Annotated Bibliography: Spatial Reasoning, Architectural Scaling, and Cognitive Limits in LLM Game-Benchmarks

Brook Animut

Department of Computer Science, Purdue University

ENGL 30400: Advanced Composition

[Instructor Name]

October 27, 2026

References

This annotated bibliography collates research and resources investigating the spatial tracking capabilities, formal logic constraints, and cognitive collapse points of large language models through automated game environments, behavioral game theory, and real-world integration frameworks.

Babbar-Sebens, M., Mukhopadhyay, S., Singh, V. B., & Piemonti, A. D. (2015). A web-based software tool for participatory optimization of conservation practices in watersheds. *Environmental Modelling & Software*, *69*, 111–127. https://doi.org/10.1016/j.envsoft.2015.03.011

Meghna Babbar-Sebens and co-authors, including computer science professor Dr. Snehasis Mukhopadhyay, specialize in interactive data science, optimization algorithms, and web-based

computational tools. In this peer-reviewed study, the authors design participatory software frameworks that allow users to interact dynamically with automated optimization engines. This article provides a valuable methodological reference for building transparent, interactive auditing interfaces, supporting the architecture used to track state changes and errors in automated testing environments.

Huang, J., & Chang, K.-W. (2023). Towards reasoning in large language models: A survey. *Findings of the Association for Computational Linguistics: ACL 2023*, 2623–2640. https://doi.org/10.18653/v1/2023.findings-acl.381

Jie Huang and Kai-Wei Chang are computer science researchers specializing in natural language processing and reasoning evaluation frameworks. In this peer-reviewed conference paper, the authors categorize the various ways language models attempt multi-step problem solving, including arithmetic, symbolic, and spatial reasoning tasks. They highlight that while prompt engineering and reasoning tokens improve short-term consistency, models consistently fail when tasks require persistent, non-degrading state tracking over long sequences. This source anchors my research paper's discussion on the boundary between surface-level text generation and genuine structural logic in AI systems.

Karvonen, A. (2024). Emergent world models and latent variable estimation in chess-playing language models. *arXiv*. https://doi.org/10.48550/arXiv.2403.15498

Adam Karvonen is a researcher focusing on mechanistic interpretability and how transformer models internally represent state spaces in rule-bound environments. In this paper, Karvonen explores how language models handle strict state-tracking tasks by analyzing internal activations

during multi-step logical tasks like chess move generation. He examines how statistical pattern-matching can drift out of alignment with the actual state over extended trajectories. This source directly supports my empirical findings on spatial drift—demonstrating that models fail not due to a lack of general language fluency, but because their internal coordinate mapping degrades over long execution horizons, eventually resulting in ungrounded moves and illegal trajectories.

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv*. https://doi.org/10.48550/arXiv.2002.06177

Gary Marcus is a scientist, author, and professor emeritus at New York University who focuses on cognitive science and the fundamental limitations of deep learning systems. In this paper, Marcus argues that current neural network architectures lack robust mechanisms for symbolic reasoning, factual tracking, and consistent rule adherence, relying instead on surface-level statistical association. This source helps frame my research argument by explaining why language models fail reliably at turn-based strategy games like chess: they simulate language patterns about chess rather than maintaining a true, internally consistent logical model of the board state.

Mostafa, J., Mukhopadhyay, S., Palakal, M., & Lam, W. (1997). A multilevel approach to intelligent information filtering: Model, system, and evaluation. *ACM Transactions on Information Systems*, *15*(4), 368–399. https://doi.org/10.1145/263479.263481

Javed Mostafa, Dr. Snehasis Mukhopadhyay, and their colleagues are prominent researchers in information systems, artificial intelligence, and intelligent retrieval architectures. In this

peer-reviewed foundational study, the authors explore multi-tiered evaluation models for filtering relevant data and managing complex operational constraints. This article provides a historical and structural lens on how automated systems process multi-level logical inputs, supporting the evaluation of architectural performance under stress.

Noever, D., Ciolino, M., & Kalin, J. (2023). Large language models on the chessboard: A study on ChatGPT's formal language comprehension and complex reasoning skills. *arXiv*. https://doi.org/10.48550/arXiv.2308.15118

David Noever and his co-authors are computational scientists specializing in artificial intelligence constraint satisfaction and automated strategic reasoning. In this study, the authors test large language models on formal rule-bound systems—specifically chess—to evaluate how strictly models adhere to rigid spatial laws and valid move validation. They analyze the frequency of illegal piece movements and rule violations under standard prompting conditions. This source integrates smoothly into my literature review to explain why language models commit illegal moves (such as trying to move through a pinned piece or ignoring existing board topology) even when explicitly prompted with strict legal parameters.

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., Ruiz, F. J. R., Ellenberg, J. S., Wang, P., Fawzi, O., Kohli, P., & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature*, *625*(7995), 468–475. https://doi.org/10.1038/s41586-023-06924-6

Bernardino Romera-Paredes and researchers from Google DeepMind specialize in neuro-symbolic optimization, evolutionary algorithms, and automated mathematical discovery.

