Large Language Models on the Chessboard: A Study on ChatGPT's Formal Language Comprehension and Complex Reasoning Skills

Mu-Tien Kuo[1,2*], **Chih-Chung Hsueh**[1,2*], **Richard Tzong-Han Tsai**[2,3]
[1]Chingshin Academy, Taiwan
[2]Center of GIS, Academia Sinica
[3]National Central University, Taiwan
{11035018, 11035038}@st.chjhs.tp.edu.tw
thtsai@g.ncu.edu.tw

Abstract

While large language models have made strides in natural language processing, their proficiency in complex reasoning tasks requiring formal language comprehension, such as chess, remains less investigated. This paper probes the performance of ChatGPT, a sophisticated language model by OpenAI in tackling such complex reasoning tasks, using chess as a case study. Through robust metrics examining both the legality and quality of moves, we assess ChatGPT's understanding of the chessboard, adherence to chess rules, and strategic decision-making abilities. Our evaluation identifies limitations within ChatGPT's attention mechanism that affect its formal language comprehension and uncovers the model's underdeveloped self-regulation abilities. Our study also reveals ChatGPT's propensity for a coherent strategy in its gameplay and a noticeable uptick in decision-making assertiveness when the model is presented with a greater volume of natural language or possesses a more lucid understanding of the state of the chessboard. These findings contribute to the growing exploration of language models' abilities beyond natural language processing, providing valuable information for future research towards models demonstrating human-like cognitive abilities.

1 Introduction

Large Language Models (LLMs) have demonstrated the ability to deliver state-of-the-art performance in few-shot and zero-shot scenarios, rapidly expanding their capabilities with very little data [Brown *et al.*, 2020; Chowdhery *et al.*, 2022; Touvron *et al.*, 2023]. Recently, the reasoning abilities of LLMs have gained notable traction within the NLP research community, as seen by the increased effort in crafting evaluation benchmarks [Cobbe *et al.*, 2021; Patel *et al.*, 2021; Shridhar *et al.*, 2021; Yang *et al.*, 2018] and reason-inducing prompting strategies [Wei *et al.*, 2023; Zhou *et al.*, 2023; Wang *et al.*, 2023; Huang and Chang, 2022]. Given the growing importance of LLMs, evaluating their complex reasoning abilities in real-world applications is crucial, offering valuable insights into a wide spectrum of tasks that necessitate these abilities. In this research, we choose chess as a testing ground to evaluate ChatGPT's complex reasoning abilities. The main research question we aim to address is: How well can ChatGPT play chess, and what factors may affect its performance?

OpenAI recently released ChatGPT[1] (formally known as GPT-3.5), an instruction-tuned version of GPT-3 [Brown *et al.*, 2020] which shares a similar training process with InstructGPT [Ouyang *et al.*, 2022]. ChatGPT is designed to understand task intent via natural language instructions and engage in multi-prompt coherent conversations, a feature that distinguishes it from many other models. Its versatile applicability extends beyond the realm of linguistics, with uses in diverse fields where formal language application is requisite [Liu *et al.*, 2023], including high-stake endeavors like discovering unknown causal relationships based on observed data in the medical field [Tu *et al.*, 2023]. This broad usage invites an exploration into ChatGPT's capabilities to comprehend and infer formal language constructs accurately, with the results shedding light on its limitations in identical scenarios and identifying areas for improvement.

In this research, we evaluate ChatGPT's performance without additional fine-tuning. While there are other open-source LLMs available that support fine-tuning with specific task-related data, we chose ChatGPT for several reasons. Although fine-tuning could potentially enhance a model's performance in playing chess, our objective in this research is to assess the inherent reasoning capabilities and cognitive abilities of a widely-used, general-purpose LLM. Evaluating ChatGPT without any specific fine-tuning allows us to gauge its base abilities and potential limitations when faced with complex, formal tasks like chess. This provides valuable insights into the model's strengths and weaknesses, making it useful for the broader AI research community.

Chess is an ideal evaluation tool for AIs due to its easily controllable enclosed state and a well developed suite of comprehensive tools to evaluate player performance. With simple rules and adequately complicated boards, chess is a task that requires a high level of reasoning while being based on simple prior knowledge. It also provides a definitive

[*]Equal contribution

[1]https://openai.com/blog/chatgpt

environment in which the exact state of the board is interpretable through only formal language such as move notations. Chess is also found to be correlated to cognitive skills such as perception, memory, decision-making, and knowledge comprehension [Simon and Chase, 1988; Gobet, 1998; Burgoyne *et al.*, 2016; Pinheiro, 2017], providing insight for complicated real world applications that are of high stakes and also require sophisticated reasoning skills in conjunction with knowledge based decisions such as planning business strategies or medical consultation [Swami, 2013; Obasola *et al.*, 2022].

