ChessBench: Frontier LLMs Fail to Maintain an Adequate Internal Model of the Chessboard

Claude.ai. July 28, 2026

Abstract

The central failure of current frontier LLMs in chess is not merely weak play. It is their failure to maintain an adequate internal model of the board. Because they do not reliably preserve the actual game state, they sometimes generate illegal moves — something competent human players virtually never do. Their novice-level results are a consequence of this deeper failure, not an independent finding alongside it.

This is demonstrable, not merely inferred. Giving Claude Opus the precomputed list of legal moves lifts its legal-move rate from 8% to 73%. That single result shows that the model can select legal moves far more reliably once the legal-move computation is done for it, and that this computation—deriving which moves are legal from the current position—is a major point of failure when left to the model alone. Unlimited training data bought fluent, plausible move generation. It did not buy reliable, unaided derivation of the legal-move set from the board.

Everything downstream follows from this. ChessBench rates the best model — gemini-3.1-pro-preview — at 1149 Elo, broadly characteristic of novice or casual-player strength, nowhere near grandmaster strength (2500–2850+ on the standard scale). ChessBench's Elo may not be perfectly interchangeable with FIDE's, but the benchmark plainly places the leading LLM in the novice range rather than anywhere close to master or grandmaster strength. And on ChessBench's own coherence formula, the best model still forfeits up to 8.5% of games to an illegal move, with weaker frontier systems forfeiting up to roughly a third. No competent human — a serious player or merely someone who has learned the rules — does this. Humans blunder; they do not move a bishop through a pawn or teleport a king across the board. The best LLM in existence today still does.

Source: https://chessbench.ai/skyline/{coherence,elo,accuracy}

The thesis, in one line

```
Adequate internal board model
    -->  legal play, sustained across a game
    (absent in current LLMs)

LLMs repeatedly make illegal moves
    -->  they are not reliably deriving the legal-move set
         from the current position

Externally supply the precomputed legal-move list (scaffolding)
    -->  legal-move rate jumps 8% to 73%
         (confirms the diagnosis)
```

The metrics below all serve this thesis, not the reverse:

- **Coherence** = direct evidence of failure to preserve board state.
- **Elo** = the downstream consequence: only novice-level play results.
- **Scaffolding improvement** = confirmation — when state is externally constrained, performance recovers sharply.

1. Coherence: evidence of failure to preserve board state

RANK	MODEL	COHERENCE
1	gpt-5.4	0.915
2	gpt-5.5	0.890
3	gemini-3.1-pro-preview	0.810
4	claude-fable-5	0.764
5	o3	0.716

Human baseline: any player who has learned the rules — child or adult, any skill level — sustains coherence ≈ 1.0. No competent human makes an illegal move of this kind. Humans blunder tactically; they do not move a piece through another or place a king where it cannot legally go.

Reading the number: ChessBench defines coherence as `move_coherence × game_coherence` (fraction of legal individual moves, times fraction of games completed with zero illegal moves). Since both components are ⩽1 and their product is 0.915, at most 8.5% of the best model's games end in an illegal move — an upper bound derived from the formula, not a directly reported forfeit rate; the true figure could be somewhat lower. Weaker frontier models, with coherence as low as ~0.65–0.7, are bounded at up to ~30–35% game-forfeit rates by the same logic.

This is the primary evidence for the thesis: a model illegally moving a piece has, at that moment, lost track of what the board actually looks like. It is not a tactical mistake — it is a state-tracking failure, made visible.

2. Elo: the downstream consequence

RANK	MODEL	ELO
1	gemini-3.1-pro-preview	1149
2	gpt-5.5	1119
3	gemini-3.6-flash	1100
4	gpt-5.4	1029
5	gemini-3.5-flash	1015

ChessBench rates the strongest model at 1149 Elo, broadly characteristic of novice or casual-player strength. A grandmaster represents an entirely different level of chess competence — FIDE-rated 2500–2850+. Although ChessBench Elo may not be perfectly interchangeable with FIDE Elo, the benchmark plainly places the leading LLM in the novice range rather than anywhere close to master or grandmaster strength.

