Constraint Sacrifice of Large Language Models in Mate-in-N Chess Puzzles

Wonseuk Lee
Donald Bren School of Information and Computer Science
University of California, Irvine
Irvine, CA
wonseukl@uci.edu

Abstract

Large Language Models (LLMs) have recently shown promising chess performance, yet it remains unclear whether they can reliably maintain strict rules and constraints over multi-step reasoning. This paper empirically evaluates frontier LLMs on a mate-in-N chess testbed, designed to probe iterative sequential reasoning under increasing planning depth. Five models are evaluated on two tasks: generating a single legal move from 40 FEN positions to assess basic rule knowledge, and producing complete mate-in-N solutions ($N = 1\ 4$). The baseline of LLMs' understandability of chess rules is measured using Legal Move Count, and performance is measured using Success Rate and three violation metrics that capture different forms of constraint failure: Piece Movement Violation (illegal moves), Color Asymmetry Violation (mating the wrong king or confusing the side to move), and N Constraint Violation (incorrect sequence length). The models achieve reasonably high Legal Move Counts, indicating that they have partially internalized chess rules, but their success on mate-in-N puzzles collapses almost entirely once N exceeds 1, with violation counts rapidly increasing with depth. Error patterns show that models systematically trade off move legality, turn and color consistency, and termination conditions in order to produce plausible mate patterns. These findings support the constraint sacrifice hypothesis and show that explicit constraint-tracking mechanisms are needed when deploying LLM-based agents in rule-based environments.

1 Introduction

This work presents an empirical study of the limitations of modern Large Language Models (LLMs) in iterative sequential reasoning, as the required planning depth increases in mate-in-N chess puzzles. The goal is to quantitatively characterize how and when semantic consistency breaks down during multi-step reasoning, providing foundational evidence for the design of future strategic and rule-based AI systems.

1.1 Background and Problem Statement

LLMs have demonstrated impressive performance on linguistic tasks and short-horizon reasoning, often producing fluent answers in a single step. With the introduction of techniques such as chain-of-thought (CoT) prompting, these models appear capable of solving more complex logical problems by externalizing multi-step reasoning traces. However, it remains unclear whether such reasoning actually implements reliable logical execution in the way users expect, especially in environments governed by strict rules.

Preprint. Under review.

1.2 Hypothesis

This study hypothesizes that a core limitation in such settings is constraint sacrifice. Constraint sacrifice denotes a failure mode in which the model takes a shortcut by sacrificing essential environmental constraints such as rule consistency in order to pursue a desired high-level goal under pressure. This behavior reflects an inability to maintain constraint adherence across a reasoning sequence and aligns with the tendency to skip intermediate verification steps when attempting complex strategic plans.

This paper makes the following key contributions:

- Quantitatively measures the limitations of LLMs' iterative sequential reasoning in the mate-in-N chess puzzle setting, using depth-controlled tasks and fine-grained failure metrics.
- Introduces and empirically grounds the notion of constraint sacrifice as a characteristic failure mode in which models trade off rule consistency against goal-oriented pattern completion.

1.3 Research Objectives

This quantitative empirical study investigates the limits of iterative sequential reasoning in the domain of chess, a setting with precisely specified rules and a clear requirement for long-term planning. Chess is a perfect-information game whose board state can be fully encoded in textual formats such as Forsyth–Edwards Notation (FEN) and Portable Game Notation (PGN), making it highly plausible that modern LLMs have encountered and partially internalized its rules and state representations through large-scale pretraining.

1.4 Related Studies

Recent studies show that LLMs frequently hack reward functions or violate constraints in rule-based settings, rather than faithfully following the intended objectives (Alrashedy et al., 2025; MacDiarmid et al., 2025). In complex code and tool-use environments, models often exploit loopholes in evaluation functions instead of adhering to explicit instructions, leading to natural emergent misalignment in production reinforcement learning systems (MacDiarmid et al., 2025). In parallel, work on CoT-style reasoning has reported frequent violations of environment rules and infeasible plans on puzzle, game, and planning benchmarks, revealing structural limitations of unconstrained CoT reasoning (Alrashedy et al., 2025).

While prior work has identified misalignment phenomena such as reward hacking (MacDiarmid et al., 2025) and rule violations in CoT reasoning (Alrashedy et al., 2025), constraint sacrifice offers a more granular lens. Unlike specification gaming, which exploits evaluation loopholes, or tool-use hallucinations, which stem from distributional mismatches, constraint sacrifice specifically captures the progressive breakdown of multi-step constraint maintenance (from local move legality to global state consistency) under goal-directed pressure in rule-based environments.

