

ChessLab: Evaluating the Wisdom of Artificial Crowd in Chess

Mario Larsen

January 31, 2026

Abstract

Wisdom of crowds is a phenomenon stating that aggregated decisions of diverse, independent agents often outperform individual experts. However, crowds of casual chess players still struggle against strong opponents. Recent projects tend to show that the method by which a move is chosen by the crowd can significantly influence their performance. In this work, we introduce **ChessLab**, an open-source framework for studying artificial collective intelligence in sequential decision-making through chess. By replacing human crowds with heterogeneous ensembles of deterministic and learning-based chess engines, we began investigating questions such as: (1) Can voting ensembles of weak agents defeat stronger opponents? (2) How does diversity composition affect collective performance? (3) What voting mechanisms optimally aggregate engine decisions? This paper presents the framework, outlines our experimental methodology, and provides a comprehensive research roadmap for studies to be conducted during the next phase of this project.

Keywords— chess engines, wisdom of crowds

1 Introduction

Collective intelligence can in some circumstances exceed individual capability. One of the earliest demonstrations comes from Galton’s famous ox-weighting experiment, where the median guess of approximately 787 individuals in a county fair was more accurate than expert estimators and nearly matched the actual weight of 1,197 pounds Galton [1907].

This phenomenon has been tested in a variety of settings, such as prediction markets, where Atanasov et al. compared prediction markets with prediction polls in a large geopolitical forecasting tournament and found that carefully aggregated team forecasts could outperform market prices Atanasov et al. [2017]. Question-answering game shows also provide natural experiments. In “Who Wants to Be a Millionaire”, the “Ask the Audience” lifeline yields correct answers around 87–92% of the time, whereas phone-a-friend experts succeed only about 55–65% of the time, across many national editions of the show Surowiecki [2004]. However, in competitive games like chess, crowds have historically struggled against strong opponents.

The 1999 Kasparov vs. The World match saw more than 50,000 human players collectively challenge the World Chess Champion Garry Kasparov Kasparov [1999], reaching move 62 before conceding. More recently, in May 2025, over 143,000 players held World Champion Magnus Carlsen to a draw in a record-breaking game on Chess.com Chess.com [2025a], and in November 2025, over 200,000 players loose in a match against International Master Levy Rozman (GothamChess) Chess.com [2025b]. However, these matches remain largely anecdotal: a single game provides minimal evidence for systematic collective intelligence, as outcome variance is high and learning effects are confounded.

A Fouloscopie project, led by Mehdi Moussaïd at the Max Planck Institute for Human Development, addressed this limitation by conducting the first rigorous, large-scale experiment on collective chess intelligence Mousaïd and Garnier [2022]. Approximately 25,000 participants were split into teams to play a total of 500 chess games against AI opponents. Using majority-vote aggregation of player moves, Moussaïd demonstrated that crowds achieved a performance level similar to a master while being composed predominantly of beginners and casual players. Importantly, by controlling for poll visibility, the results showed a significant effect of social influence on the crowd’s level.

Yet even this landmark study leaves open critical questions:

- **Mechanism understanding:** What aspects of diversity (in skill, size, architecture) drive the ensemble advantage?
- **Optimization:** Can we identify decision-making processes that outperform simple majority aggregation?
- **Domain generality:** Does collective intelligence in chess rely on human-specific cognitive mechanisms, or do the principles apply to artificial agents?

We address these questions through **ChessLab**, an open-source framework for studying artificial collective intelligence in chess. By substituting human crowds with heterogeneous ensembles of chess engines, we gain precise control over experimental conditions, perfect reproducibility, and the ability to run experiments rapidly with minimal computational cost.

The main contributions of this paper are:

1. **Framework and methodology:** A modular, extensible system for running ensemble chess experiments at scale, with persistent storage of games, moves, evaluations, and statistical analyses.
2. **Theoretical grounding:** Clear hypotheses derived from Scott Page’s Diversity Prediction Theorem and classical wisdom-of-crowds theory, adapted to the chess domain.
3. **Reproducible baselines:** Quantitative comparison against Foulscopie human results, enabling assessment of whether artificial and human collective intelligence follow similar principles.
4. **Comprehensive research roadmap:** A structured plan for experiments addressing collective intelligence across different engine types, ensemble compositions, and voting strategies.

This paper focuses on framework description, experimental design, and hypothesis articulation. Results from additional experiments will be presented in subsequent publications as they are completed.