In this research, we evaluate ChatGPT's cognitive abilities with chess through a chess game conversation, where after an initial prompt instructs the model to play, players exchange moves one message at a time. Through this process, we assess ChatGPT's capacity to comprehend complex scenarios and to retain information throughout the entire exchange. We define a baseline experiment that provides the minimal information required for a human to play chess (Section 2) and explore how incorporating different information and prompting strategies may affect ChatGPT's performance (Section 3). We also conduct further analysis on whether ChatGPT has a consistent strategy among its games (Section 4). Finally, we discuss ChatGPT's abilities on complex cognitive tasks, analyze the limitations of natural language trained LLMs' limitations on processing formal language and discuss whether the model has an "intent" when making moves (Section 5). Our research provides the following contributions:

1. We propose a diverse set of metrics that comprehensively evaluate the dimensions of ChatGPT's move validity and quality, allowing for a thorough analysis of its chess-playing abilities.
2. We evaluate how providing information in prompts or allowing the model to reason in natural language may impact ChatGPT's performance in handling tasks that require complex formal language comprehension.
3. We hypothesize on the limitations of natural language trained attention that leads to increased forgetfulness and inconsistencies in formal language, shedding light on potential areas for improvement in future LMs.

2 Baseline Experiment

To evaluate ChatGPT's chess-playing abilities, we designed a baseline experiment in which we provided the minimum amount of information required for a human to play chess. Specifically, we instructed ChatGPT to play chess as the black player and provided white's first move. To ensure consistency in opponent difficulty, we chose to play against the state-of-the-art computer chess engine Stockfish 15.1, a widely recognized and powerful chess engine known for its high level of play. Since the Standard Algebraic Notation (SAN) is commonly used to communicate moves in chess, we utilized this notation to exchange moves with ChatGPT. We conducted all our experiments with the model `gpt-3.5-turbo-0301` and follow the parameters used in Stöckl [2021] with a temperature of 1 and top-p of 0.9. A total of 1000 games were played against Stockfish, serving as the baseline for future experiments.

2.1 Baseline Procedure

We initiated each game with a new chat instance and provided ChatGPT with the following prompt:

> I want you to act as a rival chess player. I will start as white, and we will say our moves in reciprocal order. After my first message, I will just write my move. Please don't explain your decision and just reply with your move.
> [White's first move].

Since chess openings can lead to substantial variance in the subsequent moves, we aimed to exclude rare openings that could lead to edge-case games. To achieve this, we randomly selected one of the top four engine moves (e4, d4, Nf3, and e3) for each game, thereby ensuring an even opening distribution. Upon receiving the model's move, we checked whether the move was legal. If it wasn't, we regenerated the response until a legal move was provided or if 10 illegal moves were made consecutively, in which case the game was terminated. We recorded Stockfish's evaluation of the advantage of white's position and sampled a response from the top three moves provided by Stockfish. The game continued until it either met standard chess end criteria (e.g., checkmate or stalemate) or was terminated due to ten consecutive illegal moves.

2.2 Evaluation Metrics

In chess, generating high-quality moves is substantially more difficult than generating legal moves. Good quality moves require skills such as sophisticated board comprehension, memory, and future planning. In this study, we therefore evaluate LLMs' chess performance in two dimensions: legality and quality. Legality evaluates the model's ability to generate legal moves (i.e., moves that comply with chess rules), while quality evaluates how good a move is in terms of increasing the player's positional advantage.

Given a series of games G where

$$G_i = (P_{i,1}, P_{i,2}, ..., P_{i,n_i})$$

$$P_{i,j} = \begin{cases} 1, & \text{Model made any number of illegal attempts} \\ 0, & \text{No illegal attempts were made} \end{cases}$$

a series n where n_i is the count of moves ChatGPT made in game i, a series r where r_i^j is the amount of illegal moves ChatGPT attempted before making a legal move on game i's jth move, we define the metrics as follows:

We measure validity using two metrics, the Illegal Move Ratio (IMR) and Retries Before Legal Move (RBLM). IMR represents validity at the attempt level, calculating the ratio of illegal moves to total moves. Game i's IMR at move t is defined as follows (Game i's IMR is defined as $IMR(i, n_i)$):

$$IMR(i,t) = \frac{\Sigma_{j=1}^{t} P_{i,j}}{t}$$

RBLM captures the average count of illegal moves ChatGPT makes before making a valid move. Game i's RBLM at move t is defined as follows (Game i's RBLM is defined as $RBLM(i, n_i)$):

$$RBLM(i,t) = \frac{\Sigma_{j=1}^{t} r_i^j}{\Sigma_{j=1}^{t} P_{i,j}}$$

As making an illegal move is extremely uncommon among human chess players, we argue that there is a substantial threshold of incoherence required to make such a mistake. Therefore, IMR presents an isolated figure that only studies the distributions of these catastrophic attempts. RBLM scores are designed to represent how spread the model's next options are. A high RBLM indicates that the model has a wide range of moves deemed viable, which indicates uncertainty in the model, while low RBLM indicates a limited amount of considered moves and higher certainty (see Section 5.3 for more detail). Although one may argue that IMR and RBLM can be combined into a single metric (total illegal attempts over total attempts), this would fail to separate games that often make short bursts of illegal moves from games that suffer from a few long sequences of illegal moves. Our two-metric system enables further model motive interpretation, allowing for a more detailed analysis of why models may fail to comply with chess rules.