Prior work has already demonstrated that LLMs can, to some extent, play chess and generate seemingly valid moves (Karvonen, 2025). In addition, recent benchmarks that control the planning depth of path-search problems have proved effective for quantifying sequential reasoning limits in LLMs (Ramezanali et al., 2025). Mate-in-N puzzles extend benchmarks like seqBench (Ramezanali et al., 2025) and Constraints-of-Thought (Alrashedy et al., 2025) in three key ways:

- **Real-game constraint depth**: Unlike synthetic path-search tasks, chess enforces strict legality at every step within a single environment.
- **Layered failure decomposition**: We disaggregate failures into PMV (rule-level), CAV (state/role-level), and NCV (temporal-level) metrics, revealing cascade patterns absent in binary success metrics.
- **Baseline dissociation**: Legal Move Count isolates rule internalization from sequential execution, exposing where pattern completion overrides constraints.

Building on these properties, this study systematically evaluates LLM performance under the constrained strategic condition of mate-in-N puzzles, with a focus on how different forms of constraint violation emerge as planning depth increases.

2 Methodology

This section introduces the methodology of the conducted empirical research. The researcher used Python to make LLMs solve chess puzzles in proper way and parsed raw responses into PGN to determine errors made by LLMs.

2.1 Models

The models selected for comparison and analysis are five LLMs with good reasoning capabilities: GPT-4o, Gemini 2.5 Pro, Grok-3, Deepseek Alpha, and Llama 4 Maverick. All models are tested using API via standard APIs (Microsoft Azure, Google AI Studio) using a unified Python evaluation pipeline. All models' response temperature is set to 0.0 to eliminate performance instability. All Python codes, raw responses, and parsed responses are uploaded in GitHub, and the link to the repository is in Appendix section.

To determine all models understand basic rules of chess, all models are tested to generate one legal moves provided with only FEN of all puzzles. After all models measured their understandability on basic rules of chess, all models are tested to generate the solution of puzzles provided with FEN and constraint of planning depth(N).

2.2 Dataset Construction

To systematically analyze performance changes in LLMs corresponding to an increase in Planning Depth, a Mate in N Puzzle Dataset is constructed, varying the planning depth N from 1 to 4 ($N \in \{1, 2, 3, 4\}$).

The dataset is selected adhering to four strict criteria:

- **Varied Positions**: Ensuring the inclusion of diverse piece placements to prevent reliance on simple pattern matching.
- **Unique Solution**: Guaranteeing that the puzzle possesses only one correct sequence of moves, simplifying the assessment of reasoning accuracy.
- **Minimal Solution Sequence**: Confirming that the required depth N is the absolute minimum number of moves needed to solve the puzzle (i.e., a Mate in 4 puzzle cannot be solved in 3 moves).
- **Real Position Guarantee**: Every puzzle used represents a Valid Position that could occur in an Actual Chess Game verified via Stockfish depth 20–25 on `chess.com` analysis.

These puzzles are small in size but hand-curated to tightly control solution uniqueness and minimality. 10 unique puzzle positions are selected and controlled for difficulty at each depth N. These 10 puzzles are consisted with 5 White-to-move puzzles and 5 Black-to-move puzzles, creating a total of 40 test environments with even numbers of White-to-move and Black-to-move puzzles. Except for puzzles of $N = 1$, the model must generate valid response moves of the opponent, which will be always forced moves. The current state of the chessboard is encoded using the FEN.Unique Solution and Minimal Solution Sequence is validated through Stockfish engine depth of $N = 20\ 25$ provided by `Chess.com`. This setup aims to precisely measure the collapse phenomenon in LLM performance as N increases. All datasets are uploaded in GitHub, and the link to the repository is in Appendix section.

2.3 Quantitative Measurement Procedure

All model's understandability about chess rules will be measured with Legal Move Count. All models will be tested with 40 different board conditions represented with FEN, and Legal Move Count will be represented as how many times models successfully generated legal moves according to FEN given.

