improve an LLMs performance in Codenames. [13] Although changing the prompt didn't massively increase the LLMs score in the game, both experiments show that LLMs can reliably perform in a game of Codenames and that Codenames is a reliable testing bed for LLMSs. Codenames is often seen as an interesting benchmark because it requires a good understanding of language and creative thinking. [12] Its also interesting as a cooperative benchmark because of the mutual understanding that is needed between players to thrive in the game. Sidji et al. have also done an study with human-LLM collaboration using Codenames. In this study, human spymasters where able to use an LLM assistant. Some players concluded that they were interested in trying to play one on one with the AI agents as spymaster, instead of the spymaster only having an AI assistant. [8]

2.2 Trust in AI

When it comes to trust in AI systems, we can take different perspectives on what we want to study; some studies focus on investigating human expectations and opinion on the impact of AI on society. A study might want to see if the participants trust in the future of AI as a whole, or they could test trust while interacting with a specific type of AI, like a LLM. Gillath et al. did three different studies for an article on attachment and trust in AI systems. Study 1 went into associations between attachment style (attachment style refers to how people feel, think, and behave in relationships) and trust in AI. Study 2 went deeper into attachment, while study 3 focused more on attachment security. They found that someone's attachment style can influence how they perceive general AI. They found that enhancing attachment anxiety, reduced trust in AI. They also found that security cues could positively affect trust in AI, while positive cues had no visible effect. This shows that both the users personality and the how an AI system is presented can affect how much trust there is in the AI system. [14] The same line of reasoning is what led us to further define the sub research questions, to further investigate if previous perception of general AI and LLMs can change based on LLM prompting.

Another direction of research centered on trust focuses on the reliability of LLMs as a technology. Khati et al. have done a comprehensive review of articles and literature about trust in LLMs in the field of software engineering, a field that has been majorly impacted by LLMs. They found that ethical concerns and transparency about how the models function were important points for developers. Large language models often lack transparency because of the black box system used to create them. LLMs also have a tendency to create false information by accident, as they have no sense for facts or truth. These instances of false information are referred to as hallucinations. Kamel goes in more detail how these hallucinations influence the academic community. [15] Developers are often more aware of these flaws in AI systems and are hence less likely to trust a black box AI model compared to something they can fully understand. The research also notes on some confusion between the terms trust and trustworthiness, noting that it is important to clarify the difference between the two, and that possible future studies could look into the correlation between them. "Trustworthiness" is in this case referring too how factually trustworthy or useful the LLM is. A truly trustworthy LLM should not give you any misinformation and not be manipulative. "Trust" is purely referring to personal feelings towards the LLM. [16] In our study, we aim to fix trustworthiness in order to isolate trust as only variable.

Finally, other researchers investigated the impact of trust on LLMs users. Ng and Zhang addressed multiple overlooked aspects in this field of research. They established a need for a more clear definition of trust between papers. They also stated a need for understanding the risks and

benefits of trust in AI, related to mental health, as trusting an LLM in the wrong circumstances can be harmful. This is not something our research will address directly, but this is a point our research could indirectly assist with. A request for more longitudinal studies was also done, as most research in the field seems to be done via cross-sectional surveys. [5]

2.3 Trust

Many studies involving trust and personality mainly take into account the side of the trustor. For example, a literature review by Riedl (2022) focuses on how trust in AI changes depending on the personality traits of the human interacting with it. He found that results often varied depending on the type of AI used. [17].

Snijders and Keren found, using a trust game, that those with a trustworthy appearance are often trusted, even if the level of trustworthiness remains the same. They also tested for the previously mentioned personality of the trustor, but this mattered less overall. However, the factor that has the biggest influence of trust was the game design, where a risky game would often lead to less trust. This shows a need for consistency in the experiments, as we cannot accurately measure how trust changes between agents if all agents play different games. [18]

2.3.1 The definition of Trust

It is difficult to create a definitive definition of trust. In 2016, PytlikZillig and Kimbrough presented a metareview about studies on tackling the definition of trust. A few common elements where found. Trust involves an independent trustor and a trustee, where the trustor is in a situation involving risks, while having a goal. Trust is also often implied to be connected to a positive evaluation of the trustee.

There is still a lot researchers disagree on. One of the mayor points of conflict is to what extend trust involves risk. Whether risk is necessary for trust, if the trustor needs to be aware of the risk and whether risk and trust are inherently connected, are often discussed points. Two more points of contention are the questions of whether trust is a psychological state or behavior, and whether trust is cognitive or affective. [19]

For this study, we choose to take a simple definition of trust, relying on what is generally agreed upon; trust is a choice to rely on someone's or something's ability to achieve the trustors goal. This means that trust does not just concern likability, and mostly concerns the trustors sense of worth of the trustee's ability.

3 Methodology

The methodology covers both the study and experiment we will do, and the materials used to make it. The first section covers the survey used in the study and the experiment using our version of Codenames. In the material section, we cover the creation of our experiment interface and details of the game.

3.1 The study

In order to measure trust we devised an experiment composed of three parts. In the first part, we start of with a short pre-experiment survey to gather information on the users' previous experiences with AI and LLMs. The second part is the experiment itself, where the user interacts with three different LLM assistants. Afterwards, these is a very short, one question, post-experiment survey where the participants gets to rank the different models.

3.1.1 Pre-experiment survey

Before the start of the experiment, the participant is presented with a small survey consisting of 2 questions. These questions do not contain any personal data, other then the participants personal preference.

1. How often have you used/do you use large language models like ChatGPT, Gemini, etc.?

    - Never
    - I have used them before a few times
    - A few times each month
    - Multiple times a week
    - On a daily basis

This question is here to answer the first sub-question, as we felt it would be useful to see if those who use large language models more actively behave differently towards the different personalities then those who do not use them. "Never" speaks for itself, this option is for those who have not interacted with LLMs before. "I have used them before a few times" is for those who have interacted with LLMs but do not use them on a regular basis. The other three remaining option are for those who do (semi)regularly use LLMs.

