
AI Forecasting Hackathon

report template¹

Emil Schmitz
Independent

With
Apart Research

Abstract

Epoch AI's direct method assumes that lower average loss indicates better general capabilities. We posit that the loss may possibly be indicative only of higher performance on specific content. We attempt to prove this by calculating loss on high-level chess games. To calculate loss, we compare the LLM's prediction to those of open-source chess engine Leela-Zero.

At the time of submission, the experiments have not yet run through. I will try to finish them and notify you, if that works.

Keywords: Direct Method, Chess, LLM

1. Introduction

This project investigates whether an LLM's general training loss serves as a reliable predictor of its performance on specific, narrow tasks. We test this by evaluating how well language models predict chess moves in algebraic notation, comparing their probability distributions against those from Leela Chess Zero, a superhuman chess engine.

The motivation stems from Epoch AI's "Direct Approach" paper, which argues that training loss correlates with an AI system's ability to replicate human performance across all intellectual tasks. If training loss becomes sufficiently low, the theory suggests the model could reproduce any human-written work and thus handle any

¹Research conducted at the Apart Forecasting Hackathon, 2025

intellectual task humans can perform. However, this assumes balanced representation across all task types in the training data. In reality, datasets may contain abundant examples of common content (like blog posts or news articles) but sparse representation of specialized domains like notation of professional-level chess games. A model could achieve low overall loss by perfectly replicating frequent content types while remaining weak on rare, difficult tasks. However, a truly general-purpose AI should be able to predict moves in grandmaster-level games with good accuracy, indicating high chess ability.

We hypothesize that LLMs with lower general training losses will also show lower cross-entropy loss when predicting chess moves compared to Leela Zero's distributions, but that loss on chess will not scale proportionally with general loss. By testing this hypothesis, we examine whether training loss truly generalizes as a capability proxy or whether it masks significant performance gaps in underrepresented domains. Chess serves as an ideal test case: it's well-defined, measurable, has verifiable expert-level performance standards, and existing research shows LLMs struggle at chess without specific training.

2. Methods

2.1 Models Evaluated

- GPT-2 (124M parameters) - baseline small model
- GPT-Neo 1.3B - mid-size open model
- Pythia 1.4B - specifically designed with documented training
- Pythia 2.8B - larger variant with known training trajectory
- Llama 2 7B - recent state-of-the-art open model

Model training losses sourced from original papers and documentation.

2.2 Chess Ground Truth

- Leela Chess Zero (lc0) as superhuman reference
- Rationale: Open source, superhuman strength (3500+ Elo), provides move probability distributions via policy head

2.3 Dataset

- Source: Lichess Elite Database / FICS Games Database
- Number of games: 100
- Game quality: Grandmaster level (Elo > 2500)

2.4 Procedure

1. Load chess games in PGN format
2. For each position in each game:
 - Convert position to text format for LLM (full game history in algebraic notation)
 - Get LLM's probability distribution over next move (via logits over legal moves)
 - Convert position to Leela Zero input format
 - Get Leela's probability distribution over next move (via policy network output)
 - Calculate cross-entropy loss $H(\text{Leela} || \text{LLM})$ between distributions
3. Average loss per game
4. Average loss across all games for each model
5. Correlate with published training/evaluation losses

2.5 Technical Implementation

- Python with PyTorch, Transformers (HuggingFace), python-chess
- Leela Chess Zero via python-lczero or UCI interface

3. Results

...in progress

4. Discussion and Conclusion

...in progress

5. References

1. Epoch AI. (2024). Direct Approach to AI Forecasting. <https://epoch.ai/files/direct-approach.pdf>
2. Ruoss, A., et al. (2024). Grandmaster-Level Chess Without Search. arXiv:2402.04494
3. Leela Chess Zero Project. <https://lczero.org/>
4. Hoffmann, J., et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556
5. Biderman, S., et al. (2023). Pythia: A Suite for Analyzing Large Language Models. arXiv:2304.01373

6. Black, S., et al. (2021). GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

7. Touvron, H., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288

6. Appendix

Potential Limitations:

- Sample size of models tested limited by computational resources
- Chess may not generalize to other domains

Suggestions for Future Work:

- Test on multiple game engines (Stockfish, Komodo) and average their distributions. No engine is optimal. Leela-zero may be biased towards one specific type of play, while another engine may be biased towards another.
- Expand to other domains with verifiable ground truth (mathematical proofs, code correctness)
- Use constrained generation or tool-use paradigm to ensure valid move outputs
- Investigate whether including chess games in training data improves correlation