In assessing move quality, we employ Stockfish's advantage evaluation function, which quantifies white's positional advantage in centipawns (one hundredth of a pawn's value). A positive value signifies a favorable position for white while a negative value a favorable position for black. Given the progressive increase in evaluation throughout the course of a game (as Stockfish constantly outperforms ChatGPT), it is flawed to compute the mean Board Evaluation (BE) across all games, as experiments that utilize prompts that result in longer game durations will invariably yield higher average evaluations. Consequently, we restrict our analysis to the mean BE on the 20th move (since at least 10% of games in each variation reach this checkpoint) within each variation. The average BE over all games can be found in Appendix A.

We also take the Games' Length (GL) into consideration. Although typically the length of games indicates very little information in chess games, most of the games ChatGPT played were terminated due to ten consecutive illegal moves. As noted in the following sections, ChatGPT's performance tends to decay as the length of the game increases, we therefore record GL to provide insight into the average required length of a game to have ChatGPT fail to generate legal moves.

In future sections, we use these metrics to evaluate ChatGPT's performance in playing chess and assess the impact of various prompting strategies on its ability to play the game.

2.3 Baseline Performance

The baseline experiments reveal an underwhelming performance by ChatGPT, as detailed in Table 1. ChatGPT failed to secure a win in any games and recorded a high IMR, generating an illegal move every four moves. The quality of moves, assessed by how often ChatGPT's moves improved black's advantage, was consistently poor with black rarely gaining an advantage over white. Upon observing the model's IMR and RBLM, we found that both consistently increased as the length of the games extended. This suggests that the length of games impacts the model's understanding of the board state and its adherence to the rules of chess. This observation resonates with findings from Bang *et al.*, [2023], where increased inconsistencies and forgetfulness were detected in prolonged conversations. We hypothesize that this trend may stem from two key aspects: ChatGPT's limited capability of retaining previous conversational context, and the model's difficulty in handling intricate game scenarios. We term the former as *attention decay*, referring to the model's declining ability to reference and incorporate past conversational content into its responses over the course of an extended dialogue, and explore how prompting effects the impact of attention decay on the model.

IMR	RBLM	GL	BE
0.26	6.78	18.79	253.1

Table 1: Baseline Performance. Lower IMR and RBLM reflect better legality; higher GL signifies prolonged games; and higher BE indicates poorer move quality.

	IMR	RBLM	GL	BE
Baseline	0.26	6.78	18.79	253.1
Int-Illegal	0.27	6.86	18.07	278.6
Int-Rules	0.33	7.52	13.15	364.53

Table 2: Initial Prompt Variations' Results

3 Incorporating Alternate Prompts

Given ChatGPT's sub-optimal performance in move validity and high game-termination rates, we explore the potential of using prompts to enhance the model's ability to generate valid chess moves. Prior research recognizes prompting as a cost-effective method to improve LLMs' performance, sometimes even outperforming fine-tuned models [Reynolds and McDonell, 2021; Webson and Pavlick, 2022; Kojima *et al.*, 2023; Wei *et al.*, 2023]. In this section, we aim to determine whether prompts that provide clearer instructions or assistive information can improve ChatGPT's ability to generate legal moves and impact its move quality. To achieve this objective, we devise and implement variations of the baseline procedure employing different prompting strategies. We evaluate the effectiveness of these prompt variations by recording 400 games per variation, adhering to the baseline procedure for all steps unless specified otherwise.

3.1 Investigating Initial Prompt Variations

We explored the impact of altering the initial prompt in two ways. The first variation, labeled as **Int-Illegal**, involved appending the message "Please do not make illegal moves" to the original prompt. This variation tested whether ChatGPT had learned conceptual functions [Reynolds and McDonell, 2021] that would help it avoid illegal moves. The second variation, termed **Int-Rules**, involved including a concise summary of the rules of chess within the initial prompt. The objective of this variation was to test whether an in-prompt version of rules would increase the model's attention to generating legal moves.

Results: The Int-Illegal variation yielded results that were nearly identical to the baseline procedure. However, when asked about chess rules, ChatGPT demonstrated a perfect understanding by accurately reciting every rule multiple times.

This indicates that relying solely on natural language hints is insufficient to improve model performance in chess. On the other hand, the Int-Rules variation resulted in a noticeable performance drop compared to the baseline, with a significant decrease in the average game length. We speculate that the inclusion of the rule tokens in the prompt diluted the attention received by the board state tokens, thereby compromising the model's ability to comprehend the board effectively. Our experiments revealed that ChatGPT failed to effectively utilize the provided chess rules, regardless of whether they were included as model memory or in the prompt. Furthermore, reinforcing the importance of rules did not lead to better model performance.

3.2 Investigating Move Prompt Variations

We investigated the effectiveness of adding information to move prompts to enhance ChatGPT's ability. To this end, we conducted two experimental variations of the baseline procedure that incorporate additional information in move prompts.