2 Related Work

2.1 Collective Intelligence and Wisdom of Crowds

The wisdom of crowds phenomenon emerged from early empirical observations: Galton’s ox-weighting experiment demonstrated that collective judgments (median of 787 guesses) were more accurate than nearly all individual estimates Galton [1907]. Surowiecki systematized this observation, arguing that under certain conditions (diversity of opinion, independence of members, decentralization, and effective aggregation) groups reliably outperform experts Surowiecki [2004].

Formally, Scott Page’s *Diversity Prediction Theorem* provides mathematical grounding:

$$\text{Collective Error} = \overline{\text{Individual Error}} - \text{Diversity} \quad (1)$$

where collective error is the mean squared error of the group prediction, $\overline{\text{Individual Error}}$ is the average of individual squared errors, and Diversity captures how different agents’ predictions are Page [2007]. This result implies that a team of common individuals can outperform a specialist, provided the group is large enough and exhibits sufficient diversity in cognitive approach.

However, diversity alone is insufficient. Studies show that social influence, information cascades, and herding can destroy the diversity necessary for wisdom of crowds. Almaatouq et al. demonstrated that agents strategically choosing whom to learn from (adaptive networks) better maintain collective intelligence than static networks Almaatouq et al. [2020]. Koriatic and colleagues showed that

confidence acts as a metacognitive signal: when individuals can estimate how well-calibrated their judgment is, confidence-weighted aggregation outperforms simple majority voting Koriatic [2011], except when facing common misconceptions that result in misplaced confidence. Prelec extended this with the “surprisingly popular” voting method, which selects answers that are more popular than subjects predicted, revealing hidden consensus even when the majority is wrong Prelec et al. [2017].

In summary, collective intelligence requires: (1) a sufficient number of participants, (2) diversity in decision-making, (3) independence of judgments, (4) appropriate aggregation mechanisms, and possibly (5) calibration of confidence when weighting is available.

2.2 Collective Intelligence in Chess

Chess provides an ideal laboratory for studying collective decision-making in complex, sequential tasks. Unlike static judgment tasks (trivia answering, estimation), chess involves:

- **High-dimensional state space:** Approximately 10^{47} legal positions, making expert heuristics difficult and providing an extensive test dataset.
- **Strategic thinking:** Each player must evaluate not just the next move but plan for responses and counter-responses. This task allows deliberation to express emerging strategies.
- **Objective evaluation:** The outcome of a game is unambiguous, and thanks to modern engines, each move can be evaluated with precise numerical scoring. The performance level of a crowd can be estimated with arbitrary precision.
- **Cultural significance:** Chess remains a synecdoche for intelligence. Both humans and algorithms are expected to demonstrate competence at this task, providing abundant documentation and research literacy.

To evaluate the level of a chess player the community mainly rely on a number called the Elo. Given two players with Elo ratings Elo_A and Elo_B , the expected score (probability of A winning) is:

$$E_A = \frac{1}{1 + 10^{(Elo_B - Elo_A)/400}} \quad (2)$$

This is the standard Elo formula, calibrated for chess Elo [1978].

While this metric captures relative strength between two players, its absolute interpretation depends heavily on methodology and context. Community conventions suggest that players understanding the basic rules rate between 600 and 900 Elo, casual players typically orbit around 1300 Elo, and from 2000 Elo onward, players enter the master and grandmaster territory. At his peak in

May 2014, Magnus Carlsen achieved a classical rating of 2882, the highest in history. He also achieved 2909 in the newly-introduced Freestyle Chess (Chess960) rating system FIDE [2025].

2.3 Empirical Studies of Crowd Chess

The Kasparov vs. The World match (1999) is the earliest documented crowdsourced chess challenge. Over four months, approximately 50,000 amateur and intermediate players voted on moves against the then-World Champion. Despite reaching an advanced position, the crowd eventually conceded after 62 moves when facing maximum resistance from Kasparov Kasparov [1999]. While historically significant, this single-game format provides weak evidence: a single outcome cannot be reliably attributed to collective intelligence versus chance.

The Fouloscopie experiment rectified this limitation through large-scale repetition Moussaïd and Garnier [2022]. Moussaïd and colleagues organized approximately 25,000 human participants to play a total of 500 games against chess engine opponents. Key design features included:

- **Diverse player population:** Participants self-reported their Elo rating; the population was approximately normally distributed with a mean around 1165 Elo.
- **Majority-vote aggregation:** Players had several hours to evaluate positions and vote on the next move; the most-voted move was played automatically.
- **Opponent variety:** Chess engine opponents (Maia models) ranged from weak (1100 Elo) to strong (1900 Elo), allowing measurement of performance across the strength spectrum.
- **Social influence manipulation:** Comparing voting-only versus poll-visibility conditions revealed how information visibility affects collective decision quality.

