

ChessLab: A Framework for Measuring Argumentation Methodology Effectiveness

Mario Larsen

Independent Researcher

uspectacle@gmail.com

Introduction

The *wisdom of crowds* hypothesis states that aggregating the judgments of diverse and independent agents can yield decisions superior to those of individuals (Surowiecki 2004). While this effect is well documented in estimation and forecasting tasks, its applicability to complex, sequential decision-making domains such as chess remains an active area of inquiry.

Historical crowd-versus-expert chess matches and large online events (Chess.com 2025) suggest that crowds can be competitive, yet these settings are often anecdotal and lack rigorous experimental control. More recent systematic evidence demonstrated that large crowds of casual players could rival 1900 Elo artificial opponents over hundreds of games (Moussaïd and Garnier 2022). However, research indicates that simple majority voting may not always be optimal (Prelec, Seung, and McCoy 2017).

Chess engine systems, ranging from search-based algorithms to human-mimetic neural networks (Maia Team 2024), offer adversaries with precisely controllable strength, diversity, and reproducibility. At the same time, large language models are beginning to reach grandmaster levels, showing signs of an emerging internal model of the game (Karvonen 2024; Carlini 2023). This motivates the development of a controlled framework in which collective intelligence can be studied without the confounds inherent to human experiments. We propose **ChessLab** to fill this gap by enabling large-scale, reproducible experiments on artificial crowds in chess, with explicit control over ensemble composition and aggregation mechanisms.

Relevance to Argumentation

This work is intended to serve as a tool for studying voting mechanisms and deliberation strategies between artificial agents in a chess context. Chess provides an objective evaluation signal (game outcomes and engine-based position scores) that allows the effectiveness of various methodologies to be compared with arbitrary precision.

Conclusion

The modular framework supports controlled manipulation of agent diversity, aggregation rules, and opponent strength and provide a range of evaluation and visualization tools.

Results are computed and analysed in parallel, offering the ability to scale experiments and test variables independently. The source code is available at <https://github.com/Uspectacle/chesslab> under the GNU v3 License.

Although our current main contribution lies in providing this infrastructure, we also provide experimental results to compare against previous studies conducted on humans. An accompanying roadmap outlines a sequence of experiments designed to explore optimized deliberation mechanisms.

Acknowledgments

This research is conducted independently by the author while awaiting a PhD opportunity in collaborative writing and collective intelligence. Due to limited compute capacity, only lower-strength engines were experimented upon. Nonetheless, this project aims to enable further experimentation and interesting discoveries.

References

- Carlini, N. 2023. Playing chess with large language models. <https://nicholas.carlini.com/writing/2023/chess-llm.html>. Blog post.
- Chess.com. 2025. Magnus carlsen held to draw by 143,000 players. <https://www.chess.com/news/view/the-world-forces-draw-in-historic-game-vs-magnus-carlsen>. Accessed May 2025.
- Karvonen, A. 2024. Emergent world models in chess-playing language models.
- Maia Team. 2024. Maia chess engine: Human-like chess ai. Maia-2 with skill-aware attention mechanism.
- Moussaïd, M., and Garnier, S. 2022. Mechanisms of collective intelligence in chess. Research presented and partially unpublished. Conference presentation at the ACM Collective Intelligence Conference 2022. Extended findings discussed in *A-t-on besoin d'un chef?* (2025).
- Prelec, D.; Seung, H. S.; and McCoy, J. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532–535.
- Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Vintage.