# Motivation

Although chess is a “solved” problem in the field of artificial intelligence, chess puzzles can be useful as a testbed to study the capabilities of language models. In addition to puzzles having the helpful property of verifiable outcomes, the popularity of chess makes results legible to a broad audience. If a change to the scaffolding around the LLM yields a significant change in accuracy, that is easily understood and communicable.

Many studies seek to compare LLMs to humans or chess engines. Unlike those researchers, we are not interested in making generalizable claims about LLM chess-playing or strategic reasoning ability. Instead, we are demonstrating a process for improving LLM capabilities on a narrowly defined task and are simply using chess puzzles out of convenience. As a result, this article is more likely to be relevant to individuals using LLMs in applied or business contexts than to academics or researchers at a frontier lab.

# Data

Puzzles are sourced from Lichess’s [chess-puzzles](https://huggingface.co/datasets/Lichess/chess-puzzles) dataset on HuggingFace, which contains 4,679,273 chess puzzles. In addition to the board state and puzzle solution, each record contains additional metadata fields, which are described [here](https://huggingface.co/datasets/Lichess/chess-puzzles#dataset-fields) and may be useful in subsequent analyses. This dataset is updated monthly and our test set was derived from the 2/2/2025 [commit](https://huggingface.co/datasets/Lichess/chess-puzzles/commit/76ea53ea668026b572a07d32d23b5056de62fbe9).

# Baseline Models

For baseline evaluations, we are focused on non-reasoning LLMs (i.e., not DeepSeek’s R1, Google’s gemini-2.0-flash-thinking-exp. or any models from OpenAI’s o-series). For more information on why it wouldn’t make sense to include reasoning models in our reference set, see this [thread](https://x.com/lateinteraction/status/1886814702833164378). While it isn’t necessary to have a perspective on what frontier labs “should” be doing, Omar’s points about how the behavior of reasoners is different from that of language models is salient as we construct our control sample.

We limit our initial list to the following:

* openai/gpt-4o-2024-08-06
* openai/gpt-4o-mini-2024-07-18
* anthropic/claude-3-5-sonnet-20241022
* anthropic/claude-3-5-haiku-20241022
* google/gemini-2.0-flash-001
* google/gemini-2.0-flash-lite-preview-02-05
* google/gemini-2.0-pro-exp-02-05

These models represent the latest generally available non-reasoning models from a variety of providers.

# Baseline