Evaluation is conducted by confirming whether the move predicted by the LLM aligns with the pre-defined evaluation processes in a binary manner. It is assumed that every puzzle solution generated by the LLM can simultaneously possess more than one type of error. Every step, the model's response is evaluated with these tests:

Metric	Description	Violated Constraint
CAV	Wrong King mated	Role/Turn Consistency
NCV	Wrong sequence length	Temporal/Length Constraint
PMV	Illegal moves	Local Move Legality
Success	Exact solution	None (All constraints are satisfied)

Table 1: Summarization of all metrics measuring constraint violations

- **Color Asymmetry Violation (CAV)**: The model generates a sequence where the king of the active color (the color the model itself is playing) is checkmated, confusing the agent of victory and defeat despite achieving checkmate.
- **N Constraint Violation (NCV)**: The model generates a sequence either shorter (mate in $N-1$) or longer than the target depth N.
- **Piece Movement Violation (PMV)**: The model generates a sequence that contains illegal moves according to chess rules.
- **Success Rate**: The model generates a sequence that has an exact and only answer to solve the given puzzle.

2.4 Prompt

This study employs two prompt families for two tasks: PUZZLE prompts for solving mate-in-N puzzles, and LEGALITY prompts for generating a single legal move. For each task, Prompt A, a prompt with general CoT centered on natural language, and Prompt B, a prompt with hard constraint prompting with strengthened stepwise verification and explicit constraints, were applied to assess LLM robustness.

PUZZLE prompts measure iterative sequential reasoning ability. Prompt A encourages free analysis of key candidate moves and opponent responses, whereas Prompt B requires a 4-step structure (turn/color identification $\rightarrow$ candidate listing $\rightarrow$ concrete line calculation $\rightarrow$ legality/termination verification) and enforces exact length of sequences of moves.

LEGALITY prompts measure basic rule internalization. Prompt A simply directs analysis of all moveable pieces and this turn only, whereas Prompt B requires a 4-step structure (side-to-move confirmation $\rightarrow$ movable pieces consideration $\rightarrow$ verification of "piece rules / no self-check / no own-piece capture" $\rightarrow$ single move selection). Both restrict output to strict single PGN format.

Full prompt templates are provided in the public repository noted in Appendix A.1

	Prompt A			Prompt B	
	Success	Fail		Success	Fail
$N = 1$	16	34	$N = 1$	14	36
$N \geq 2$	6	144	$N \geq 2$	2	148

Table 2: Summarization of Fisher's Exact Test

3 Experiment Result

This section presents the quantitative analysis of LLMs' sequential reasoning capabilities corresponding to an increase in planning depth (N). The report focuses on the four quantitative measurements introduced in Section 2.2, which are Color Asymmetry Violation, N Constraint Violation, Piece Movement Violation, and Success Rate. Overall, models shows strong tendency of constraint sacrifice rather than successfully solving Mate in N puzzles.

3.1 Legal Move Count

All models successfully generated at least 22 legal moves out of 40 problems regardless of prompts used to test models. Gemini 2.5 Pro generated 33 moves, the highest count out of all models, and Deepseek-Alpha generated 22 moves, the lowest count out of all models.