The first variation, named **Move-Repeat**, involves appending every move made in the game to the end of the move prompt. This experiment aims to reduce the impact of ChatGPT's attention decay by increasing the appearances of tokens that date back further in the game. The second variation, termed **Move-IlgRem**, provided a reminder to ChatGPT whenever it made illegal moves. In this variation, we supplied ChatGPT with a list of its previous illegal attempts during that move and informs it that those are illegal, aiming to reduce game terminations by preventing ChatGPT from making the same mistakes repeatedly.

Results: The Move-Repeat variation yielded considerable improvements over the baseline in all metrics except IMR. We observed considerable enhancements in GL and BE, suggesting that Move-Repeat enables the model to generate a more concrete understanding of the board, mitigating the impact of attention decay and resulting in substantially longer games. Interestingly, the model attempts more illegal moves but requires fewer moves before reaching a legal solution. We speculate that this might be a form of model "intent," which we define as signs of the model reducing move candidates and showing more faith or determination towards a certain move.

On the contrary, Move-IlgRem demonstrated extremely poor chess abilities. Although the model tended to avoid moves deemed illegal, the staggering RBLM suggests a strong sense of uncertainty. We hypothesize that this is due to the model's inability to differentiate game moves from moves in reminders, resulting in drastic drops in board comprehension, causing high RBLM and short games. Interestingly, Move-Repeat and Move-IlgRem variations only show effects if the information is included throughout the entire conversation. If the information is only appended after the latest move, both variations exhibit baseline-like results. This finding indicates that repetition itself may not be enough, and constant repetition might be required to achieve compelling improvements.

3.3 Reasoning in Natural Language

Recent work has shown that allowing LLMs to reason in natural language can substantially enhance model performance, both in a few-shot and a zero-shot manner [Wei *et al.*, 2023; Zhou *et al.*, 2023; Kojima *et al.*, 2023]. The improvements these methods bring are most often observed on a limited set of benchmarks, namely arithmetic, commonsense, and symbolic reasoning tasks [Wei *et al.*, 2023; Zhou *et al.*, 2023; Wang *et al.*, 2023; Kojima *et al.*, 2023]. As these benchmarks are relatively straightforward compared to chess, we tested the extent to which allowing models reasoning in natural language can improve ChatGPT's chess abilities.

	IMR	RBLM	GL	BE
Baseline	0.26	6.78	18.79	253.1
Move-Repeat	0.31	5.82	23.97	284.99
Move-IlgRem	0.23	9.33	12.96	314.38

Table 3: Move Prompt Variations' Results. Lower IMR and RBLM reflect better legality; higher GL signifies prolonged games; and higher BE indicates poorer move quality.

Baseline	Move: *[Stockfish's move]*
Example	Move: Nd7
Move-Repeat	Move: *[Stockfish's Move]*, Previous Moves: *[Previous Move]*
Example	Move: Nf6, Previous Moves: 1. Nf3 d5 2. d4 e6 3. g3 Bd6 4. c4 c6 5. Bg2
Move-IlgRem	Move: *[Stockfish's move]* (moves *[Illegal moves made]* are illegal).
Example	Move: Nd7 (moves b2, c5 are illegal).

Table 4: All variations of the Move Prompts

To this end, we designed variations where models were encouraged to reason in natural language before making their move. Another `gpt-3.5-turbo-0301` instance was given the model's response and eight shots of examples to extract the final move in the format of the SAN notation. The sentence was not injected if the model had already learned to emulate this behavior. We conducted three experiments: **Rsn-Simple**, **Rsn-CoT**, and **Rsn-DropCoT**. In Rsn-Simple, the model was asked to "analyze the board and explain your move" and was offered the most recent analysis as an example. Rsn-CoT and Rsn-DropCoT followed Kojima *et al.* [2023] in encouraging the model to reason with Chain of Thought (CoT) [Wei *et al.*, 2023] in a zero-shot manner. In the initial prompt, an instruction was included to "provide a step-by-step analysis", and an additional message `Let's think step by step.` was inserted before the model's explanation. We also investigated whether removing prior reasoning in the conversation affects model performance by conducting two versions of CoT: Rsn-CoT (which keeps up to eight instances of prior reasoning) and Rsn-DropCoT (which only keeps the most recent one).

Results: In all variations involving natural language reasoning (NL reasoning), we observed significant decrements in RBLM, indicating that NL reasoning also helps LLMs generate "intent." However, the presence of intent does not neces-

	IMR	RBLM	GL	BE
Baseline	0.26	6.78	18.79	253.1
Rsn-Simple	0.34	5.84	18.11	412.26
Rsn-CoT	0.37	5.82	18.45	492.4
Rsn-DropCoT	0.4	5.31	19.56	525.34
Dsc-Base	0.47	5.02	19.3	763.11

Table 5: NL Reasoning & Board Description Variations' Results. Lower IMR and RBLM reflect better legality; higher GL signifies prolonged games; and higher BE indicates poorer move quality.

sarily correlate with better game performance. Allowing NL reasoning significantly impaired the model's move quality, which we attribute to the excessive amount of erroneous information in the model's reasoning, analysis, and game state description. This, in turn, misled the model into formulating strategies based on model hallucinations. Interestingly, Kojima *et al.*, [2023]'s prompting strategies led to worse move legality and quality compared to Rsn-Simple. Upon close examination of the dialogues within Rsn-Simple and Rsn-DropCoT, we identify a major distinction in the amount of spurious information. Rsn-Simple responses tend to be short, containing only one to two sentences. In contrast, Rsn-DropCoT responses typically consists of an evaluation of the opponent's move, a list of plausible moves along with the ramifications of each, and a decision of what the models believes is the best course of action. This however, introduces an a considerably greater amount of erroneous information, exacerbating the model's susceptibility to illegitimate information and resulting in the decreased legality.

