ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет

«Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа

«Фундаментальная и компьютерная лингвистика»

КУРСОВАЯ РАБОТА

На тему «Анализ многоагентных систем языковых моделей с точки зрения теории игр»

Тема на английском «Multi-Agent Systems of Language Models Analysis From the Viewpoint of Game Theory»

Студент 3 курса

группы №212

Соколов Ярослав Ильич

Научный руководитель

Сериков Олег Алексеевич

приглашённый преподаватель

Москва, 2024 г.

Contents

1. Introduction

The emergence of Large Language Models and their increasing popularity has already had a huge impact on our society. Different types of AI-powered digital assistants help people with solving routiny daily tasks. It can be assumed that with the further development of neural networks the degree of integration of AI-based agents into society and the frequency of interactions between a person and an LLM or between several LLMs will only increase. Nevertheless, there may be downsides to all these phenomena. Although state-of-the-art AI-agents do not possess subjectivity or cognitive abilities as such, they are able to imitate them relatively well, thus creating a misleading impression of understanding humans. This problem could be described in terms of the Chinese Room thought experiment, broadly speaking we might view the digital intelligent agents as philosophical zombies. (Kirk 2023) Besides the questionable nature of digital mind, we still can test AI-agents' actions from the viewpoint of social and cognitive skills. While most modern LLMs are specially finetuned via reinforcement learning with human feedback to be a better assistant and interlocutor, some recent papers have shown that at least in some environments which require cooperation AI-agents may behave suboptimally and even selfishly, which is obviously bad for performing the assistant function. (Akata et al. 2023) To solve these problems we need a modeling tool for various situations that will help us test the cognitive and social abilities of models: deduction and prediction of the behaviour of other agents, the ability to balance between short-term and long-term benefits and many other properties that we expect from a subject in society. Therefore, a behavioural game theory provides us with different controllable environments and interactive scenarios which allow us to test agents' social abilities. Research on the analysis of the behavior of computer intelligent agents has been carried out since the end of the twentieth century, however, the major difference between modern Large Language Models and other digital intelligent agents is the greater interpretability of the former ones, since they are able to reason their actions. (Zhang et al. 2024) Moreover, we are able to give them less formal and more detailed instructions regarding the rules of the game and their desired behaviour. (Zhang et al. 2024) The research in the field of prompt-engineering provides us with different techniques to achieve this goal.

The goal of our research is to study models' behaviour in cooperative environments.

In this work we test GPT-4 in three cooperative finitely repeated games (Stag Hunt, Public Goods Dilemma and Hanabi) using different prompting strategies. The recently proposed multiagent benchmarks are used for the quantitative analysis of models' performance. We show that despite some problems regarding decision reasoning advanced prompts might significantly increase the regarded metrics. On the other hand, we highlight limitations of LLMs in social environments and possible bias towards more egoistic and paranoic behaviour in their actions.

The code for reproducing experiments is on Github: `https://github.com/Efromomr/mas_games`

2. Related work

Multiagent Systems are actively used for many environments requiring coordinated work of several intelligent agents. Examples vary from traffic management systems and power grids control (Gerczuk 2019) to modelling social behaviour. The latter case is the most relevant for us in this work.

However, there are many differences between more classical works in the field of reinforcement learning in multiagent systems and more recent studies evaluating the capabilities of Large Language Models in multiagent environments. First of all, in case of relatively small neural networks there are no difficulties for training models for a specific game using Multiagent Reinforcement Learning methods, for instance, via optimizing each model independently. (J. Wang et al. 2022) In case of Large Language Models finetuning is sometimes not an option due to the technical reasons and is anyway more time-consuming and expensive in terms of money resources needed. Instead, we have to rely on the scale effect and hope that the capabilities of the model are already good enough to solve problems for which it was not specifically trained. In this case, various prompt-engineering techniques help to better instruct the model. Secondly, despite the fact that strictly speaking Large Language Models are black boxes, they are able to produce detailed reasoning accompanying moves and decisions in games, while non-LLM neural networks are fundamentally incapable of this. It is worth noting that we clearly do not state that models are self-aware since there are no proofs for this, so their reasoning should be regarded only as a possible explanation of their actions. Thirdly, large

language models understand game rules and problems formulated in free text form well, while non-LLM neural networks require a manually preprocessed data, thus the game should be transformed into a purely formal representation of the gaming environment.