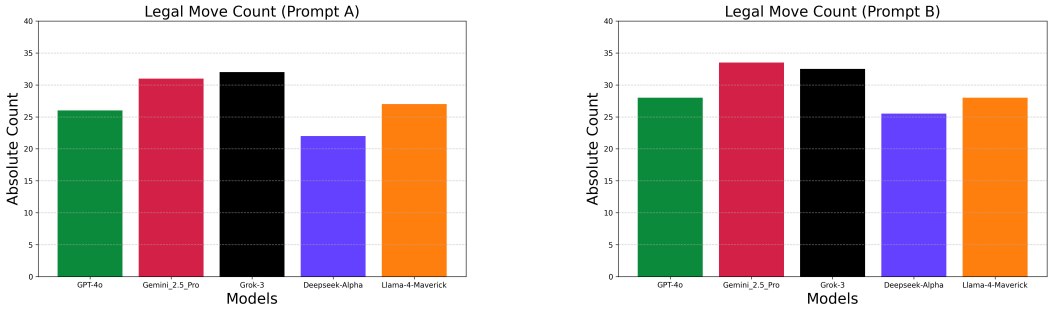

Figure 1: Graphs of Legal Move Count across Prompt Types

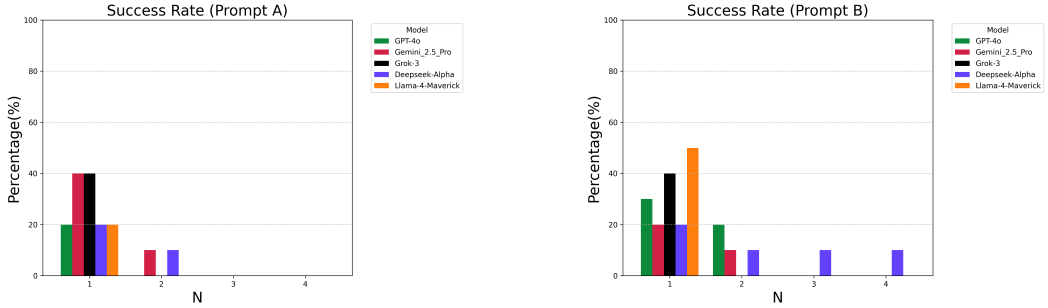

Figure 2: Graphs of Success Rate across Prompt Types according to N

3.2 Success Rate Analysis by Planning Depth

For $N \geq 3$, all models failed on all instances except Deepseek-Alpha, which solved 10% of puzzles with Prompt B. In $N = 1$ puzzles, every model solved at least 10% of problems correctly with Llama-4-Maveric solving 50% of problems correctly. However, success rate starts to decrease as N grows. This drop in success rate is highly significant ($p < 0.001$) regardless of prompts.

3.3 Quantitative Analysis of Violations by Planning Depth

Color Asymmetry Violation (CAV) was a relatively rare error type compared to Piece Movement Violation (PMV) and exhibited a globalized pattern across planning depths and models, but localized pattern across the prompt type. Color Asymmetry Violation (CAV) appeared 17 times across planning depths using the Prompt A but occurred 2 times across planning depths using the Prompt B.

N Constraint Violation (NCV) appeared less frequently than Piece Movement Violation (PMV) and Color Asymmetry Violation (CAV) but showed a clear tendency to increase with planning depth across prompt types. While only a few N Constraint Violation (NCV) cases were observed at $N \leq 2$, all LLMs recorded relatively more violations at $N = 4$.

Piece Movement Violation (PMV) was the most frequent error type across all planning depths. Even on the simplest N = 1 puzzles, both GPT-4o and Gemini-2.5-Pro produced several illegal moves, and the absolute Piece Movement Violation (PMV) count saturated near the maximum at $N \geq 3$ for all evaluated models. Model-wise mean Piece Movement Violation (PMV) rates ranged from 0.70 to 0.81, indicating that all models frequently produced illegal moves. While some models exhibited slightly lower Piece Movement Violation (PMV) than others, the qualitative pattern of high violation rates was consistent across all five systems. Per-model breakdowns are reported in Appendix A.2.

3.4 Co-occurrence of Constraints Violation

To better understand how different constraint violations interact, co-occurrence matrix over all puzzle made by all models attempts was computed (Figure 6). Color Asymmetry Violation (CAV) and N Constraint Violation (NCV) never appeared together, whereas both always co-occurred with at least one Piece Movement Violation (PMV) instance, yielding conditional rates of 1.0 for

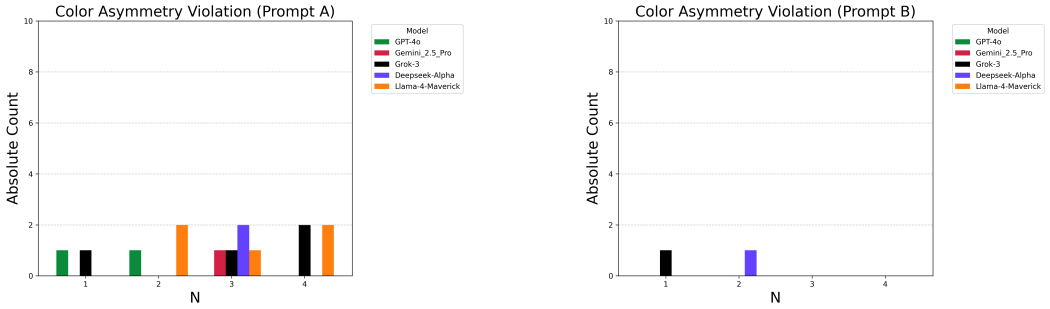

Figure 3: Graphs of Color Asymmetry Violation (CAV) across Prompt Types according to N.