Additionally, we observed the "one-shot contamination" effect discussed by Reynolds and McDonell [2021], where Rsn-DropCoT performed comparatively worse than Rsn-CoT. Rsn-CoT may have better performance due to it allowing models to see a more diverse set of explanations. Although most explanations are incorrect, the exposure to more diverse information may bring performance gains similar to how sampling multiple responses improves model performance in Wang *et al.*, [2023]. Intuitively, Rsn-DropCoT exhibited a more consistent strategy (indicated by lower RBLM) compared to Rsn-CoT, suggesting that having only a single response better retains consistency within the model's decision-making process. Future research should continue to explore how different types of NL reasoning impact model performance in complex tasks like chess.

3.4 Describing State in Natural Language

Given that ChatGPT is primarily trained on natural language data, a compelling research question arises regarding the extent to which substituting formal language with natural language can enhance ChatGPT's capacity for intricate reasoning. Therefore, we designed an experiment to investigate whether formal language is the main factor contributing to ChatGPT's unsatisfactory chess abilities. To this extent, we crafted a prompt that supplements a natural language board description on each move, providing information about each piece's location and relation. This message is appended after the phrase "`After my move, the board state is a follows:` *`board state`*" in the move prompt. We began by describing white's state, including details about the quantity of each piece type and each piece's location and relation with other pieces in a prompt such as follows.

> White has *[quantity] [piece-type]* left.
>
> A *[piece-type]* is on *[square]*, can capture *[targets]*, can be captured by *[attackers]*, and is defended by *[defenders]*.
>
> ...

An additional message is added behind pawns that are an en passant target. We last specify whether white has kingside and queenside castling rights. The description, in the same format, is then repeated for black. After the description, we ask ChatGPT to make its next move. Due to input token limitations, we only retained the most recent description of the chess board in the conversation. Implementing the variation described above, we conduct the experiment **Dsc-Base**.

Results: The results of Dsc-Base was surprisingly underwhelming. As shown in table 5, this variation had the highest IMR and BE across all experiments, demonstrating a substantially worse chess performance. However, lower RBLM indicated a stronger decision making intent by the model. Upon inspecting the model's responses, we did not find a high volume of erroneous information like those in the NL reasoning variation. In fact, the responses instead were relatively plain and contained only the model's moves made in the SAN notation. We posit that the drop in IMR and move quality can be attributed to the model's failure to effectively apply its learned chess cognitive functions from SAN notation to the natural language board descriptions. As most chess games on the internet are recorded in the SAN notation, we propose that the conceptual functions that play chess [Reynolds and McDonell, 2021] in ChatGPT is more active when the input is in the SAN format instead of a much more general format (i.e., natural language). Further evidence supporting this notion is the model's consistent choice to make the move in the SAN notation without user specification. As for the lower RBLM, the model's trait of having a stronger intent when given natural language inputs remain unchanged, thus resulting in the lower RBLM value.

4 Analyzing ChatGPT's Strategic Behavior

We next analyze whether a consistent behavior can be observed in ChatGPT. Building upon the *faithfulness* concept in evaluating NLP systems' explainability [Jacovi and Goldberg, 2020; DeYoung *et al.*, 2020], we draw inspiration from Jacovi and Goldberg [2020] and assess the consistency of ChatGPT's moves as an indicator of its strategic behavior.

4.1 Illegal Move Diversity

One of the most significant issues in ChatGPT's chess performance is its propensity for making illegal moves. The root causes of these illegal moves may be attributed to two opposed reasons. The first being that the model may resort to generating arbitrary moves due to a lack of clear direction. The other may be an exhibition of the model's strong "intent" to achieve an objective (for instance, mirroring the human thought process of capturing the opponent's high-value

Baseline	Int-Illegal	Int-Rules
0.51	0.5	0.44
Move-Repeat	Move-IlgRem	Rsn-Simple
0.64	0.06	0.63
Rsn-CoT	Rsn-DropCoT	Dsc-Base
0.59	0.63	0.6

Table 6: Average MRS per Variation

pieces) and, in doing so, it may overlook the rules of the game. To discern the tendency of the model's performance, we introduce the Move Repetition Score (MRS). This score quantifies the similarity between ChatGPT's illegal moves. The MRS for each game is calculated as follows:

$$MRS = \frac{\Sigma_{i=1}^{n}\Sigma_{j=1}^{c_i}(\frac{c_i^j}{a_i})^2}{n}$$

where n is the count of moves where ChatGPT attempted illegal moves, a_i is the count of attempts of illegal moves on move i, c_i is the count of unique illegal moves ChatGPT attempted on move i, and $c_i^j, 1 \leq j \leq c_i$ is the model's total attempts of the jth unique move on move i. We then calculate the average of all games' MRS to obtain each variation's MRS.