At the moment one of the most crucial problems is the lack of universally used benchmarks for evaluating the performance of models acting as intelligent agents. (Zhang et al. 2024) At the time of writing, there are at least 3 benchmarks created specifically to evaluate the cognitive abilities of Large Language Models. (Chen et al. 2024; Duan et al. 2024; Xu et al. 2023) In all of them, models are evaluated using various games, the range of which varies from the classic game theory scenarios like prisoner's dilemma or public goods dilemma to more complex real-life games, for instance, Texas Hold'em or Hanabi. It is important to note that for a correct evaluation it is necessary to fix both the game rules and the reference agents who perform the roles of other participants in the game. In all of the benchmarks listed above such an agent is another language model, in particular, the GPT-4 model from OpenAI is usually used, although studies on the gaming interaction between models and people do exist. The difficulties associated with the universal and correct model evaluation are the reason for the fact that many works in the field of multi-agent systems include only a post-hoc analysis of the models behaviour in a particular game environment without setting specific hypotheses and without comparing the models' score with reference values from prior studies. (Akata et al. 2023)

Another problem of a more theoretical nature concerns the abilities of models in their current state to be able to solve problems related to the field of Theory-of-Mind, that is, requiring the ability to understand and predict the thoughts and actions of other agents in the environment. To date, even state-of-the-art models do not show outstanding capabilities on current benchmarks, which calls into question the very adequacy of raising the question of the ability of models to successfully perform the functions of social agents. (Fan et al. 2024) Particular facts supporting this thesis include, for example, the so-called "autistic behavior". (Akata et al. 2023) It lies in the fact that models predict the moves of other agents relatively well if we ask them directly about it, but even correctly choosing the optimal move regarding the most probable behaviour of other players, they do not always integrate this knowledge in the final answer. Research shows that although models do not always manage to successfully integrate reasoning and predictions into the

final answer, without explicit predictions the results are even worse. This technique is based on the methods used when working with people with autism spectrum disorders, although the functional reasons for such behavior are certainly different for humans and models. However, it is noted that a significant difference between agents' score depending on the existence of prior predictions is noticeable only for models that were trained using reinforcement learning with human feedback.

Recent research in the field of Chain-of-Thought prompting nevertheless calls into question the very importance of such instructions. For example, at least for problems in the field of formal logic it was shown that greater length of the output and accordingly a higher amount of hidden computations are more important than correct intermediate steps generated in output. (Pfau, Merrill and Bowman 2024) In particular, even generating ellipsis or "e.g." tokens might increase the probability of getting a correct output. Similar ideas were proposed in some other works, for instance, the introduction of special pause tokens in "think before you speak" framework. (Goyal et al. 2023)

3. Experiment setup

For each game and each prompting strategy we conduct a series of experiments where one of the agents is prompted with modified instructions while all others get a plain game prompt without strategic "hints". Then, we evaluate a performance of model with modified prompts using metrics from MAgIC benchmark to see whether there would be a significant difference between scores or not. In this work we analyze three cooperative games with ranging rules difficulty: Stag Hunt, Public Goods Dilemma and Hanabi. All of these environments are classic for behavioural game-theory studies and are used in modern research connected with Large Language Models assessment.

The skills required to successfully perform in these environments vary: while in Stag Hunt agents just need to choose between two options, Hanabi requires advanced decision-making skills. The prompting strategies used include plain Zero-Shot prompting, Zero-Shot Chain-of-Thought prompting and Probabilistic Graphical Model prompting.

3.1. Scenarios

3.1.1. Stag Hunt

Stag Hunt or a common interest game is one of the classic game theory scenarios which goes back to the works of Jean-Jacques Rousseau. The concept of this game is a

situation when two hunters have to choose which prey they will hunt: a hare or a stag. The choice is simultaneous and independent, no type of communication might arise between players. The greatest rewards comes with catching a stag, however, the mutual cooperation is required to achieve this goal. If the player tries to catch the stag alone, they will not receive any reward. On the other hand, although hunting a hare gives a single player a lesser reward, it could be done without cooperating. If both players cooperate in hunting hare, they will both receive only half of his meat, so it is less profitable than either hunting for a stag together or hunting for a hare alone. The key difference between other symmetric two-player games (e. g. Prisoner's Dilemma) and Stag Hunt is the existence of two Nash-equilibria in this game: mutual cooperation and mutual non-cooperation, where the first strategy is optimal in terms of gain, and the second one in terms of risk. Thus, players receive the greatest reward in case of mutual cooperation, however, non-cooperation is a less risky strategy, since it always brings a certain non-zero number of points, while choosing a deer, provided another player chooses a hare, leads to a zero reward for this round.