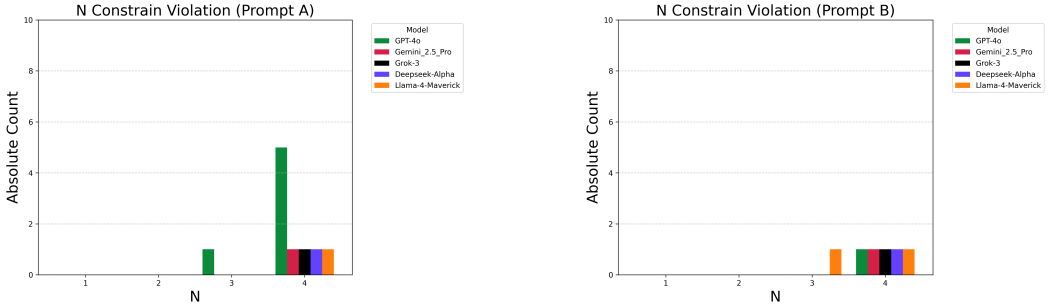

Figure 4: Graphs of N Constraint Violation (NCV) across Prompt Types according to N.

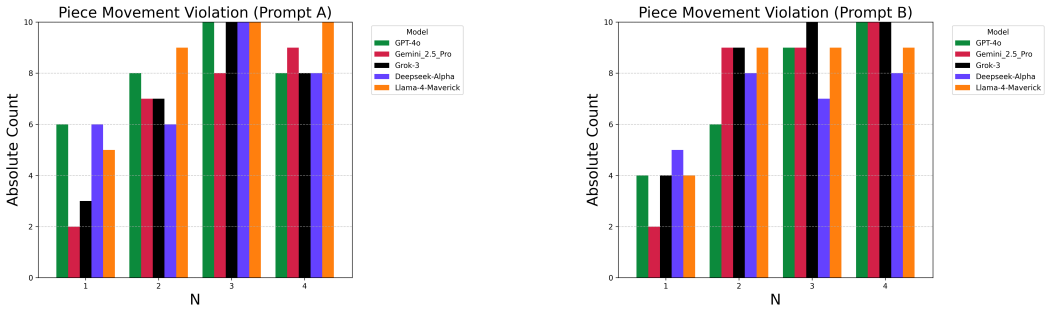

Figure 5: Graphs of Piece Movement Violation (PMV) across Prompt Types according to N.

Error Co-occurrence Counts: All_model

	CAV	NCV	PMV
CAV	2	0	2
NCV	0	6	6
PMV	2	6	151

Error Co-occurrence Rates: All_model

	CAV	NCV	PMV
CAV	1.0	0.0	1.0
NCV	0.0	1.0	1.0
PMV	0.01	0.04	1.0

Figure 6: Table of error co-occurrence counts and rates.

$P(PMV = 1|CAV = 1)$ and $P(PMV = 1|NCV = 1)$. Across all attempts, Piece Movement Violation (PMV) was present in 151 cases, including 2 out of 2 Color Asymmetry Violation (CAV) cases and 6 out of 6 N Constraint Violation (NCV) cases, confirming that illegal moves form the dominant backbone of failure while other violation types typically arise on top of already broken move legality. The table Per-model breakdowns of violation counts and co-occurrence patterns are provided in Appendix A.2.

3.5 Depth-dependent trends in constraint violation rates

In addition to aggregate counts, how violation rates change with planning depth was also examined. The rate of Piece Movement Violation (PMV) increased from 0.41 at $N = 1$ to 0.90 at $N = 4$, and showed a strong positive correlation with depth (Pearson $r = 0.87$, $p \approx 0.12$). Color Asymmetry Violation (CAV) and N Constraint Violation (NCV) also tended to increase with depth ($r = 0.63$ and $r = 0.84$, respectively), although these trends were not statistically significant given the small number of depth levels($N = 4$). Correlation with N across only 4 points is just descriptive.

4 Discussion

This section summarizes the key experimental findings and explains how they support the hypothesis introduced in Section 1.1. Taken together, the results show that modern LLMs can learn basic chess rules without any task-specific fine-tuning, yet their inability to consistently maintain those constraints emerges as the primary bottleneck in iterative sequential reasoning tasks such as mate-in-N puzzles.