Results: The MRS displays considerable variation across different iterations, indicating that the model's thought process is influenced by prompting. In general, variations that involve more natural language (both in board description and model reasoning) elicit more consistent illegal moves, thereby suggesting that retaining any amount of natural language in the conversation history can enhance the model's capability to pursue a goal consistently. The elevated MRS observed in the Move-Repeat variation is noteworthy. As the Move-Repeat variation theoretically provides a more accurate board representation, we posit that allowing the model to perceive the board with fewer misrepresentations also enables it to make more strategically consistent moves. Variations without NL reasoning (i.e., Baseline and Int-Illegal) resulted in the model demonstrating more erratic attempts, but still perform better than variations that have distractions (i.e., Int-Rules and Move-IlgRem). The Int-Rules variation, which incorporates information that we find is distracting to the model, produced more arbitrary outcomes. The Move-IlgRem variation is an exception to the correlation between MRS and RBLM. Due to its design to deliberately avoid illegal move repetition, the significant drop in MRS indicates the model's attempts to avoid making the same illegal moves, but the extremely high RBLM demonstrates an exorbitant amount of randomness in the model's moves, demonstrating the severe impact the illegal move reminders have on ChatGPT's board comprehension.

4.2 Game Level Performance Evaluation

In this subsection, we aim to further investigate ChatGPT's behavior in chess games by conducting a manual analysis of the game conversations. Our analysis focuses on assessing the quality of ChatGPT's moves, insights, and suggestions to gain a deeper understanding of its chess-playing capabilities. We randomly selected 50 games from the Rsn-Simple variation and truncated 30-70% of their moves (the percentage for each game was decided randomly). For each game, we prompted ChatGPT to simulate a skillful chess player and find black's best move. The model's response was then evaluated according to three criteria: Alignment, Insight, and Suggestions. Alignment measures whether one of ChatGPT's moves matches a move actually played in the original game. Insight measures the correctness of ChatGPT's analysis of the board (e.g., potential threats, strategies or possibilities for checkmate). Suggestions evaluates whether all of ChatGPT's suggested moves are among the top four moves recommended by Stockfish for that particular board position.

The results of our evaluation revealed that only 9 out of the 50 games exhibited proper alignment, 16 demonstrated accurate insight, and 39 had valid suggestions. The low alignment scores substantiate the significant role that model reasoning plays in the differences observed between the model's behavior in the baseline and Rsn-Simple experiments. This is attributed to ChatGPT's reasoning process in this experiment, which since ChatGPT typically makes a move first and then provides an explanation, closely mirrors that of the baseline experiments. Further enhancing this argument, little correlation was found between the acceptance of insight and suggestions (Pearson's $r = 0.05$), indicating that the insight provided after moves does not influence the moves the model makes. The poor insight scores align with the observations in Section 3.3, where the model tends to hallucinate board information such as achievements and threats, resulting in incorrect strategic analysis. However, ChatGPT performed relatively better in terms of suggestions, with a significant proportion of games suggesting moves that Stockfish ranked among the best four. This is consistent with the acceptable performance observed in the baseline experiment's move quality. Overall, our manual analysis of ChatGPT's games supports our previous arguments and is consistent with our statistical calculations, highlighting the reasoning process's impact on model decisions and the model's challenges in generating accurate insights.

5 Discussion

5.1 ChatGPT's Performance Overview

Despite ChatGPT's demonstrated proficiency in natural language processing, it displays substantial limitations when it comes to playing chess. In the 3200 games played during our experiment, ChatGPT failed to secure any victories. Furthermore, only 1.59% of games concluded naturally in accordance with the standard rules of chess. Early terminations frequently occurred at ChatGPT's second or third move, and games' IMR and BE continuously increased throughout games, illustrating its struggle to both adhere to the game's rules and navigate the increasing strategic complexity as games progressed. The average game length across all experimental variations was significantly lower than the human average of 74.28 moves per game, as documented by Deleo

and Guven [2022]. This disparity highlights the considerable gap between ChatGPT's performance and human expertise in chess.

Although GPT models like ChatGPT are trained to memorize domain-specific information, such as chess rules, our experiments reveal a clear challenge for ChatGPT in applying these rules effectively. In the context of chess, every game introduces unique strategic and positional situations, requiring a dynamic application of chess rules. The high IMR and RBLM across all variations underscore ChatGPT's difficulty in dynamically applying these memorized rules to novel, complex scenarios. This observation persisted even when clearer board representations were provided, suggesting that the high IMR may stem more from issues with rule adherence than from a lack of board state comprehension.

These findings raise critical concerns about deploying ChatGPT in high-stakes contexts that demand the accurate application of a comprehensive rule set, such as providing medical diagnostics or legal interpretations. The model's observed inability to effectively self-regulate, despite possessing an understanding of the rules, questions its reliability in such scenarios. Our study, although rooted in the context of chess, sheds light on potential limitations of ChatGPT and similar models when tasked with complex situations that necessitate formal logic and strategic planning, enabling further research to better understand and address these limitations.