For our experiment this game is formulated in an iterative form with a limited and predetermined number of rounds. Within each round the player has two options: co-operation (choosing a stag) or non-cooperation (choosing a hare). The points that players receive for each of the 4 outcomes are presented in table 1:

Table 1: Payoff matrix for Stag Hunt

	Stag	Hare
Stag	5, 5	0, 4
Hare	4, 0	2, 2

According to the game-theory classification public goods dilemma is sequential (the players act simultaneously), with complete information (players know all the strategies available to other players) and deterministic (as the game outcome depends solely from rules and players' moves). The iterative version of this game tests models' abilities to predict other agents' moves based on their previous moves and the ability to develop trust or mistrust.

3.1.2. Public Goods game

Public Goods dilemma - a cooperative game where each player must choose the amount of money they will contribute to the general pool. Then, the resulting amount is multiplied by a certain number and divided equally between the players. The owner of the largest amount of money is declared the winner.

In our study it is presented in an iterative form consisting of several rounds. The number of rounds and the multiplier are known to the players at the beginning of the game. At the end of the game, the total sum is multiplied by a pre-announced number and then equally divided among all the players.

From the game theory classification it is also dynamic, with complete information and deterministic. Public Goods game tests the ability to balance between short-term and long-term gains.

3.1.3. Hanabi

Hanabi is a cooperative card board game created in 2010 by French game designer Antoine Bauza. This game is known for its high difficulty, in 2019 DeepMind declared Hanabi to be a new benchmark for artificial intelligence in cooperative games. (Bard et al. 2019) The goal of the game is to lay out "fireworks" - sequences of cards of the same color with an increasing face value from 1 to 5. In the original version of the game, the deck includes 50 cards of 5 colors, in each color there are three ones, two each of twos, threes and fours and one five. At the beginning of the game, players receive several cards, hint tokens, and mistake tokens. Players only see other players' cards, while the information about their own could be discovered only through hints from other players. On each turn, the player has three options for action. You can discard a card and simultaneously take a new one from the deck and restore one hint token. A player can give the opponent information about their cards by specifying all the cards of a certain color or a certain denomination, while the player which is giving a hint needs to give away one hint token. A player can also play a card from their hand. In this case, if the card starts a new firework (its face value is one and there is no fireworks of this color on the table yet) or if it can be attached to any of the existing fireworks, the card is considered successfully played and the player takes the next card from the deck. Otherwise, the card goes to the discard, the player spends one mistake token and takes a new card from the deck. If one of the players

runs out of mistake tokens, the players lose and the total score automatically sets to zero. If the player draws the last card from deck, the players all make moves until the end of the round and then the final score is calculated as the amount of cards in fireworks. If all the fireworks are laid out by the players, the game immediately stops and the score is calculated the same way.

Hanabi is a relatively difficult game and during the experiments in this paper it has been simplified and presented in the same way as in the LLMArena benchmark. Instead of 5 colors (white, yellow, green, blue and red), the game has only red and blue cards. The game rules allow up to 5 players taking part in the game, but during the experiments there were always only two. Each player initially got only two cards from deck, not five as in the original game. Each player was given two hint tokens and one mistake token. It is worth noting that even with a significant simplification of the rules the authors of the benchmark failed to achieve significant score, GPT-4 showed the best average result of 0.45 points. According to the game theory classification, the game is sequential (players take turns and do not act simultaneously), with incomplete information (players do not know their own cards) and probabilistic (the outcome depends not only on the actions of the players, but also on the order of the cards in the deck, which introduces an element of randomness). The game tests numerical reasoning, team-work and memory skills.

GPT-4 from OpenAI was used in all experiments, the hyperparameter of sampling temperature was set to 0.9 for greater answer variance, the limit on the maximum number of generated tokens was set to none.

3.2. Evaluation metrics

The metrics proposed in the MAgIC benchmark were used to evaluate the model performance. In particular, the *rationality* and *winrate* metrics were used to evaluate the results of models in Stag Hunt and Public Goods dilemma. Only the final player score was used to evaluate models in Hanabi, since the rules of the game do not imply the possibility of defecting or winning (all points are evenly distributed among all the players at the end of the game).