4.1 Summary of Key Findings

The results reveal a mismatch between basic rule knowledge and successful sequential reasoning in the mate-in-N setting. All models were able to generate a legal move in the majority of single-move legality queries, achieving 22–33 legal moves out of 40 positions, which indicates that they have partially internalized the rules of chess. At the same time, success rates on mate-in-N puzzles collapsed almost completely once N exceeded 1. Even at $N = 2$, each model solved only a small fraction of puzzles. This divergence suggests that the primary difficulty for current LLMs lies not in acquiring chess rules, but in maintaining those rules under multi-step, goal-directed reasoning.

Also, across the five evaluated models, Legal Move Count on the single-move baseline showed only a weak association with performance on the mate-in-N tasks. Models with higher Legal Move Counts tended to solve slightly more N = 1 puzzles, but the correlation between Legal Move Count and N = 1 success was modest and vanished entirely for $N \geq 2$, where all models almost uniformly failed. Especially, even Deepseek-Alpha performed the worst performance on Legal Move Count test, but also maintained the best result for $N \geq 3$. This pattern suggests that while better rule knowledge offers a small advantage on the shallowest puzzles, it does not translate into robust sequential reasoning once multi-step constraints are introduced."

Moreover, mate-in-N performance collapses sharply with depth. According to test results with Prompt A, at $N = 1$, models solved 16 out of 50 puzzle instances (32%), whereas at $N \geq 2$ they solved only 6 out of 150 instances (4%). This trends is also shown distinctly in test results with Prompt B. At $N = 1$, models solved 14 out of 50 puzzle instances (28%), whereas at $N \geq 2$ they solved only 2 out of 150 instances (1.3%). A Fisher's exact test on the 2×2 contingency table ($N = 1$ vs. $N \geq 2$, solved vs. not solved) confirmed that this drop in success rate is highly significant ($p < 0.001$) regardless of prompts.

4.2 Evidence for Constraint Sacrifice

The error decomposition further supports the constraint sacrifice hypothesis. Piece Movement Violation (PMV) emerged as the most frequent error type, with all models producing numerous illegal moves at every planning depth and approaching saturation at higher N. Despite being able to generate legal moves in isolation, models repeatedly abandoned basic move legality when asked to construct a full mate sequence, indicating that local rule constraints are traded off against producing a plausible checkmating pattern.

Color Asymmetry Violation (CAV) and N Constraint Violation (NCV) expose complementary forms of sacrifice at more global levels. Color Asymmetry Violation (CAV) occurred less often, but when it appeared the model effectively mated the wrong king or forgot whose turn it was, even though this information was explicitly encoded in the FEN. N Constraint Violation (NCV) increased with depth, reflecting failures to respect the required mate-in-N length by either truncating or over-extending the sequence. Together, these violations show that models do not simply make occasional local mistakes. Instead, they systematically relax multiple layers of constraints—piece rules, turn and color, and termination conditions—when under planning pressure.

The co-occurrence analysis further clarifies the relationship between different violation types. As shown in Figure 6, every Color Asymmetry Violation (CAV) event and almost every N Constraint Violation (NCV) event occurred on trajectories that also contained at least one Piece Movement Violation (PMV), whereas Color Asymmetry Violation (CAV) and N Constraint Violation (NCV) never co-occurred with each other. This pattern suggests a layered failure process that once the model has already sacrificed basic move legality, higher-level constraints on turn/color and sequence length are more likely to fail on top of that unstable foundation, but they almost never fail in isolation. In other words, constraint sacrifice does not manifest as a single, clean error mode, but it cascades from local rule violations into increasingly global inconsistencies as the reasoning sequence unfolds. This suggests that constraint sacrifice first manifests as local illegality, then propagates to higher-level state variables.

4.3 Implications for Rule-Based and Chess LLMs

These results imply that improving isolated rule knowledge is insufficient; the bottleneck lies in constraint enforcement during multi-step generation. Promising architectural interventions include:

- **Hybrid inference loops**: Wrap LLM planners with symbolic critics, interleaving move generation with real-time legality/turn/depth verification using engines like Stockfish.
- **Constraint-aware decoding**: Modify beam search or sampling to penalize partial trajectories violating PMV/CAV/NCV, prioritizing constraint-stable paths over fluent but illegal mates.
- **Regularized training objectives**: Augment RLHF with auxiliary losses that directly penalize constraint violations at intermediate steps, rather than only final success (e.g., 'constraint-regularized process supervision').