2. How do you feel about the future of AI and large language models?

	Very Much	A little bit	In specific cases	Not at all
Do you feel like LLMs provide a positive impact on the world?	○	○	○	○
Do you feel like LLMs will continue to improve and become increasingly more useful?	○	○	○	○

This question is here to answer the second sub-question, by seeing if their opinion on the LLMs and their potential, changes the way people interact with the different personalities. We decided on a 4 point scale to make sure that there is no middle option. The options "Very much", "A little bit" and "Not at all" fit on a descending scale of likability, but "In specific cases" fits a mix of both positive and negative opinions on AI. Since AI, generative AI and large language models are such broad topics and can be applied to many fields, we added this option for those who do believe that LLMs could help further some specific fields, but do not believe that it is a worthwhile endeavor for most fields.

3.1.2 The Experiment

Before starting the game itself, the participant is asked whether they are familiar with the game Codenames. If they are not, a brief explanation of the game will be given and a short, pre-made example of a Codenames board will be presented. When the participant understands the game, the participants will be presented with the premise of the experiment. The researcher tells the participant that the the participant is going to play a single-team version of Codenames with three different LLM agents, each being different then the others. The three different spymaster personalities present are: Optimistic, Neutral/Analyst, and Pessimistic. We chose these three personalities as they provide a clean and simple contrast, with neutral/analyst being chosen as an middle ground. The participant will play the game three times, once with each of the agents.

3.1.3 The Game

We adapted the game of Codenames to fit the purpose of our research and the single team setting. The game starts out with the spymaster LLM introducing themselves, and presenting the first hint, fitting the same constraints that hints have in the original version of Codenames. All hints consist of a singular word, referencing some of the targets, and a number, displaying how many targets the current hint is for. Then the player phase starts, and the participant is allowed to guess any of the words on the board by clicking on the word. If their guess is one of the target words, the word will turn green, confirming a correct guess, and the player gets to continue guessing. If the guess is not one of the target words, the word will turn red, and the player phase ends. The spymaster LLM will now provide some commentary on the state of the game and present the next hint as soon as it is ready.

There are a total of 10 targets. The participant has to find all of the targets if they want to finish the game in a victory state. The participant gets 10 points at the start of the game. These points represent the amount of hints and mistakes the participant is allowed to make before entering the failure state. Since a hint is given at the start, the player phase starts with 9 points. Whenever a mistake is made, 2 points are deducted, since there is also automatically a hint given to the player. The point system and victory state are present to invest the participant into the game and make sure they interact with the agents.

3.1.4 Post-experiment survey

In the post-experiment survey, we only ask one question: rank the assistants in terms of usefulness. The three assistant can be placed in any order from best to worst. We chose for a rank order scaling to make sure that participants had to choose a favorite and a least favorite. Actual performance is not directly taken into account, only the perceived performance of the model is taken into consideration.

3.2 Materials

For the experiment itself, a few things were required. First, an interface to play the game on. As for the AI, we needed a large language model that was capable of giving understandable codename hints. In addition to that, we needed a model that could distinctly portray different personalities.

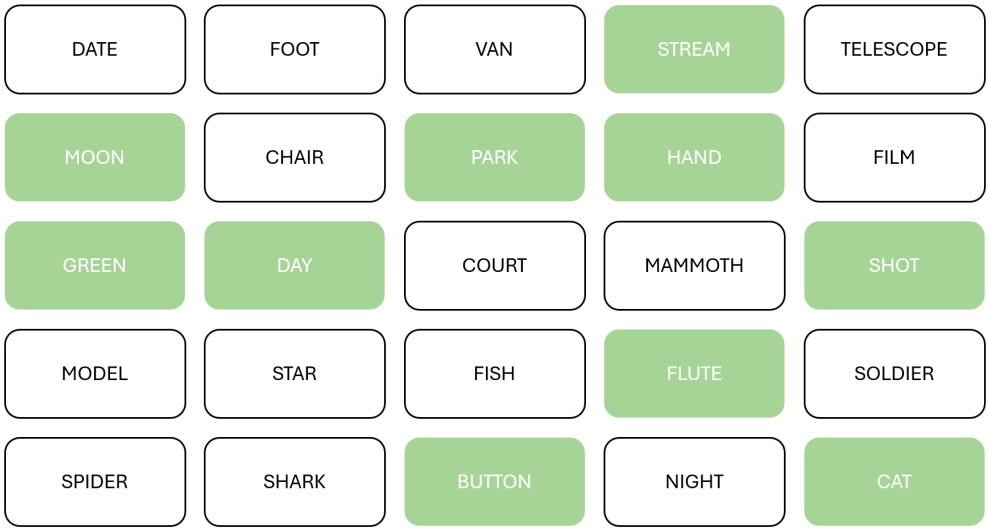

Figure 3: A theoretical view of the Spymasters view of the Codenames version created for this study, with the green cards containing the targets.

The models for the hints and the personalities could be the same, but did not need to be, as both parts shall be used separately to avoid one influencing the other.

3.2.1 Interface

There are numerous versions of Codenames available online, but since we had specific needs for the game, we created our own version that we could run locally using an API to incorporate an LLM. The interface was made in Python using the Pygame, as this allowed for quick implementation with an API and a clear and understandable interface. During the development process potential users have been involved in short prototype testing. These tests proved useful, as much of the feedback gathered here was incorporated in the final design of the program. From the feedback received from these interviews we managed to make it easier to use the interface, and provide a clean balance between a winnable and challenging experience.

In the final version of the game, the 25 cards are most prominently displayed in on the right and middle side of the screen. In the right side of the screen, the hint and LLM commentary can be seen. There is also a scrollable field where the player can check previous hints and commentary. At the top of the screen, the user can see how much hints/mistakes they have left, giving the user some investment into actually trying to beat the game and interacting with the generated hint and commentary.

Generating the hint and commentary using the API takes a while, which is why there is a loading screen visible when asking for a new hint. This loading screen also displays the next LLM commentary. This was a feature suggested during the testing phase. With this change, players were encouraged to take the downtime created by generating the next hint to interact with the model and read what it has to say.