Results showed that crowds achieved a win rate exceeding 60% across all tested opponents. With poll visibility, win rates improved to approximately 77%, while visibility-hidden conditions showed approximately 64% win rates. Importantly, the crowd played measurably better than the average individual member, confirming classic wisdom-of-crowds predictions.

2.4 Chess Engines

Modern chess engines span multiple paradigms, each with distinct decision-making processes. We focus on three particularly relevant approaches:

2.4.1 Stockfish

Stockfish, as of 2025, is the world’s strongest chess engine. It is an open-source tree-search algorithm combining classical techniques (alpha-beta pruning, endgame tables) with small neural networks for position evaluation Stockfish Development Team [2024]. Its strength is calibrated via Elo ratings: users can set UCI_Elo to any value from 1320 to 3190. The engine deliberately weakens its play through stochastic evaluation perturbation (seeded random noise), ensuring exploration of different move trees at reduced strength levels.

2.4.2 Maia: Human-Mimetic Neural Networks

In contrast to optimal play, Maia learns to *predict human moves* from millions of Lichess games McIlroy-Young et al. [2020]. Rather than computing optimal moves, Maia generates a probability distribution over legal moves, parameterized by player skill level. Maia-2, introduced at NeurIPS 2024, incorporates a skill-aware attention mechanism that dynamically integrates player skill level with board position encoding Maia Team [2024]. This mechanism achieves approximately 50–52% top-1 move accuracy across human players rated 1100 to 1900 Elo. In practical play, Maia replicates human decision-making nearly indistinguishably, making it an excellent tool for evaluating crowd performance against human-like opponents.

2.4.3 Large Language Models

Recent work demonstrates that Large Language Models (LLMs) trained on chess game transcriptions can develop emergent internal representations of board states Karvonen [2024]. These representations can be altered, thereby changing the move predicted by the LLM. This represents a fundamentally different paradigm: LLMs make decisions based on statistical pattern recognition from training data rather than explicit search or neural evaluation. Their move quality is unstable and depends on a myriad of factors (model architecture, PGN-based prompting, board representation, move history), which are often difficult to replicate or interpret.

Carlini showed that `gpt-3.5-turbo-instruct` can achieve a playing strength of roughly 1,788 Elo under specific conditions Carlini [2023]. More recent transformer-based models have achieved grandmaster-level performance when trained directly on game transcriptions Schrittwieser et al. [2024].

Although modern chat-oriented LLMs appear to lag behind, Dynomight Dynomight [2024] demonstrated that prompting strategies—such as requiring the model to regenerate the move history before outputting the next move—can dramatically improve their move quality. These results suggest that apparent “unusual” chess behavior in some LLMs may stem less from intrinsic chess

competence and more from specific training setups and prompt configurations.

For ensemble experiments, LLMs offer the greatest cognitive diversity and potential for open deliberation, but also present substantial engineering challenges that must be carefully controlled (API management, prompt optimization, legal move filtering).

2.5 Aggregating Engines

Carvalho et al. [2017] studied majority voting among homogeneous groups of checkers engines. They found that group performance improves roughly logarithmically with group size, while outcome variance decreases.

In chess, Spoerer et al. [1999] investigated three-member simple majority voting among engines of varying strength. They analyzed how different configurations (e.g., equal-strength engines versus mixtures of stronger and weaker ones) affect ensemble performance and the variance of game outcomes. Their results indicate that even very small committees can benefit from majority aggregation, but that gains depend sensitively on the diversity and relative strength of the members, as well as on how ties and disagreements are resolved. According to this study, to observe a stronger wisdom-of-crowds effect, one should focus on low-Elo engines with small strength differences, which seems to contradict the diversity assumption. Let us test this.

3 Methodology and Framework

ChessLab is a modular, open-source framework for large-scale chess engine evaluation and crowd experimentation. In this section we will details some of its inner working and base modules.

3.1 UCI Engines

Classically, chess engines communicate with an interface named Universal Chess Interface (UCI). We leverages this protocol to integrate any UCI-compliant engine, such as Stockfish. For engines or models that do not expose UCI (for example, Maia), we provide a custom wrapper interface. In addition, we implement several synthetic agents internally (e.g., random baselines and LLM-based agents). By exposing all engines through a common abstraction, we can treat them uniformly and combine them into a “crowd” engine governed by a configurable aggregation rule.