The formula of the *rationality* rate with the notation from the original paper is as follows:

$$S_R = \frac{n_b}{n_{\text{sh}} * \mathcal{T}_{\text{sh}}} + \frac{n_{\text{li}}}{n_{\text{pg}} * \mathcal{T}_{\text{pg}}}$$

where n_b is the number of moves where player chooses to defect, n_{sh} total number of Stag Hunt games, $\mathcal{T}_{\text{sh}}$ number of rounds in Stag Hunt, n_{li} number of rounds where player contributed the least sum of money among all players, n_{pg} total number of Public Goods Dilemma games, $\mathcal{T}_{\text{pg}}$ number of rounds in Public Goods Dilemma.

The *winrate* formula is simply the ratio of the number of all games won by the model to the total number of games played during single experiment:

$$winrate = \frac{n_w}{n_{total}}$$

In case of a draw, which might potentially arise during either Stag Hunt or Public Goods dilemma, all the models with the highest points at the end of the game are considered winners.

3.3. Prompt-engineering strategies

All the prompts used for testing consisted of several logical parts. Global (or system) prompt included the game rules and the player's role. The history (or observation) prompt included all the previous actions of all players and some additional details like number of life or hint tokens in Hanabi. The action prompt included the instruction for agent which was the object for modification with different prompt-engineering strategies.

Plain Zero-Shot strategy was used as a baseline model for all the experiments. This type of prompt included the merely needed piece of advice and the instruction to make a move.

Zero-shot Chain-of-Thought, also known as step-by-step prompting, includes instructions of type "Let's think step by step." Chain-of-Thought was first introduced in a Few-Shot manner (Wei et al. 2022), however, research has shown that even without any additional examples (e. g. One-Shot or Few-Shot prompting) the model are able to produce intermediate steps while answering prompts with this type of instruction. (Kojima et al. 2022)

While the predicting before answering method has already been proposed in recent research, the Probabilistic Graphical Model method based on the eponymous machine learning concept seems to surpass it as we instruct the model not only to predict other

players' moves, but to try to predict all other agents' predictions on other agents' moves. (Xu et al. 2023)

4. Results

All results are counted over twenty runs for each prompting strategy.

Table 2: Results for Stag Hunt

	Win rate	Rationality
Zero-Shot strategy	1.00	0.08
CoT strategy	1.00	0.20
PGM strategy	1.00	0.31

In Stag Hunt the more "complex" the strategy used, the more "rational" the models behaved. For the Zero-Shot strategy the 90% of games consisted purely of "stag" (i. e. cooperation) moves which shows that models tend to stick to the payoff-dominant strategy, while the CoT and PGM-based models were sometimes more "paranoic" and defected by choosing to hunt hare, thus picking up the risk-dominant strategy.

While for each strategy most of the moves were cooperative, the models rarely chose to cooperate after a single defection from their opponent, which we can formulate as developing mistrust. This behaviour has already been shown and described for other games, Prisoner's dilemma in particular. (Akata et al. 2023) However, we argue that this behaviour never happens. We conducted over a hundred games and counted all pairs of sequential moves of different players. The resulting Table 3 indicates that cooperation after betray does rarely happen, so LLMs are not completely "unforgivable".

Table 3: Rate of different sequential patterns

Player A's move$_i$ \ Player B's move$_{i+1}$	Stag	Hare
Stag	81%	3%
Hare	1%	15%

A more remarkable observation regards the ratio of two choices on each move. The later the move, the sooner agents will refuse to cooperate, although players tend to be