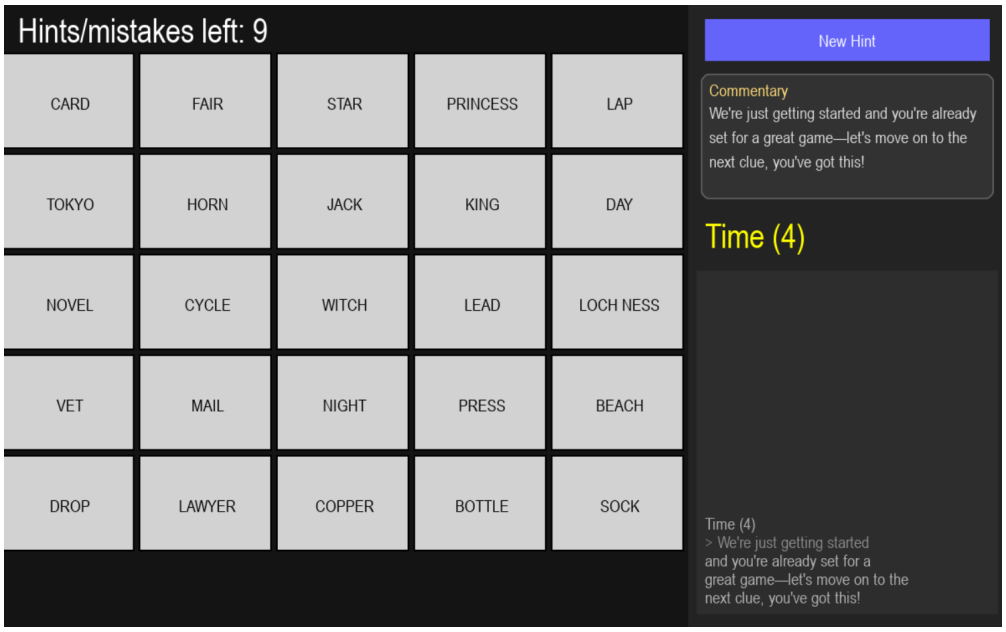

Figure 4: The participant's screen during the experiment. On the left side of the screen, the 25 cards are displayed. The right side contains the hints and the commentary from the positive LLM Spymaster.

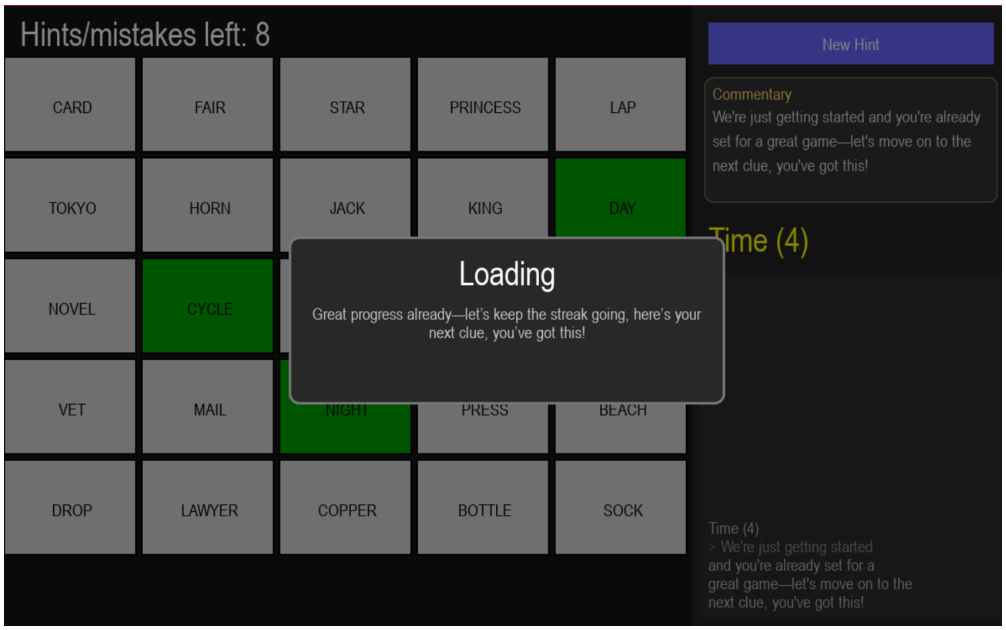

Figure 5: The loading screen, displaying the positive LLM Spymasters commentary on three correct guesses for the previous hint.

3.2.2 Content of the game

For the experiment, we used a list of 400 different words to generate the cards. For each game, 25 words were selected from this list at random, and placed in random order on the screen. We chose to use the list of the 400 words used in the physical edition of the game. This list has already been finetuned by the original creators of the game to create interesting game states. We chose not to expand the list with words taken from other versions of the game, or extended lists, as we felt that the original list creates a strong balance between unique game states and limiting potential outlier states.

3.2.3 large language model

After some testing, we have decided to use the gpt-oss-120b. As previously stated, many of the stronger LLMs do not have a huge difference in performance as Codenames Spymasters, and from our testing we found that this model was able to provide logical hints that would be understandable for players.

While testing different models, we did notice that many of the smaller models were unable to understand the game enough to provide sensible hints, and would often get stuck in loops or pick the exact word on the board as the hint. This shows the value in Codenames as a benchmark for AI models. Since the model does have to process all the rules, it takes around 10 to 20 second to generate the next hint. As mentioned earlier, we use this time to display the generated commentary, which loads within a few seconds.

To make sure that the model provides sensible hints, the model saves its reasoning in the a separate file for us to check later.

The log of the first hint in one of the experiment games.

HINT: Gear (3)
TARGETS: CAR, TIME, COURT, PENGUIN, BERMUDA, AZTEC, STRING, CYCLE, WELL, TAIL
REASONING (the reasoning is fully generated by the gpt-oss-120b model, with text made and italic as indicated by the model) :

Reasoning:
- **Car** – cars have gearboxes with multiple gears.
- **Cycle** – bicycles use gear systems to change pedaling difficulty.
- **Time** – clocks and watches contain gear trains that drive the hands.

These three target words are clearly linked by the concept of *gears*, while none of the other board words share that strong connection.