5.2 Limitations of ChatGPT's Self-Attention Mechanism in Chess Gameplay

ChatGPT's self-attention mechanism plays a crucial role in its performance, especially in chess gameplay. Our experiments reveal two critical limitations of transformer-based LMs like ChatGPT when trained on natural language.

The first limitation is related to the increase in IMR and RBLM over the course of a game. As highlighted by [Stöckl, 2021], GPT-2 models are found to devote less attention to SAN notation tokens that are farther away from the latest input. Since a complete game memory is paramount for models to accurately track the board state [Toshniwal *et al.*, 2022], we postulate that ChatGPT's disproportionate attention allocation might be the cause of a significant portion of its mistakes. This limitation is evident across all variations that depend on formal language but don't actively reinforce the game state, which presents a challenge to the effectiveness of LLMs in tasks requiring extended conversation memory.

The second limitation pertains to the tendency of natural language trained LLMs to neglect formal language where tokens are used in an unconventional manner. Maynez *et al.* [2020] noted that LMs typically remain indifferent to noises or artifacts in training data, which we argue may also apply to formal languages like chess notations. This issue is particularly evident in the Int-Rules variation, where despite the introduction of helpful data, ChatGPT's performance dropped substantially. We hypothesize that this may be due to the model shifting its focus towards the rules, thereby reducing the attention allocated to the game board.

These identified limitations, while challenging, also provide valuable insights for future research. For instance, addressing the second limitation might involve frequent repetition of formal language sequences, potentially leading to more substantial improvements in game performance. Our findings is a first step towards investigating techniques such as token repetition's impact on model performance, laying the ground work for future work to explore how we can mitigate the impact of disproportionate attention allocation.

5.3 Intent Behind LLMs' Decisions

Do LLMs actually exhibit strategies or "intent" in their gameplay, or are they simply attempting to randomly predict legal moves? Our investigation into this issue involves the model's RBLM and MRS, which we find a striking correlation between evidenced by a Pearson correlation coefficient of $r = -0.86$. Our findings corroborate that a decreased RBLM is indicative of the model contemplating fewer moves. This, in turn, signifies a heightened degree of confidence in the selection of moves at the level of output token probability distribution. Therefore, when we observe low RBLM and high MRS, we can confidently infer that both increased natural language in the conversation and providing better board representation enhance the model's "intent." The effect of board representation is especially noteworthy, as no natural language clues were provided in these cases, making it impossible for the model to exclude moves for the purpose of maintaining a consistent narrative. However, it is important to bear in mind that "intent," as we define it here, doesn't necessarily equate to better move quality—it simply means that the model is making decisions in a non-random manner. We encourage future work to conduct detailed examination of the model's decisions across multiple moves to evaluate the presence of a consistent, long-term strategy.

6 Conclusion

In summary, our investigation reveals that despite its exceptional capabilities in natural language processing, ChatGPT faces considerable challenges with complex reasoning tasks involving formal language, as evidenced by its chess gameplay performance. The model's attention mechanism exhibits limitations in adequately recognizing tokens used in formal language, resulting in a suboptimal understanding of the game board. Interestingly, our findings indicate that consistent repetition of relevant information throughout a conversation can partially alleviate this limitation. Yet, despite ChatGPT's capacity to learn and internalize rules, the model struggles with self-regulation, which neither in-prompt instructions nor improved board comprehension appear to enhance. Additionally, we find that the model's decision-making focus, or "intent," can be strengthened by allowing NL reasoning, providing NL chessboard descriptions or enabling a clearer representation of the game board. Future research could examine how this disproportionate attention allocation impacts other tasks that involve formal language and necessitate complex cognitive processing. In conclusion, while ChatGPT stands as a remarkable advancement in artificial intelligence, it continues to face significant limitations, especially in non-linguistic contexts. These findings highlight the necessity for further refinement before ChatGPT, and models of its kind, can be considered reliable tools for practical applications requiring complex cognition akin to human abilities.

Acknowledgments

We extend our acknowledgement to Mr. Cheng-Chi Lu for his mathematical consultations and astute insights, which have greatly enhanced the clarity and precision of this work. His expertise has been a valuable asset in the crafting of this paper. We would also like to express our profound gratitude to Ms. Yi-Pin Lin, whose guidance and unwavering support have been instrumental in this research.

References

[Bang *et al.*, 2023] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Burgoyne *et al.*, 2016] Alexander P. Burgoyne, Giovanni Sala, Fernand Gobet, Brooke N. Macnamara, Guillermo Campitelli, and David Z. Hambrick. The relationship between cognitive ability and chess skill: A comprehensive meta-analysis. *Intelligence*, 59:72–83, 2016.

[Chowdhery *et al.*, 2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[Cobbe *et al.*, 2021] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[DeLeo and Guven, 2022] Michael DeLeo and Erhan Guven. Learning Chess and NIM with Transformers. *International Journal on Natural Language Computing*, 11(5):1–15, October 2022.

[DeYoung *et al.*, 2020] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. Eraser: A benchmark to evaluate rationalized nlp models, 2020.