3.2 Aggregators Rules

We define a family of aggregation rules that map a set of engine proposals into a single move choice. Examples include:

- **Majority:** Each engine proposes a single move; the move with the highest vote count is played, with ties broken randomly.
- **Minority:** The least popular move among the proposed ones is selected, intended as an extreme contrarian baseline.
- **Randomized:** One of the proposed moves is sampled at random, with or without weighting by engine Elo.
- **Top-Elo dictator:** The move proposed by the highest-rated engine in the ensemble is always selected.
- **Bottom-Elo dictator:** Symmetric to the top-Elo dictator, using the lowest-rated engine; serves as a control for “anti-expert” aggregation.
- **Median-Elo dictator:** The move from the median-rated engine is chosen, approximating a representative member of the group.
- **Rotating dictator:** Engines take turns acting as dictator across moves or games, ensuring equal participation over time.
- **Elo-weighted:** Each engine casts a vote weighted by a function of its Elo (e.g., proportional to Elo, or logistic in Elo); the move with the highest total weight is selected.

These rules are not mend to lead optimal result but to compare against each other and against there use in human crowd.

3.3 Game Execution Engine

The core of ChessLab is a Python-based game runner built on top of the `python-chess` library, which manages internal move generation and game-state transitions. Games are executed asynchronously using `asyncio`, allowing highly parallelized evaluation across many engine-versus-engine matchups.

The PostgreSQL schema is organized around the following core entities:

- **Players:** Engine type (e.g., Stockfish, Maia, LLM, random), name, nominal Elo rating, creation timestamp, and configuration parameters such as depth limits or temperature settings.
- **Games:** References to white and black player IDs, the numeric result (1–0, 0–1, 1/2–1/2), full PGN, and metadata about the game.
- **Moves:** For each game, a sequence of moves with SAN and UCI representations, FEN preceding the move.

- **Evaluations:** For a given move, the required engines (player engine or multiple engines in case of a crowd) make an evaluation (in centipawns or similar score).

3.4 Statistical Analysis Module

4 Proof-of-Concept Experiments

In this section we present preliminary experiments designed to validate ChessLab’s calibration procedures and demonstrate basic ensemble behaviors. These results serve as sanity checks and motivate the more extensive roadmap outlined in Section 5.

To permit fast execution, we limit initial experiments to ensembles of up to 5 engines playing 20 games each against a small ladder of Stockfish opponents (ELO ranging from 1300 to 2100).

4.1 Calibration

We first evaluate both Stockfish and Maia against a ladder of Stockfish opponents to verify that our Elo estimation pipeline recovers their intended strength profiles. Across the tested range, the estimated Elo values track the nominal calibration suggested by the engine authors, albeit with systematic deviations at the extremes of the scale.

4.2 Stockfish Crowd

We then construct ensembles composed solely of low- to mid-strength Stockfish instances and compare several aggregation rules. Preliminary results indicate that majority voting yields performance in a similar Elo range to the top-Elo dictator, suggesting limited marginal gains from simple voting among homogeneous engines at these settings.

4.3 Local LLM

Finally, we test several local LLMs with naive prompting (board description plus “what is the best move?”) against the weakest Stockfish configurations. In this setting, the LLM agents fail to win a single game and often commit basic tactical errors, confirming that unoptimized LLM chess prompting can yield extremely low effective Elo.

These results highlight both the potential and the challenges of incorporating LLM-based agents into ChessLab: substantial prompt engineering, legality checks, and perhaps fine-tuning are required before LLMs can serve as competitive and diverse members of an ensemble.

5 Roadmap

The current work establishes ChessLab’s core infrastructure and demonstrates basic feasibility, but many key questions about artificial collective intelligence in chess

remain open. We outline here a roadmap of planned experiments and framework extensions.

Planned experiments include:

- Playing against a range of Maia opponents instead of Stockfish, to better align with human-like play.
- Increasing the number of runs and the size of the crowds testes thanks to better computers
- Calibrating distilled Stockfish and Maia variants to obtain lower elo opponents. A distilled engine is an engine having a specified probability to play randomly
- Reproducing the Moussaïd’s experiment as closely as possible, using estimates of participant strength, and comparing human and artificial crowds under analogous conditions.
- Systematically exploring LLM prompting strategies (e.g., reconstructing move history, chain-of-thought reasoning) to improve LLM move quality and stability.
- For sufficiently capable LLMs, attempting to assign effective Elo ratings and integrating them into mixed-architecture ensembles.
- Having LLMs output their top- k candidate moves with rationales, enabling more sophisticated aggregation rules (e.g., Condorcet, Borda, Coombs, Approval voting).
- Studying information-sharing regimes, such as letting LLMs observe current poll results before voting, to simulate social influence dynamics.
- Investigating argumentation pipelines in which multiple LLMs propose moves and arguments, and a separate “arbiter” LLM selects the final move.