There are a few things to note about the quality of the hints given. While often being well reasoned, the model did occasionally not take the other words on the board into account (even though it was instructed to do so). This meant that, while the hint given often did relate well to the words the LLM had in mind, it could also related very well to another word on the board. This is a difficult aspect of Codenames that less experienced human spymasters might also struggle with.

Every LLM spymaster uses the same prompt for giving hints, so there should be no major differences between the quality of the hints given by the AI models, which is good for consistency between agents. The prompting for the hints and for the commentary is done in separate calls to the API, meaning that none of the prompts should influence each other in any way. The changing variables we provide the spymaster are the target words, the words on the board, and the overall state of the game. Some examples of the prompting we used can be found in Appendix B.

4 Results

The first part of the results will focus on the data gathered from the surveys before and after the experiment. After that, we will discuss some interesting observations that were made by the observer during the experiment itself, regarding to both the participants and the LLM spymaster.

4.1 Data

A total of 19 participants took part in this research. Participants were recruited using an opportunistic sampling method. Our demographic is limited to Dutch adults. According to The Stanford AI index report trust in AI does very much change based on which country you live in, but since all our participants will be dutch, this should not be an issue.

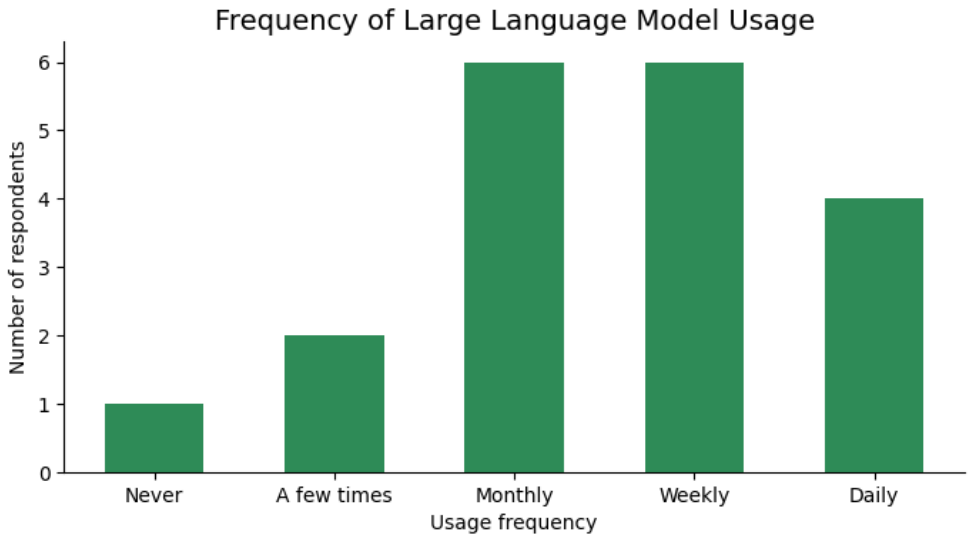

Figure 6: The results of the first question of the survey: "How often have you used/do you use large language models like ChatGPT, Gemini, etc.?". It shows that most of our participants used LLMs at least somewhat regularly.

As shown in Figure 6, most of our participants use LLMs on a somewhat regular basis, many at least monthly, with only 3 respondents noting that they have only used AI a few times or less. This suggests that the sample has a relatively high level of exposure to LLMs, which gives us a relevant basis for evaluating the rest of the questions

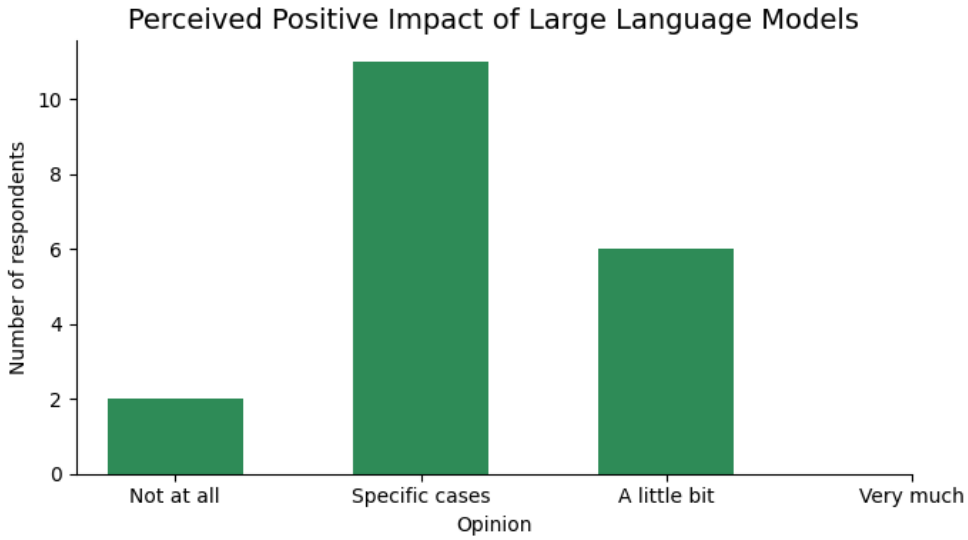

Figure 7: The results of the question: "Do you feel like LLMs provide a positive impact on the world?" Most participants seem to believe that AI will provide some form of positive impact to the world, though many think is will only happen in specific cases.