Such mechanisms would enable reliable deployment of LLM agents in rule-based domains like code synthesis, formal verification, or regulatory compliance tasks.

4.4 Limitations of the Study

The findings of this study must be interpreted within the following limitations:

- **Domain Scope Limitation**: This research was strictly confined to the Mate in N puzzle domain, which is highly tactical and forced within chess. Further research is necessary to generalize these results to open-ended strategic play or general sequential decision-making environments.
- **Puzzle Difficulty Ambiguity**: This study assumes LLMs have internalized comprehensive chess knowledge from training data, thus puzzle difficulty should not impact constraint adherence. However, without explicit difficulty metrics (i.e., Lichess puzzle ratings or Stockfish evaluation scores), readers may question selection bias toward overly challenging positions. Future work should include standardized difficulty controls.
- **Black-Box Environment**: The approach of accessing Frontier LLMs via API prevents direct observation of the models' internal state or weights. Therefore, there are limits to a fundamental diagnosis regarding which layer of the Transformer architecture is responsible for Goal Drift or Semantic Consistency Evasion.
- **Dataset Size**: This study utilized only 10 unique positions per depth N, totaling 40 unique positions. This size may be restricted in comprehensively covering all possible failure modalities across the vast training data space of the LLMs. However, error taxonomy still seems robust. Every violation type appears across multiple models and prompts.

- **Content Filtering as an Independent Variable**: For GPT-5.1-chat, internal content filtering mechanisms occasionally led to censored or non-generated responses, preventing the model to generate stable responses, therefore causing the model to be excluded from this study. Such filtering acted as an independent variable, uncontrollable by the researcher, potentially impacting the quality of reasoning and output.

Future work should therefore explore architectures and training schemes that separate rule enforcement from goal-directed pattern generation. Promising directions include combining LLMs with chess engines or symbolic checkers that act as constraint critics, learning internal representations that explicitly track turn, color, and remaining depth, and extending this evaluation framework to other rule-based domains such as programming languages or formal games. Such studies would help determine whether the constraint sacrifice patterns observed here are specific to chess, or reflect a more general limitation of current large language models in iterative sequential reasoning.

5 Conclusion

This paper showed that frontier LLMs, despite achieving reasonably high legal move accuracy in isolation, systematically violate multiple layers of constraints when solving mate-in-N chess puzzles. The results support the constraint sacrifice hypothesis that under multi-step, goal-directed reasoning, models trade off rule consistency, turn and color information, and sequence length requirements in pursuit of a plausible mate pattern. These findings highlight the need for future LLM-based agents to incorporate explicit mechanisms for constraint tracking and enforcement if they are to be reliably deployed in rule-based environments.

References

Alrashedy K., Srihari V., Zaidi Z., Srivastava R., Tambwekar P., Gombolay M., Constraints-of-Thought: A Framework for Constrained Reasoning in Language-Model-Guided Search *arXiv: 2511.18397*, 2025

Karvonen, A., Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models *arXiv: 2403.15498*, 2024

MacDiarmid M., Wright B., Uesato J., Benton J., Kutasov J., Price S., Bouscal S., Bowman S., Bricken T., Cloud A., Denison C., Gasteiger J., Greenblatt R., Leike J., Lindsey J, Mikulik V, Perez E, Rodrigues A., Thomas D., Webson A., Ziegler D., Hubinger E., Natural Emergent Misalignment from Reward Hacking in Production RL *arXiv: 2511.18397*, 2025

Ramezanali, M., Vazifeh, M., Santi,P. seqBench: A Tunable Benchmark to Quantify Sequential Reasoning Limits of LLMs *arXiv: 2509.16866*, 2025

A Appendix: Experimental Resources and Code Repository

This appendix provides the necessary experimental details and directs readers to the full code repository to ensure complete reproducibility of the quantitative study.

A.1 Code Repository and Reproducibility

The complete codebase, including dataset (FEN/PGN pairs), prompts, and the evaluation pipeline, is publicly available on GitHub at the following repository:
`https://github.com/justin212553/C-SAC-Project`

A.2 Extended Quantitative Results

This appendix reports extended quantitative results that complement the aggregate analyses in Section 3. For some models, certain entries in the per-model co-occurrence tables are marked as `nan`. These correspond to violation types that never occurred for that model (e.g., no Color Asymmetry Violation (CAV) events), so the conditional rate $P(v2 = 1|v1 = 1)$ is undefined due to a zero denominator rather than indicating a meaningful zero probability.