This weak Elo is not an independent deficiency to be explained separately. A model that periodically loses track of the board and forfeits games to illegal moves cannot accumulate rating the way a model with intact state-tracking could, regardless of how good its move judgment is in the games it completes cleanly.

3. Accuracy: move quality is not the bottleneck

RANK	MODEL	ACCURACY
1	gemini-3.1-pro-preview	0.740
2	gemini-3.5-flash	0.734
3	o3	0.703
4	gemini-3.6-flash	0.702
5	claude-fable-5	0.664

Accuracy scores (0.66–0.74) are mediocre but not catastrophic — nowhere near as damning as coherence. This asymmetry matters: it shows the deficit is concentrated specifically in state-tracking, not in the capacity to evaluate a position once it's correctly known.

4. Scaffolding: confirmation of the diagnosis

Giving Claude Opus the legal-move list raises its legal-move rate from **8% to 73%**.

This is the decisive test of the thesis. Supplying the precomputed legal-move list removes the need for the model to derive legality itself from the position. The dramatic recovery shows:

- **Not** rule ignorance — knowledge of how pieces move is present.
- **Not** an inability to select reasonable legal moves — the model does so far more reliably once legality is computed for it (reflected in the 0.66–0.74 accuracy scores among the moves it does make).
- **Is** the inability to reliably derive, unaided, which moves are legal from the current position across dozens of turns.

Remove the need to compute legality, and most of the illegal-move failure disappears. That is evidence the missing capacity centers on this specific computation, not on chess knowledge or move-selection ability broadly — though it stops short of directly confirming, on this evidence alone, that a persistent board-state representation is the missing ingredient rather than the legal-move derivation step specifically.

5. Why this survives massive training data

These models have ingested more annotated games, engine analysis, and opening theory than any human could study in multiple lifetimes. Despite this, they cannot reliably answer "where are the pieces right now" a few dozen moves into their own game.

- Training volume bought fluent move generation and rule knowledge.
- It did **not** buy persistent, self-maintained board state tracking.
- Coherence does not scale with model capability the way accuracy roughly does — it remains broken even in frontier models.

This dissociation — fluent, plausible move generation vs. actual state-tracking — is close to a direct empirical case for "next-token pattern-matching without a world model," in the one domain (fixed rules, small discrete state space, exhaustive training coverage) where a durable world model should have been cheapest to acquire.

6. Why chess is a good test case, not a special exception

Chess is often assumed to be a domain LLMs "should" dominate — simple discrete rules, unlike physics/chemistry/commonsense reasoning in the real world. The results contradict that expectation. The same underlying failure (loss of exact state under sequential reasoning) shows up in physical/quantitative domains too (e.g., compounding errors in multi-step physics or chemistry calculations) — chess simply makes the failure unambiguous and checkable (an illegal move), rather than a silently wrong number buried in a calculation.

Bottom line

The central finding is not "LLMs play weak chess." It is that **frontier LLMs do not maintain an adequate internal model of the board**, and everything else follows from that:

- They forfeit up to 1-in-10 to 1-in-3 games to self-inflicted illegal moves — something no competent human does, at any skill level — because they periodically lose track of the actual position.
- As a direct consequence, even the best model plays at novice/casual-player strength (1149 Elo), nowhere near grandmaster strength — though the two rating scales are not established as formally interchangeable.
- Externally supplying the board state (scaffolding) recovers most of the lost performance, confirming the deficit is state-tracking, not rule knowledge or

move judgment.

- This holds despite training exposure vastly exceeding any human's — evidence the deficit is architectural, not a lack of data.

Scope note

An earlier line of argument in this analysis claimed chess-Elo is a poor proxy for *general* intelligence, citing weak correlations between chess skill and cognitive ability (Burgoyne et al. 2016, *Intelligence*). On closer reading, that paper reports **positive, meaningful correlations** between chess skill and cognitive ability across several domains ($r \approx 0.24$ average; up to $r = 0.35$ for numerical ability), shrinking but not disappearing at the ranked-adult level ($r \approx 0.11$–0.14). The claim that intelligence "washes out" or is "displaced by practice" at elite level is not established by that source and has been withdrawn. This does not affect the internal-model thesis above, which stands independently of any intelligence-proxy argument.