In this peer-reviewed Nature study, the authors introduce FunSearch, a framework that pairs a pre-trained large language model with an automated deterministic evaluator to guard against hallucinations and incorrect logical leaps during complex combinatorial searches. They demonstrate that while standalone language models routinely confabulate invalid proofs over extended reasoning chains, coupling token generation with strict external verification enables genuine algorithmic discoveries. This source directly supports my thesis that overcoming late-stage state collapse in LLM reasoning requires hybridizing probabilistic language models with deterministic rule-checking engines.

Rozman, L. (2026, October). *GPT-6 Astra vs Claude Opus 5.5 (10,000 Elo chess)* [Video]. YouTube. https://www.youtube.com/@GothamChess

Levy Rozman (known professionally as GothamChess) is an International Master, chess educator, and prominent digital media analyst who evaluates high-level human and machine chess play for millions of viewers. In this public media breakdown, Rozman analyzes a high-stakes, deep-horizon self-play match between frontier reasoning architectures (gpt-6-astra and claude-opus-5.5), highlighting tactical brilliant moves juxtaposed against sudden, unforced spatial hallucinations in the late midgame. While traditionally categorized as popular media rather than a peer-reviewed journal, this source provides vital qualitative context on how modern frontier reasoning models behave under pressure in public deployments, validating the real-world occurrence of late-stage cognitive cliffs outside of isolated laboratory scripts.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general

reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

David Silver and the DeepMind research team are pioneers in artificial intelligence, reinforcement learning, and automated game-tree search architectures. In this peer-reviewed foundational paper, the authors establish the principles of self-play reinforcement learning, demonstrating how agents learn optimal spatial strategies and long-horizon planning through recursive trial rather than static imitation. This source provides the theoretical baseline for why automated self-play is the gold standard for stress-testing complex reasoning limits, directly supporting the methodology used in my empirical study of language model cognitive bottlenecks.

Torstensson, O., Prokopovych-Tkachenko, D., Desiatko, A., Hbur, Z., & Britchenko, I. (2026). Risk-driven deployment and forensic readiness framework for AI-enabled security operations centers. *Journal of Cybersecurity and Privacy*, *6*(5), 149.

Olga Torstensson and co-authors are security and information technology researchers focusing on risk management and operational AI deployment. In this peer-reviewed journal article, the authors examine the challenges of integrating automated intelligence systems into operational environments where system opacity and logic failures can introduce critical vulnerabilities. This source provides essential perspective on where it is hazardous or ineffective to rely purely on opaque machine learning systems without rigorous human oversight or deterministic verification frameworks.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, *625*(7995), 476–482. https://doi.org/10.1038/s41586-023-06747-5

Trieu H. Trinh and his colleagues at Google DeepMind and New York University are artificial intelligence researchers specializing in automated theorem proving and neuro-symbolic spatial reasoning. In this peer-reviewed Nature article, the authors present AlphaGeometry, a theorem prover that overcomes the spatial and deductive limitations of pure neural language models by pairing a language model with a symbolic deduction engine. They show that while standalone transformer models struggle to maintain geometric and spatial consistency over deep multi-step proofs, a symbolic engine grounds the search tree and prevents structural drift. This source reinforces my research argument regarding why pure autoregressive models suffer spatial collapse in board games and how symbolic state-tracking mitigates horizon degradation.

Wang, S., Ji, L., Wang, R., Zhao, W., Liu, H., Hou, Y., & Wu, Y. N. (2024). Explore the reasoning capability of LLMs in the chess testbed. *arXiv*. https://doi.org/10.48550/arXiv.2411.06655

Shu Wang and a team of artificial intelligence systems researchers focus on the limits of multi-step planning, computational lookahead, and constraint degradation in modern language models. In this paper, they investigate the horizon limits of deep-turn reasoning using chess as a primary stress-test sandbox. They evaluate how compounding computational overhead leads to attention degradation and sudden logical drop-offs. This source provides theoretical grounding for my observation of the cognitive cliff around plies 50–66, proving mathematically and

empirically that token generation constraints and spatial memory decay place a hard upper bound on a model's capacity for deep endgame calculation.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. https://doi.org/10.1038/s41562-023-01659-w

Taylor Webb, Keith J. Holyoak, and Hongjing Lu are cognitive scientists and computational psychologists investigating abstract relational and spatial reasoning in artificial neural networks. In this peer-reviewed study, the authors evaluate large language models on visual-spatial matrix problems and formal analogical reasoning benchmarks to compare machine performance against human cognitive patterns. While they find strong emergent zero-shot reasoning on short-horizon relational mappings, they note that model robustness remains sensitive to structural complexity and sequential representation limits. This article provides a foundational cognitive science benchmark for my paper, contrasting early-stage emergent reasoning competence with the late-stage state tracking collapse observed in extended multi-turn games.