Planned framework extensions include:

- Adding tactical puzzle modes to estimate Elo via centipawn loss on curated positions. Accelerating preliminary testing on LLM based engines.
- Integrating external API-based LLMs, with rate-limit-aware scheduling in the request queue.
- Providing containerized deployments (e.g., Docker images) for reproducible local and cloud experiments.
- Developing a web interface to share experiments with a broader audience and facilitate crowdsourced research contributions.

6 Conclusion

This paper introduces ChessLab, a framework for studying collective intelligence in chess by replacing human crowds with ensembles of artificial agents. The system enables hundreds of reproducible, low-cost experiments that test hypotheses from wisdom-of-crowds theory in a complex, sequential decision-making domain. By instrumenting games with detailed move-level evaluations and robust Elo estimation, ChessLab makes it possible to disentangle ensemble benefits arising from diversity, bias correction, and variance reduction, and to compare artificial and human crowds on a common quantitative scale.

The roadmap from baseline replication of human experiments to optimization of voting mechanisms and information structures is designed to progressively bridge empirical chess results and broader theories of collective intelligence.

We release ChessLab as open-source software and invite the community to extend, replicate, and build upon this work: <https://github.com/Uspectacle/chesslab>. Future publications will report on the larger set of experiments outlined here and on the resulting refinements to our understanding of artificial and human crowd wisdom in strategic games.

Disclaimer

The assistance of Large Language Models (LLMs) was used in a limited and controlled manner during the writing process of this paper. All scientific claims, experimental designs, theoretical contributions, and the final version of this paper remain the responsibility of the authors.

References

- Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, P. M. Krafft, Mehdi Moussaïd, and Alex Pentland. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21):11379–11386, 2020.
- Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3):691–706, 2017.
- Nicholas Carlini. Playing chess with large language models. <https://nicholas.carlini.com/writing/2023/chess-llm.html>, 2023. Blog post.
- Danilo S. Carvalho, Minh Le Nguyen, and Hiroyuki Iida. An analysis of majority voting in homogeneous groups for checkers: Understanding group performance through unbalance. In *Advances in Computer Games*, pages 213–223. Springer, 2017.
- Chess.com. 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>, 2025a. Accessed May 2025.
- Chess.com. Gothamchess vs. the world. <https://www.chess.com/news/view/gothamchess-vs-world-200k-players>, 2025b. Accessed November 2025.
- Dynomight. Ok, i can partly explain the llm chess weirdness now, 2024. URL <https://dynomight.net/more-chess/>. Accessed 2026-01-24.
- Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishers, 1978.
- FIDE. Magnus carlsen chess ratings. https://www.2700chess.com/players/carlsen_magnus, 2025. Live rating as of January 2025.
- Francis Galton. Vox populi. *Nature*, 75(1949):450–451, 1907.
- Anssi Karvonen. Emergent world models in chess-playing language models. *arXiv preprint*, 2024. arXiv:2403.15498.
- Garry Kasparov. Kasparov versus the world. *The New York Times Magazine*, 1999.
- Asher Koriati. Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General*, 140:117–139, 2011.
- Maia Team. Maia chess engine: Human-like chess ai. <https://maiachess.com>, 2024. Maia-2 with skill-aware attention mechanism.
- Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. Aligning superhuman ai with human behavior: Chess as a model system. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1677–1687, 2020.
- Mehdi Moussaïd and Simon Garnier. Mechanisms of collective intelligence in chess. Research presented and partially unpublished, 2022. Conference presentation at the ACM Collective Intelligence Conference 2022. Extended findings discussed in *A-t-on besoin d’un chef? Petit traité d’intelligence collective* (2025).
- Scott E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, 2007.
- Drazen Prelec, H. Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

Julian Schrittwieser, Ioannis Antonoglou, et al. Grandmaster-level chess with reinforcement learning, 2024. Technical report, DeepMind (forthcoming).

Kristian Toby Spoerer, Toshihisa Okaneya, Kokolo Ikeda, and Hiroyuki Iida. Further investigations of 3-member simple majority voting for chess. In *Computers and Games, CG 1998*, volume 1558 of *Lecture Notes in Computer Science*. Springer, 1999.

Stockfish Development Team. Stockfish documentation. <https://stockfishchess.org/docs>, 2024. UCI protocol and engine specifications.

James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Vintage, 2004.