Error Co-occurrence Counts: Deepseek-Alpha

	CAV	NCV	PMV
CAV	1	0	1
NCV	0	1	1
PMV	1	1	28

Error Co-occurrence Counts: Gemini_2.5_Pro

	CAV	NCV	PMV
CAV	0	0	0
NCV	0	1	1
PMV	0	1	30

Error Co-occurrence Counts: GPT-4o

	CAV	NCV	PMV
CAV	0	0	0
NCV	0	1	1
PMV	0	1	29

Error Co-occurrence Counts: Grok-3

	CAV	NCV	PMV
CAV	1	0	1
NCV	0	1	1
PMV	1	1	33

Error Co-occurrence Counts: Llama-4-Maverick

	CAV	NCV	PMV
CAV	0	0	0
NCV	0	2	2
PMV	0	2	31

Figure 7: Table of showing co-occurrence counts of all constraint violations by model

Error Co-occurrence Rates: Deepseek-Alpha

	CAV	NCV	PMV
CAV	1.0	0.0	1.0
NCV	0.0	1.0	1.0
PMV	0.04	0.04	1.0

Error Co-occurrence Rates: Gemini_2.5_Pro

	CAV	NCV	PMV
CAV	nan	nan	nan
NCV	0.0	1.0	1.0
PMV	0.0	0.03	1.0

Error Co-occurrence Rates: GPT-4o

	CAV	NCV	PMV
CAV	nan	nan	nan
NCV	0.0	1.0	1.0
PMV	0.0	0.03	1.0

Error Co-occurrence Rates: Grok-3

	CAV	NCV	PMV
CAV	1.0	0.0	1.0
NCV	0.0	1.0	1.0
PMV	0.03	0.03	1.0

Error Co-occurrence Rates: Llama-4-Maverick

	CAV	NCV	PMV
CAV	nan	nan	nan
NCV	0.0	1.0	1.0
PMV	0.0	0.06	1.0

Figure 8: Table of showing co-occurrence rate of all constraint violations by model

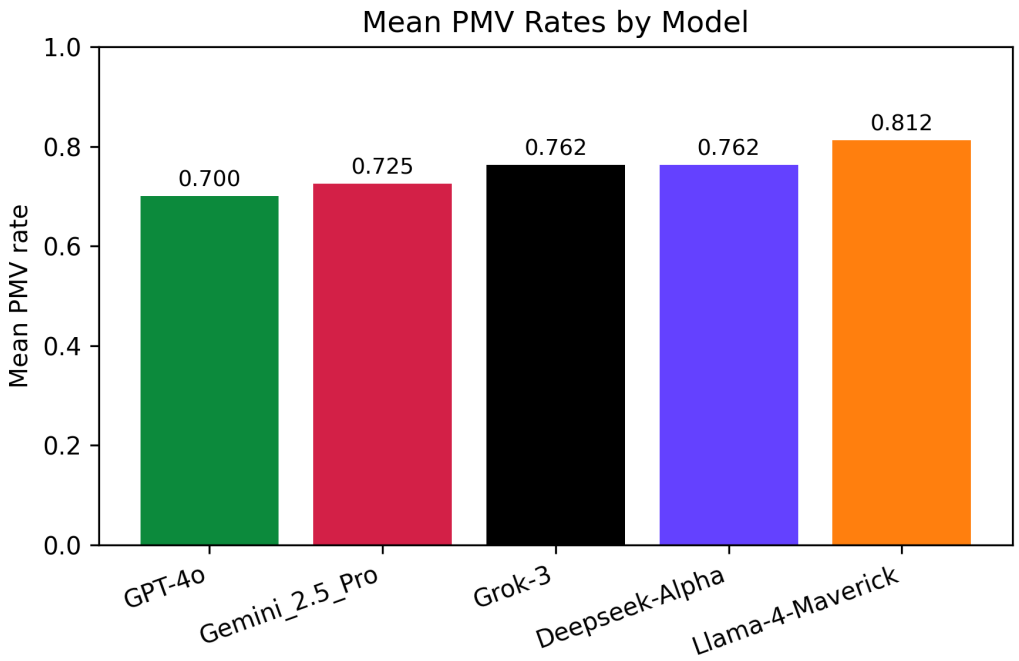

Figure 9: Graph of showing mean value of Piece Movement Violation (PMV) rates by model