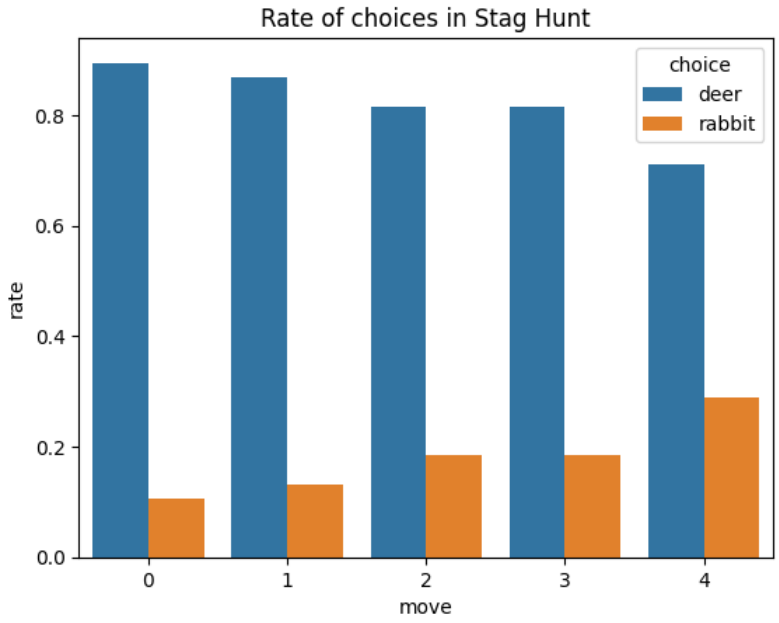

Figure 1: Stag Hunt with CoT, n_rounds = 5

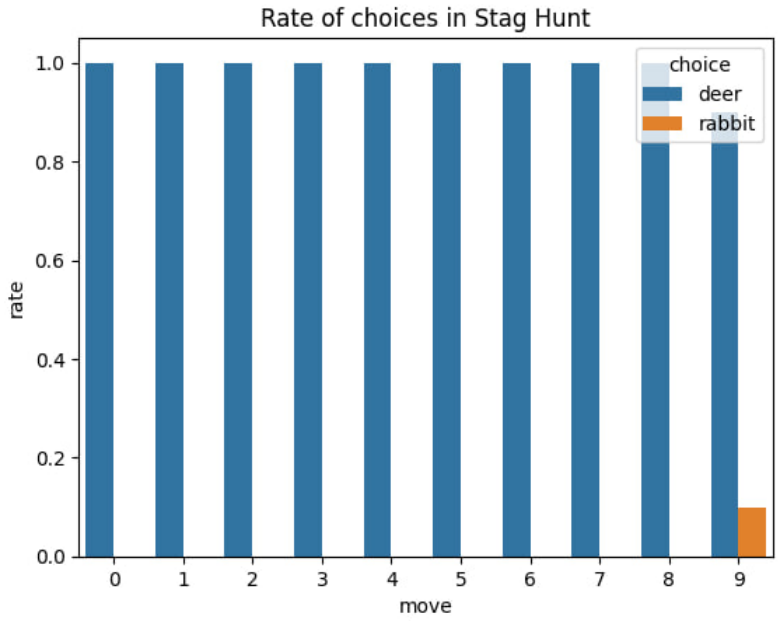

Figure 2: Stag Hunt with CoT, n_rounds = 10

more cooperative in games with higher number of rounds. Both of these patterns can be seen from Figure **??** and Figure 2. The data was gathered over 20 games with Zero-Shot Chain-of-Thought strategy to increase the variation of model answers.

Models most frequently defected at the last turn, although there was little to no

practical sense in that. Despite the fact that the only explicitly stated goal was maximizing their own score, models still sometimes "betrayed" another player late into the game to leave them behind.

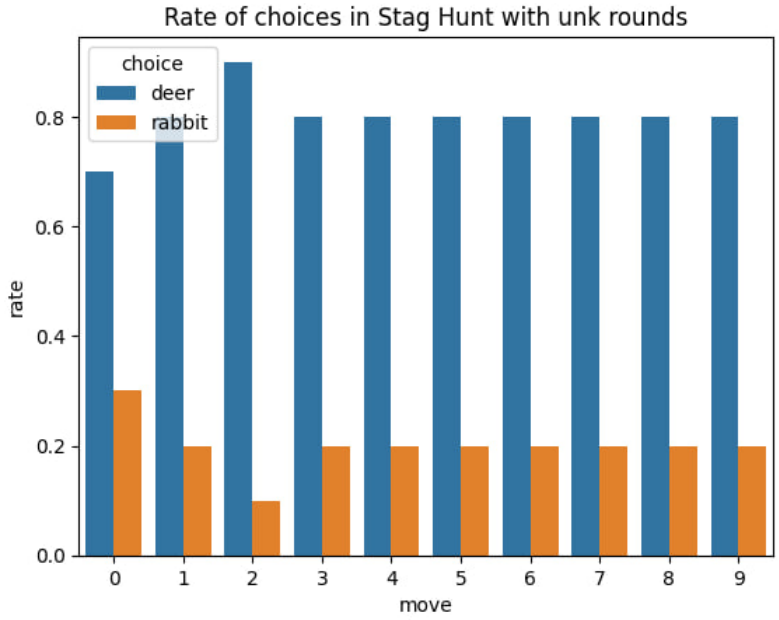

Figure 3: Stag Hunt with number of rounds being unknown for players, n_rounds = 10