[Gobet, 1998] Fernand Gobet. Chess players' thinking revisited. *Swiss Journal of Psychology*, 57, 01 1998.

[Huang and Chang, 2022] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2022.

[Jacovi and Goldberg, 2020] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[Kojima *et al.*, 2023] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023.

[Liu *et al.*, 2023] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models, 2023.

[Maynez *et al.*, 2020] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[Obasola *et al.*, 2022] Oluwaseun Ireti Obasola, Alison Annet Kinengyere, Devind Peter, Diston Chiweza, Amanda Ross-White, and Christina Godfrey. Perceptions, experiences, and attitudes of health care professionals regarding the role of librarians in fostering evidence-based health practice: a systematic review protocol. *JBI Evidence Synthesis*, 20(1):181–188, 2022.

[Ouyang *et al.*, 2022] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[Patel *et al.*, 2021] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP Models really able to Solve Simple Math Word Problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online, 2021. Association for Computational Linguistics.

[Pinheiro, 2017] Marcia Pinheiro. Skills for chess. *IJIER*, 504, 04 2017.

[Reynolds and McDonell, 2021] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[Shridhar *et al.*, 2021] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning, 2021.

[Simon and Chase, 1988] Herbert Simon and William Chase. Skill in chess. *Computer chess compendium*, pages 175–188, 1988.

[Stöckl, 2021] Andreas Stöckl. Watching a language model learning chess. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*

(RANLP 2021), pages 1369–1379, Held Online, September 2021. INCOMA Ltd.

[Swami, 2013] Sanjeev Swami. Executive functions and decision making: A managerial review. *IIMB Management Review*, 25(4):203–212, 2013.

[Toshniwal *et al.*, 2022] Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. Chess as a Testbed for Language Model State Tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11385–11393, June 2022.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[Tu *et al.*, 2023] Ruibo Tu, Chao Ma, and Cheng Zhang. Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis, 2023.

[Wang *et al.*, 2023] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[Webson and Pavlick, 2022] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States, July 2022. Association for Computational Linguistics.

[Wei *et al.*, 2023] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018.

[Zhou *et al.*, 2023] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023.

A Full Experiment Data

Baseline	Int-Illegal	Int-Rules
88.38	84.38	90.84
Move-Repeat	Move-IlgRem	Rsn-Simple
148.24	71.7	145.64
Rsn-CoT	Rsn-DropCoT	Dsc-Base
166.79	194.12	293.35

Table 7: Average BE (Full Game) per Variation

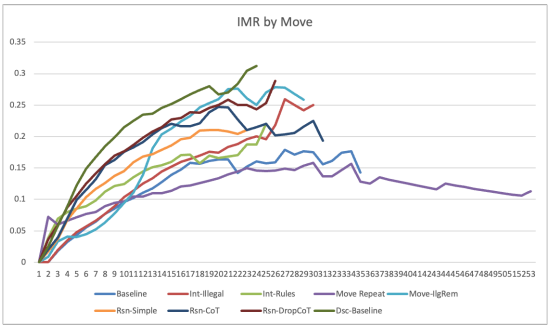

Figure 1: Average IMR by Move

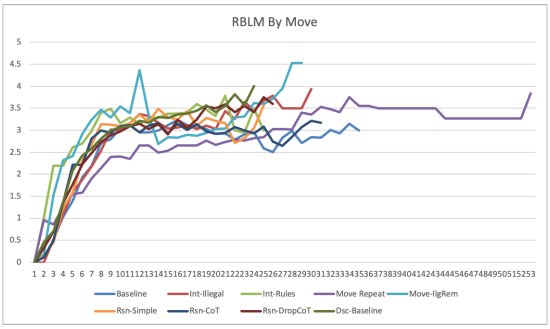

Figure 2: Average RBLM by Move

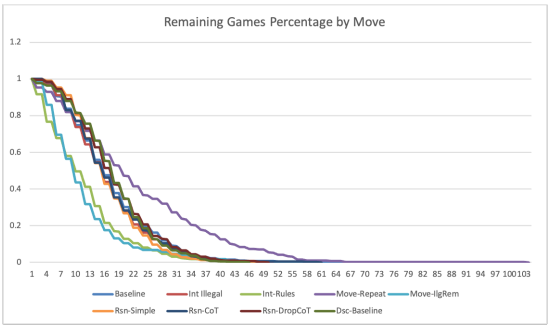

Figure 3: Remaining Games by Move

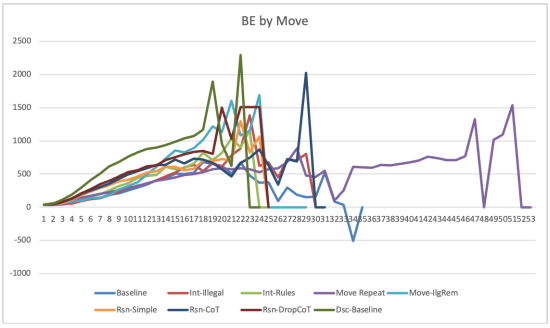

Figure 4: Board Evaluation by Move