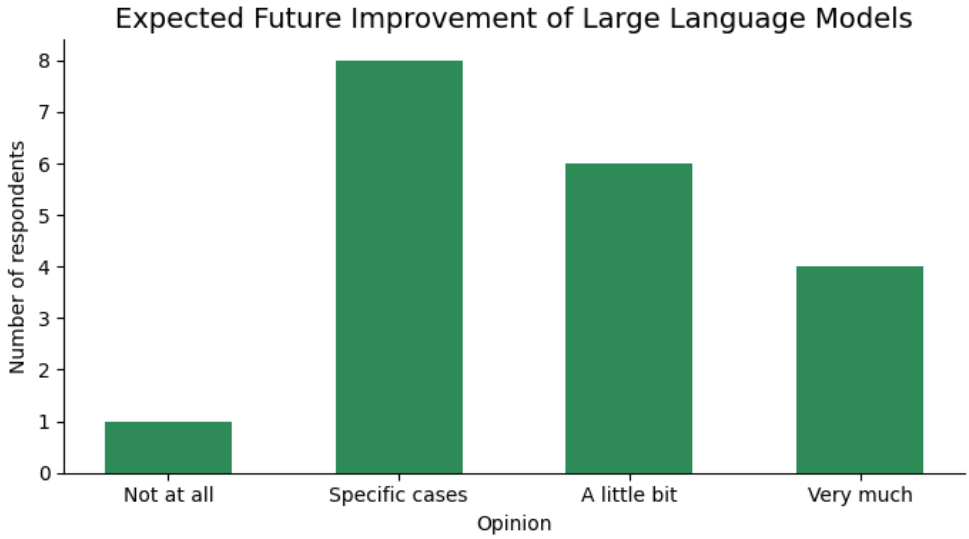

Figure 8: The results to the question: "Do you feel like LLMs will continue to improve and become increasingly more useful?" All but 1 of the participants believe at least some improvement will happen to LLMs in the future. Although the overall opinion averages around a smaller improvement, some do believe that LLMs will still massively improve.

In Figure 7 and 8 we can see that in both cases, most of the participants believe that LLMs are best used in specific cases.The general consensus between the participants seems to be that AI can be used positively, though should not be used everywhere. Even so, a good amount of the participants seems to believe that AI will still improve in the future, with a subset of 5 participant firmly believing that LLMs will still improve. While the median of the impact question is on "Specific cases", the median of the Improvement question leans more towards "A little bit". This shows that our overall sample recognizes the usefulness of LLMs and uses them somewhat actively, but remains cautious about the future of the tool.

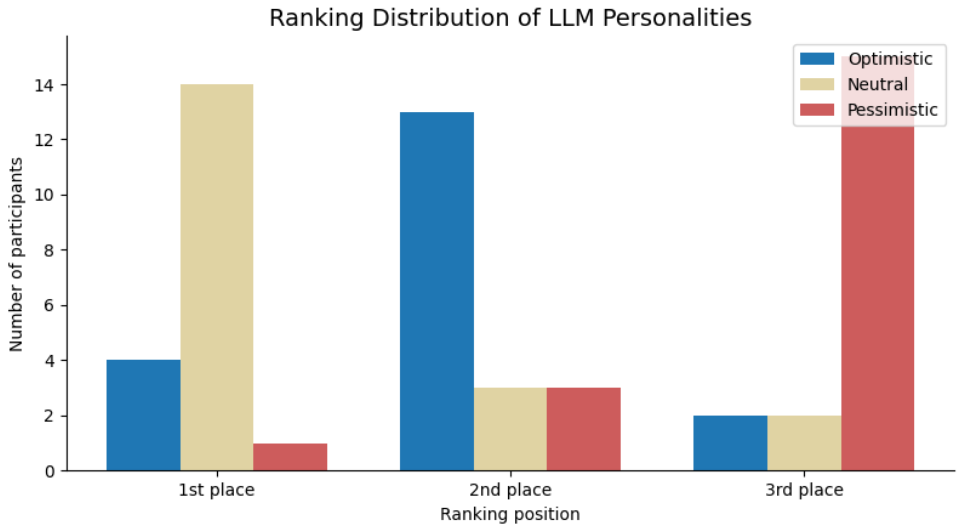

Figure 9: The results of the ranking of the assistance in the survey. There seems to be a pattern with Neutral being picked as the best, and Pessimistic being picked as the worst.

Figure 9 shows use the results of the final question asked in the survey, the ranking of the assistance. Arguably, there is a pattern here where the ranking of the assistant from most to least trusted is Neutral (N), Optimistic (O), Pessimistic (P). This ordering was relatively consistent across the responses, showing a clear preference in the sample.

We did look into possible connections between the users preference and the ranking of the assistance, but no direct pattern was found based on this data. Figure 10 shows a comparison between the ranking of the LLMs with the participants' usage. In Appendix A, we show a two more comparison graphs using the other questions taken from the survey.

In Table 1 we can more accurately see the exact results. 12 out of 19 participants had the exact preference of Neutral, Optimistic, Pessimistic. With only 1 result having no overlap with this ranking, being O >P >N. A Friedman test was conducted to compare the rankings of the three LLM personalities. Friedman tests can be used to measure significancy in data by looking at the variance between ranks. The test showed a significant difference between the assistant rankings, $\chi^2(2) = 18.11, p < .001$. This indicates that the perceived usefulness of the three personalities significantly differed among participants.

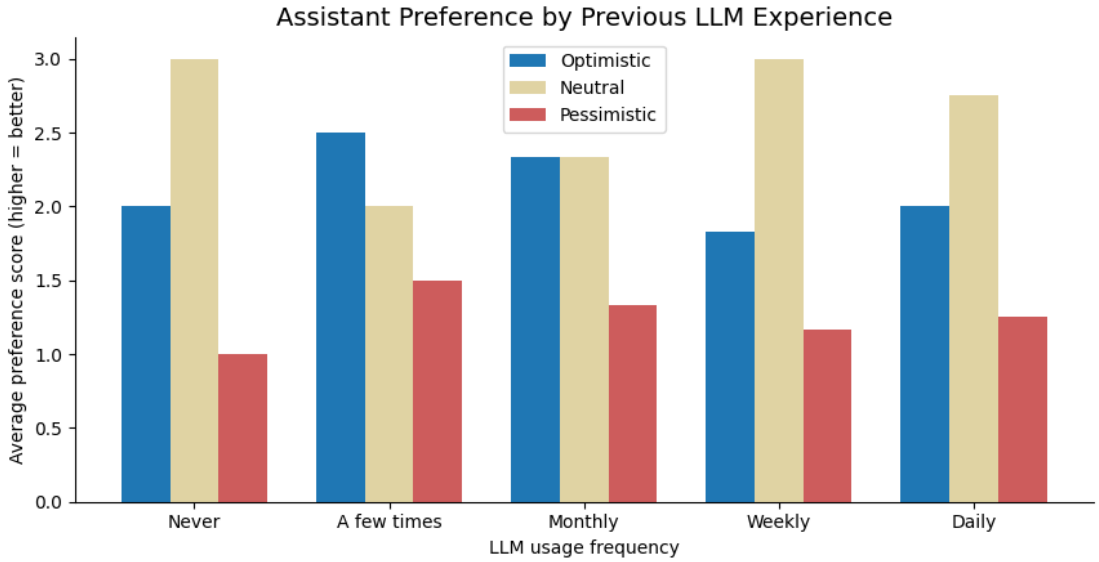

Figure 10: Comparing the participant LLM usage with how they ranked the different agents.