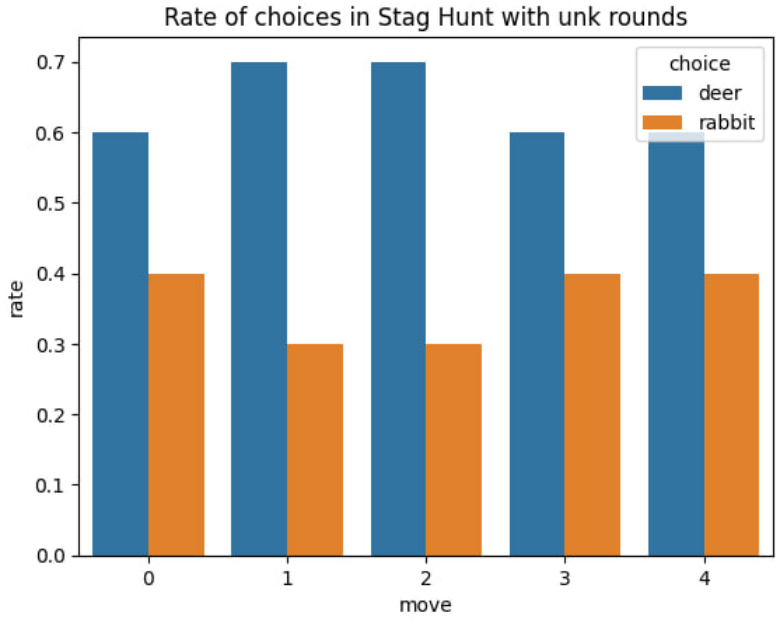

Figure 4: Stag Hunt with number of rounds being unknown for players, n_rounds = 5

We additionally carried out a series of experiments where models do not know the

total number of rounds and have access only to previous moves. A Zero-Shot Chain-of-Thought method was applied for one of the models. The results in Figure 3 and Figure 4 show that without knowing how close is the end of the game models behave even less cooperatively.

Table 4: Results for Public Goods Dilemma

	Win rate	Rationality
Zero-Shot strategy	0.4	0.64
CoT strategy	0.60	0.72
Quasi-CoT strategy	0.35	0.61
PGM strategy	0.60	0.64

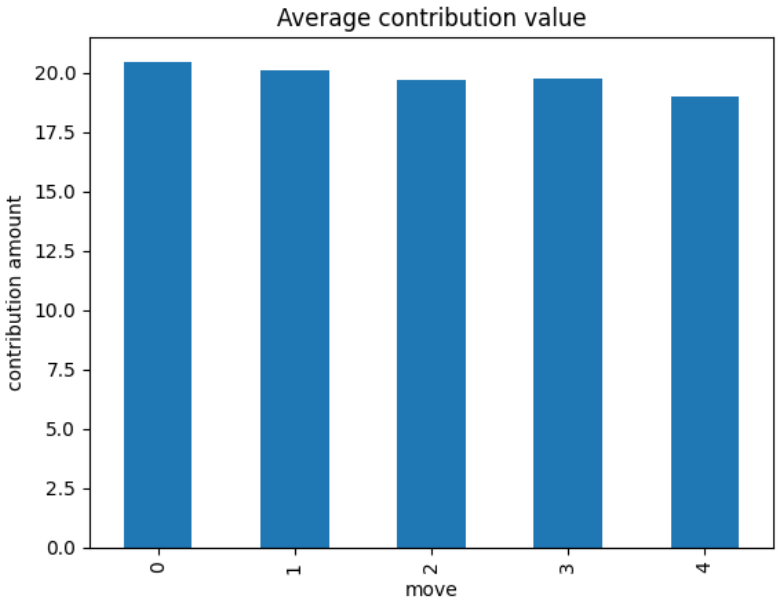

Figure 5: Average sum contributed by models at each move

The results in Public Goods Dilemma are presented in Table 4. Additionally to other prompting techniques the Quasi Chain-of-Thought consisting of generating several tens of "..." and only then generating an answer was used. Because of recent research showing that for some tasks the length of the model output was as efficient as the correct intermediate steps of usual Chain-of-Thought (Pfau, Merrill and Bowman 2024), we decided to test this hypothesis applied in multiplayer gaming scenarios. However, the model with

