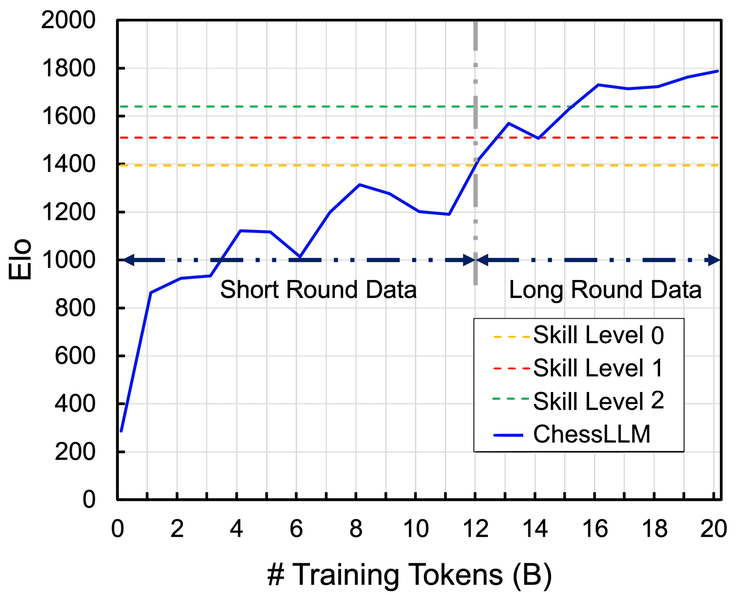
Complete Chess Games Enable LLM Become A Chess Master

Yinqi Zhang[1,4], *Xintian Han*[5], *Haolong Li*[2,5], *Kedi Chen*[1,4], *Shaohui Lin*[1,3]

[1]*School of Computer Science and Technology, East China Normal University*
[2]*School of Computer Science and Technology, Tongji University*
[3]*Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education, China*
[4]*Xiaohongshu,*[5] *Bytedance*

Abtract

Large language models (LLM) have shown remarkable abilities in text generation, question answering, language translation, reasoning and many other tasks. It continues to advance rapidly and is becoming increasingly influential in various fields, from technology and business to education and entertainment. Despite LLM's success in multiple areas, its ability to play abstract games, such as chess, is underexplored. Chess-playing requires the language models to output legal and reasonable moves from textual inputs. Here, we propose the Large language model ChessLLM to play full chess games. We transform the game into a textual format with the best move represented in the Forsyth-Edwards Notation. We show that by simply supervised fine-tuning, our model has achieved a professional-level Elo rating of 1788 in matches against the standard Elo-rated Stockfish when permitted to sample 10 times. We further show that data quality is important. Long-round data supervision enjoys a 350 Elo rating improvement over short-round data.



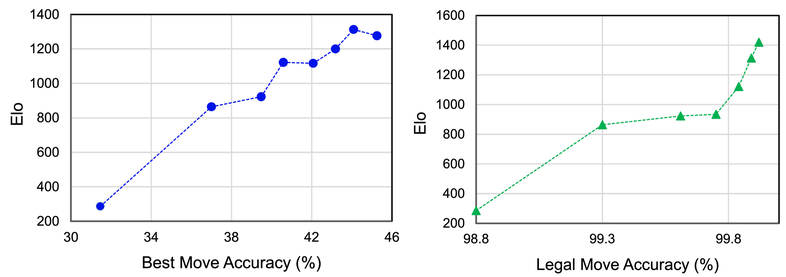
Contributions

- **Dataset.** We collected a large dataset of chess games with over 20B tokens from open-source platforms. Data quality matters; long round data supervision outperforms short-round data by 350 Elo points.
- **Model.** Our ChessLLM is designed to play entire chess games through dialogues. After fine-tuning, it achieved an Elo rating of 1788, winning 61% of games against Stockfish at skill level 0, 56% at skill level 1, and 30% at skill level 2.
- **Eval Method.** We propose evaluation methods based on full games against Stockfish, including move validity, Elo rating, and win rate. We are the only ones using a large language model for chess that can complete full games.

Experiment

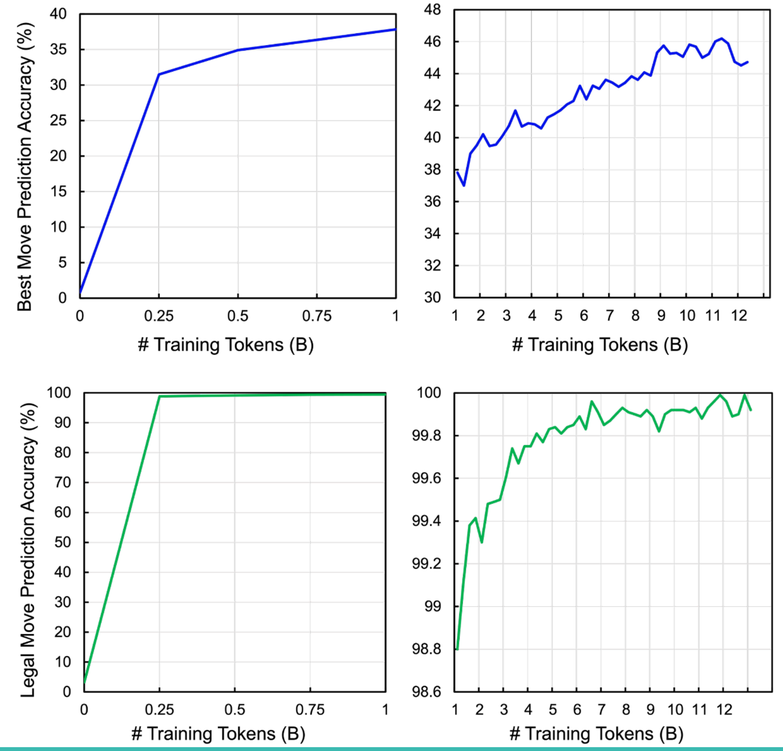
Structure of the data sample.

INPUT: r4rk1 pp6 8 4p2Q 3n4 4N3 qP5P 2KRB3 w - - 3 27 {what is the next best move} <sep>
Target: Qg6+ <eos>



Conclusion

In this paper, we convert chess to a text game and introduce a large-scale Fen-Best Move pair dataset. With the dataset, we propose the Large language model ChessLLM that can play a complete chess game. Considering the limitation of the evaluation set in out-of-distribution data, we propose the need to evaluate model capabilities in actual games. ChessLLM finally achieves an Elo rating of 1788 through the SFT method. In subsequent work, we will discuss how to improve ChessLLM by improving the data quality.



Contact: zhang.inch@gmail.com