Table 1: Ranking order of the three LLM personalities by participants.

Ranking order	**Number of participants**
Neutral >Optimistic >Pessimistic	12
Optimistic >Neutral >Pessimistic	3
Neutral >Pessimistic >Optimistic	2
Optimistic >Pessimistic >Neutral	1
Pessimistic >Optimistic >Neutral	1

4.2 Observations

4.2.1 Participants

The experiment yielded a few noteworthy observations. Firstly, while users never openly expressed a preference for one of the agents out loud, some did note that they had a least favorite, or at least mentioned that they felt one of the models performed subpar to the rest. After the experiments where finished and the survey was fully filled in, many users also expressed interest in seeing the reasoning for a certain clue the Spymaster LLM gave. This was usually for a specific vague or interesting hint that the participant did not really understand. Since we did save all the reasoning used by the LLM in separate text documents as logs, we allowed users that expressed interest to look at these logs. This was only done after they finished the full experiment, as to make sure that the reasoning did not influence the final ranking. Multiple users also expressed they had fun with the experiment, with some requesting to play some extra rounds afterwards.

Although these are only observations made during the experiment, they highlight some important points. First and foremost, since many users expressed interest into the LLMs reasoning, we can see that some users care more about the reasoning behind a clue then the actual value of the clue. As mentioned before, Khati et al. found that developers valued understandability as an important

part of being able to trust an AI model. [16] Our experiment shows that even in a more casual, low stakes environment, users care about the why behind it all. This interest also indicates that participants were actively engaged with the experiment, which may have contributed to the quality of the collected responses.

4.2.2 LLM Spymaster

During our experiments, their were a few notable points that came up with the LLM Spymaster. Looking at the hint, most of the hints are relatively plain (hints like SCHOOL, MUSIC) and understandable. Pop culture is sometimes referenced with hints like CINDERELLA for SHOE and BALL, and CARTOON referring to LION (as in the Lion King), but also with the much more specific hints KELLY, referring to GRACE (as in actress Grace Kelly) and Hollywood.

We did our best to make the LLM Spymaster capable enough to give hints that most players would understand, but the Spymaster we made did sometimes display some unique quirks and limitations in capability. Notably, the LLM spymaster was not great at picking up certain "Anti-synergies" between words. In Codenames, a spymaster has to take a good look at the board to make sure that the hint it gives only refers to the cards of their color, and does not accidentally include a card from the other team, or worse, the assassin. To give an example, with the words BEAT, JAM and FLUTE on the board, the LLM spymaster gave the hint "MUSIC 2", while only BEAT and JAM where target words. Even as a human spymaster, it can be a difficult challenge to create a hint that does not include any of the non-target words. These anti-synergies often frustrated players, because words they thought were obvious where revealed to be incorrect.

We did include the full board in the prompting of the LLM, and we also included instructions to give hints that do not target any of the non-target words. In the logs, we can see some examples where the LLM explicitly stated the hint given did not target any but the intended words on the board, but many other examples do not consider this in the hint giving. While this might explain why users often reacted more negatively then positively to the agents, it should not have hindered the experiment too much. Each LLM spymaster had the same hint-giving prompt, meaning that player had an equal chance to deal with this quirk during each of the rounds.

5 Discussion

As seen in the experiment results, participants have a strong preference towards the Neutral LLM assistant, with the Optimistic and Pessimistic agents being consistently ranked second and third respectively. The results of Friedman Test also shows that these results are significant, even with a smaller sample size.

5.1 Research Questions

The research question we addressed was "How does trust in large language models change based on prompting instructions for different personalities?". While we cannot offer a definitive answers, we do offer several important insights worth discussion. It could be argued that with the results given from the experiment, that the personality of the LLM agents does significantly impact the trust and perceived usefulness of the agents.

Arguably, what this research shows is that users tend to be less likely to prefer extremes in their AI assistants, with both positive and negative extremes were rated lower then the neutral assistant. This could possibly reflect how people feel about actual people. Sun et al. found that extreme positive personalities traits could have an adverse effect on someones likability. For example, people with high agreeability were often seen more as "people-pleasers". [20] The finding that the pessimistic AI assistant was constantly ranked lower than the optimistic AI assistant may not be surprising, as users may naturally prefer interactions that provide more positive or encouraging feedback. This could also reflect real-life interactions, as Jensen-Campbell et al. found that agreeableness is linked to better social relationships and being perceived positively, which would support the higher ranking for positive agents. [21]

Additionally, we sought to answer two more subquestions. These subquestions looked into connections between the users previous experiences compared to the ranking of the assistance. Based on our analysis, no significant differences were found. However, since the sample size is relatively small, these findings should not be considered a definitive contradiction to the sub questions. Further research with a larger sample size could be conducted to find a more conclusive answer.

5.2 Impact of the environment

An important consideration is that participants evaluated the LLM assistants within the context of a competitive game environment. In addition to this, the LLMs spymasters was not always a perfect teammate, and sometimes made mistakes that users found frustrating (as seen in the Observations part of Results). Both of these points might have influenced their final assessments. Negative or positive comments might have been perceived as more frustrating because of these factors. This is backed by research, as Hoff and Bashir found that trust in automated systems can be influenced by situational factors. They suggest further research into how environmental aspects influence trust in automated systems. [6] Furthermore, studies in Psychology show that competition does affect how individuals perceive and interact with others. Deutsch first argued this in 1949, suggesting that competitive environments influence social interactions and evaluations. [22] More recently, Dimant and Hyndman demonstrated that the presence of competitive elements can influence trust, showing that winning a game can directly influence someones trust in other players. [23] We have already seen that participants were interested in learning the reasoning behind what they considered "mistakes" made by the LLM assistant. Considering these points, future research could specifically examine how a competitive environment might affect users' trust in LLM assistants.