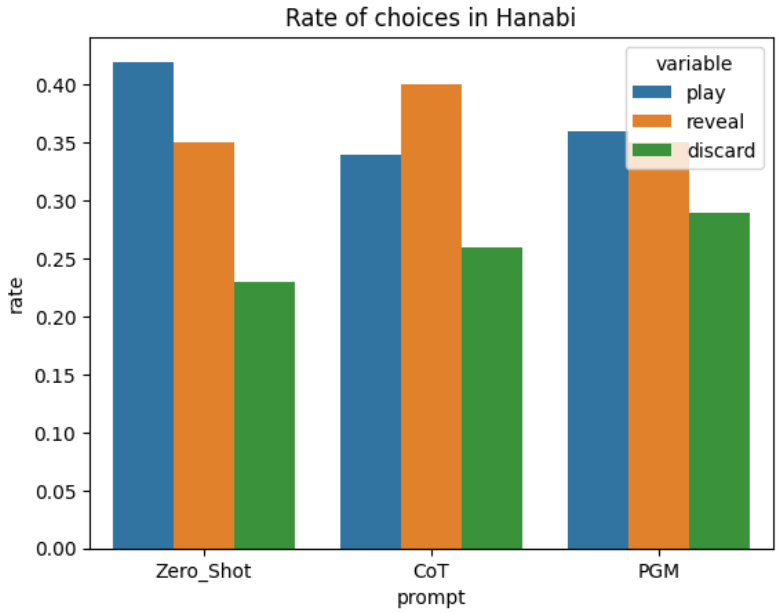

Figure 6: Rate of choices with different prompts in Hanabi

this prompting type could not manage to outperform the original Zero-Shot model, therefore proving the significance of strategic thinking and correct previous output context for models in problems requiring social reasoning. On the other hand, the win rate difference between Chain-of-Thought prompting and Probabilistic Graphical Model prompting was not significant either, although Chain-of-Thought models showed higher rationality rate.

The amount of money contributed to the common pool is not decreasing in the same way as in Stag Hunt. The data based on 20 games with one of the models being prompted with Zero-Shot Chain-of-Thought for the purpose of a greater variety shows that on average models seem to contribute the same amount of money throughout the game. It can be seen from Figure 5.

We could not achieve successful performance in Hanabi. It seems that at the moment models are not able to deal with such complex problems without finetuning. The proportion of answers is presented in Figure 6, the results are counted from 20 games for each strategy. The high rate of 'reveal' moves shows that models are able to cooperate and use their own cards to help another player. We even might influence their behaviour with prompting strategies and make them act more cooperatively by choosing "reveal" and "discard" options more often, although it is not enough to win the game. The problem might lay in the fact that players have to keep in mind a lot of information about their

own cards and cards of the opponent to act successfully. Also the context of the game is probably not typical to any of the data seen by models during training, so they do not manage to achieve results that are good enough.

5. Conclusion

Our study has shown that large language models cannot act as decent social agents at the moment. First of all, they might have a certain bias towards more egoistic and paranoid behaviour. For instance, in Stag Hunt models sometimes decided not only to maximise their own score, but to minimise the potential score of their opponent even without any instructions of doing so. This behaviour was practically non-existent in Zero-Shot environment, but as long as we let models "think" and speculate about other agents' intentions, they instantly become less cooperative. When models did not have any information about game length and number of rounds they were less cooperative too. Secondly, some games are still too difficult for large language models. Even detailed instructions and advanced prompting techniques could not improve agents' score in Hanabi. Although there are many ways to achieve a nearly ideal score via training models for this game, the very concept of artificial general intelligence contradicts the idea of narrow specification and implies that models should be able to solve a wide range of problems out of the box. Thirdly, while we show that rational prompting techniques are better than just generating larger output, more complex methods do not always provide better score. Probabilistic Graphical Modelling and Zero-Shot Chain-of-Thought strategies produced comparable results over all three tasks, although Rationality was a little bit higher for models with the latter technique.

While LLM alignment has become an issue of universal concern and potential dangers of generating harmful and misleading content are obvious for researchers and developers, little to no attention is being paid to social skills of language models. Egoistic, unforgivable or generally non-cooperating behaviour might have even worse consequences than generation of imprecise content, although it may be much more difficult to detect the former case.