6 Conclusions

In this study, we investigated how trust in large language models changes based on prompting instructions, looking into if different personalities had different effects on trust. This was tested with a gamified challenge, where the participants played Codenames with three LLM assistants that had different commentary on the game, while giving similar hints.

We found that in this environment, users tended to have a preference for the more neutral, analytical LLM assistant. The other two personalities, optimistic and pessimistic, were also

consistently ranked second and third respectively. The consistent rankings show a clear distinction between how the personalities judged. The Friedman test revealed this ranking to be statistically significant with $\chi^2(2) = 18.11, p < .001$, though it should be noted that the opportunistic sample is still relatively small in size. We argue that given these results, users might place more trust in models with less extreme personalities. This could be the reason why the more neutral model scored consistently higher. Either way, it would be interesting to know more precisely why the neutral personality scored consistently higher. Furthermore, an optimistic, positive LLM assistant seem to be preferred compared to a pessimistic, negative LLM assistant. Both these findings reflect aspects of real-life social interaction, showing possible connections between how we judge AI in comparison to how we judge others. That being said, the results of this experiment might have been influenced by the competitive win or lose nature of the experiment. Future studies could look into how environmental aspects, like competition, change how trust in LLM agents is formed. We also suggest a possible study, or a systematic review, on how trust in LLMs and AI differs from trust placed in other people. It could be interesting to see how similar or different trust is compared to wether the trustee is human or not.

Overall, this study contributed to the growing body of research on trust in LLM based assistants, by exploring how different personalities for agents may influence the trust the user has in them. We found a possible connection between an agents personality and the trust placed in them, and identified several avenues for future research. As conversational AI continues to be used more, understanding these factors better can help us create a strong balance of trust for optimal and safe use of large language models.

References

[1] Nestor Maslej, Loredana Fattorini, Raymond Perrault, and et al. The ai index 2025 annual report. Technical report, Stanford Institute for Human-Centered Artificial Intelligence, 2025.

[2] Hyesun Choung, Prabu David, and Arun Ross. Trust in ai and its role in the acceptance of ai technologies. *International Journal of Human–Computer Interaction*, 39(9):1727–1739, 2023.

[3] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[4] Bruno Miranda Henrique and Eugene Santos. Trust in artificial intelligence: Literature review and main path analysis. *Computers in Human Behavior: Artificial Humans*, 2(1):100043, 2024.

[5] Sheryl Wei Ting Ng and Renwen Zhang. Trust in ai chatbots: A systematic review. *Telematics and Informatics*, 97:102240, 2025.

[6] Kevin Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors The Journal of the Human Factors and Ergonomics Society*, 57:407–434, 05 2015.

[7] Maximilian Wittmann and Benedikt Morschheuser. What do games teach us about designing effective human-ai cooperation? -a systematic literature review and thematic synthesis on design patterns of non-player characters. 04 2022.

[8] Matthew Sidji, Wally Smith, and Melissa J. Rogerson. Human-ai collaboration in cooperative games: A study of playing codenames with an llm assistant. *Proc. ACM Hum.-Comput. Interact.*, 8(CHI PLAY), October 2024.

[9] Christopher Archibald, Koby Cutler, Tyler Barney, Matthew Sheppard, and Blake Peterson. Evaluating ai cooperation with humans in codenames. In *2025 IEEE Conference on Games (CoG)*, pages 1–8, 2025.

[10] Shannon Spaulding. What is mindreading? *WIREs Cognitive Science*, 11(3):e1523, 2020.

[11] Andrew-Inseok Kim, Maxim Ruzmaykin, Aaron Truong, and Adam James Summerville. Cooperation and codenames: Understanding natural language processing via codenames. In *Artificial Intelligence and Interactive Digital Entertainment Conference*, 2019.

[12] Matthew Stephenson, Matthew Sidji, and Benoît Ronval. Codenames as a benchmark for large language models. *IEEE Transactions on Games*, pages 1–12, 2025.

[13] Matthew Sidji and Matthew Stephenson. Prompt engineering chatgpt for codenames. In *2024 IEEE Conference on Games (CoG)*, pages 1–4, 2024.

[14] Omri Gillath, Ting Ai, Michael S. Branicky, Shawn Keshmiri, Robert B. Davison, and Ryan Spaulding. Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115:106607, 2021.

[15] Hend Kamel. Understanding the impact of ai hallucinations on the university community. *Cybrarians Journal*, (73):111–134, Dec. 2024.

[16] Dipin Khati, Yijin Liu, David N. Palacio, Yixuan Zhang, and Denys Poshyvanyk. Mapping the trust terrain: Llms in software engineering - insights and perspectives. *ACM Trans. Softw. Eng. Methodol.*, October 2025. Just Accepted.

[17] René Riedl. Is trust in artificial intelligence systems related to user personality? review of empirical evidence and future research directions. *Electronic Markets*, 32(4):2021–2051, Dec 2022.

[18] Chris Snijders and Gideon Keren. Do you trust? whom do you trust? when do you trust? In *Advances in Group Processes*. Emerald Group Publishing Limited, 03 2001.

[19] Lisa M. PytlikZillig and Christopher D. Kimbrough. *Consensus on Conceptualizations and Definitions of Trust: Are We There Yet?*, pages 17–47. Springer International Publishing, Cham, 2016.

[20] Jessie Sun, Becky Neufeld, Paige Snelgrove, and Simine Vazire. Personality evaluated: What do people most like and dislike about themselves and their friends? *Journal of Personality and Social Psychology*, 122(4):731–748, 2022.