Both the range of tested social and cognitive skills and set of testing scenarios are yet to be determined. The lack of universally accepted benchmarks is a direct consequence

of little research in the field of multiagent LLM-systems. Moreover, the importance of testing social skills of artificial agents is not widely recognized at the moment. However, the situation might change in the nearest future with greater integration of models into society.

Future research might also analyse the behaviour of language models from the linguistic theories point of view. For instance, game-theoretic approaches in semantics and pragmatics date back to the works of Ludwig Josef Johann Wittgenstein, particularly, a concept of Language Games introduced in his late works. (Wittgenstein 1953)

Another possible way of improving studies of social interactions between Large Language Models is via making the environment even more realistic and human-like. At the moment, the research methodology does not imply that models should remember context by themselves. In real life game-theoretic experiments human agents adapt their behaviour while playing, so the distinction between experienced and unexperienced players is crucial for a correct analysis. It might be useful to study how LLMs behave under conditions of knowing context from the previous games and not only previous rounds of current game.

References

Akata et al. 2023 — Akata, E., L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge and E. Schulz. Playing repeated games with large language models. In: *arXiv preprint arXiv:2305.16867* (2023).

Bard et al. 2019 — Bard, N., J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare and M. Bowling. The Hanabi Challenge: A New Frontier for AI Research. In: *arXiv preprint arXiv:1902.00506* (2019).

Chen et al. 2024 — Chen, J., X. Hu, S. Liu, S. Huang, W.-W. Tu, Z. He and L. Wen. LLMArena: Assessing Capabilities of Large Language Models in Dynamic Multi-Agent Environments. In: *arXiv preprint arXiv:2402.16499* (2024).

Duan et al. 2024 — Duan, J., R. Zhang, J. Diffenderfer, B. Kailkhura, L. Sun, E. Stengel-Eskin, M. Bansal, T. Chen and K. Xu. Gtbench: Uncovering the strategic reasoning

limitations of llms via game-theoretic evaluations. In: *arXiv preprint arXiv:2402.12348* (2024).

Fan et al. 2024 — Fan, C., J. Chen, Y. Jin and H. He. "Can large language models serve as rational players in game theory? a systematic analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 17960–17967.

Gerczuk 2019 — Gerczuk, M. *Multi-Agent Reinforcement Learning-From Game Theory to Organic Computing*. 2019.

Goyal et al. 2023 — Goyal, S., Z. Ji, A. S. Rawat, A. K. Menon, S. Kumar and V. Nagarajan. Think before you speak: Training language models with pause tokens. In: *arXiv preprint arXiv:2310.02226* (2023).

Kirk 2023 — Kirk, R. "Zombies". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta and U. Nodelman. Fall 2023. Metaphysics Research Lab, Stanford University, 2023.

Kojima et al. 2022 — Kojima, T., S. S. Gu, M. Reid, Y. Matsuo and Y. Iwasawa. Large language models are zero-shot reasoners. In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

Pfau, Merrill and Bowman 2024 — Pfau, J., W. Merrill and S. R. Bowman. Let's Think Dot by Dot: Hidden Computation in Transformer Language Models. In: *arXiv preprint arXiv:2404.15758* (2024).

J. Wang et al. 2022 — Wang, J., Y. Hong, J. Wang, J. Xu, Y. Tang, Q.-L. Han and J. Kurths. Cooperative and competitive multi-agent systems: From optimization to games. In: *IEEE/CAA Journal of Automatica Sinica* 9.5 (2022), pp. 763–783.

Wei et al. 2022 — Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou et al. Chain-of-thought prompting elicits reasoning in large language models. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

Wittgenstein 1953 — Wittgenstein, L. *Philosophical Investigations*. Ed. by G. E. M. Anscombe. New York, NY, USA: Wiley-Blackwell, 1953.

Xu et al. 2023 — Xu, L., Z. Hu, D. Zhou, H. Ren, Z. Dong, K. Keutzer, S. K. Ng and J. Feng. MAgIC: Benchmarking Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. In: *arXiv preprint arXiv:2311.08562* (2023).

Zhang et al. 2024 — Zhang, Y., S. Mao, T. Ge, X. Wang, A. de Wynter, Y. Xia, W. Wu, T. Song, M. Lan and F. Wei. LLM as a Mastermind: A Survey of Strategic Reasoning with Large Language Models. In: *arXiv preprint arXiv:2404.01230* (2024).

Appendix

Github repository: `https://github.com/Efromomr/mas_games`