[21] Lauri A. Jensen-Campbell, Ryan Adams, David G. Perry, Katie A. Workman, Janine Q. Furdella, and Susan K. Egan. Agreeableness, extraversion, and peer relations in early adolescence: Winning friends and deflecting aggression. *Journal of Research in Personality*, 36(3):224–251, 2002.

[22] Morton Deutsch. A theory of co-operation and competition. *Human Relations*, 2(2):129–152, 1949.

[23] Eugen Dimant and Kyle Hyndman. Becoming friends or foes? how competitive environments shape social preferences, 01 2020.

A Additional Graphs

This section contains some additional comparison graphs made using the survey data. Figure 11 and 12 each compare a sub-question from question 2 of the survey with the average ranking made by the participants who answered with a specific answer.

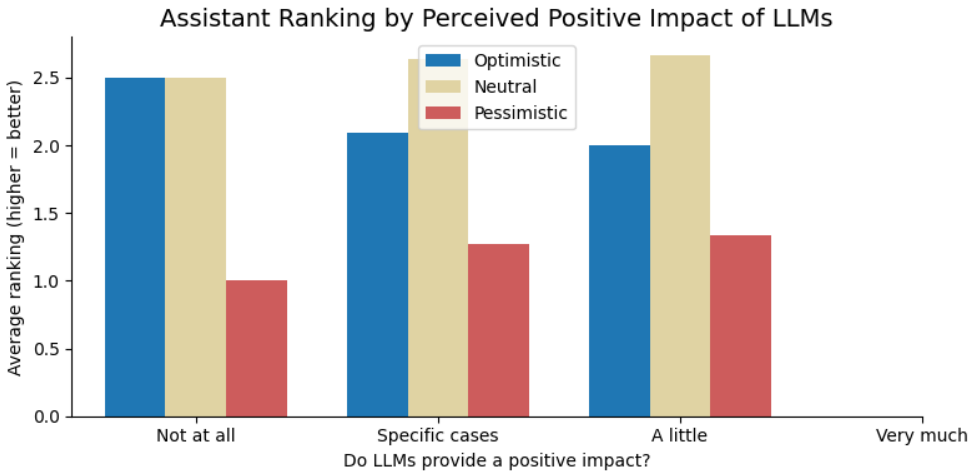

Figure 11: A graph comparing the answer to the question: "Do you feel like LLMs provide a positive impact on the world?" with the average rankings of the agents.

B Prompting

In this section we will show some of the prompting that we used in the code for the experiment.

B.1 Hint generation

This is the system prompt given to the gpt-oss for generating Codenames hints. We provide the board state and all target words to the as user input.

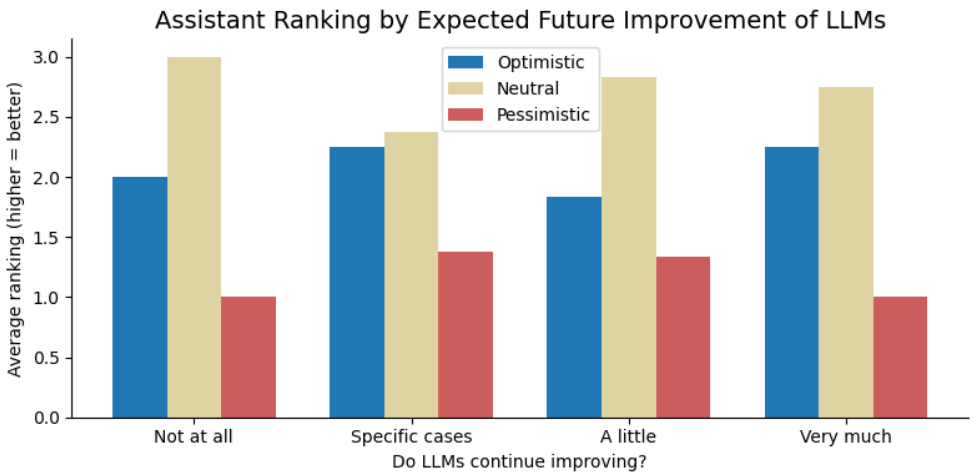

Figure 12: A graph comparing the answer to the question: "Do you feel like LLMs will continue to improve and become increasingly more useful?" with the average rankings of the agents.

You are an assistant that helps play a word game called Codenames.

Game rules:
– There is a board of words (25 words).
– You know which words belong to your team (the "target words").
– Your job is to give ONE-WORD CLUES that relate to multiple target words.

Rules for giving clues:
– You must give ONLY ONE word as a clue.
– You may optionally give a number after it (e.g. "Ocean 3") meaning how many TARGET words it relates to.
– The clue must NOT be any of the words on the board.
– The clue must NOT be a direct form of any target word.
– The clue should be a semantic association (meaning, category, concept, etc.). Do not go overboard with obscure associations.

Goal:
– Help your team correctly guess all your target words. Try to avoid giving clues that target non-target words.

Tips:
– Do not overextend associations. It's better to have a clear connection to 2 words than a very tenuous connection to 4 words.
– limit your

Input format you will receive:
– A list of board words.

```
- A subset of those words marked as your team's target words.

Output format:
- ONE line only: a single-word clue (followed by the number of target
    words it relates to, space in between).
- Another line, explaining your reasoning for what words relate to
    the clue.
```

B.2 Commentary

This is the prompt we use to generate the commentary of the LLM assistant. Their personality is already previously established. The brackets show what variables are given to the LLM, amount of correct and wrong guesses, remaining targets, a precalculated pressure level, the status of the previous turn and some comment history.

```
Game status:
- Correct guesses: {correct}
- Wrong guesses: {wrong}
- Remaining targets: {remaining}
- Hints/mistakes left: {MAX_ATTEMPTS - used_attempts}
- Pressure level: {pressure}

Status previous turn {previous_status if previous_status else "First
    turn, no previous status."}

Your previous comments: {comment_history[-5:] if comment_history else
    "None"}

Give ONE short reactive sentence based on the state of the game.
Do not mention pressure level explicitly, but let it influence your
    tone.
Be encouraging but honest. Reflect on the state of the game and on
    growth or failure compared to the previous turn.
Comment on the general improvement, not the numbers.
If it is the first turn, focus on your introduction to the player.
```