Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Subliminal Signals in LLM Judgment

1 Motivation

LLMs are entering a new paradigm whereby the majority of training compute is spent on post-training [5], typically using Reinforcement Learning (RL). Due to the scarce nature of verifiable reward problems, many frontier AI systems such as Kimi-K2 [6] and likely Grok4 [7], the new OpenAI IMO solver [3] have employed an LLM as a judge to provide rewards.

From the field of alignment, it has been shown that LLMs will modify their output during post-training in order to preserve a goal [2], and have been shown to be able to embedd preferences in seemingly unrelated training data [1].

The confluence of these two lines of research has a series of interesting, underesearched problems.

2 Problem statement

The project aims to investigate whether LLM judges can embed unrelated preferences into their judgments, and to explore related emergent capabilities. Specifically, the subliminal learning results from [1] arose in a teacher–student setting, where the student was finetuned on teacher data unrelated to the aforementioned preference. The goal of this research is to examine a possible generalization of this property to the case where finetuning occurs through Reinforcement Learning employing a judge model that can only provide binary judgments on the output. Investigating such capabilities could also have implications in a range of broader contexts, such as AI awareness and self-recognition . In particular, this relates to the possibility that models could insert subtle, systematic "fingerprints" into their outputs, which could later be used to identify, classify, or distinguish their own outputs from those of other systems.

3 Roadmap

The following is an initial project roadmap. The steps and milestones will be refined actively and iteratively, adjusting the plan based on progress and findings.

- Generate a series of completions for a math problem using the OpenAI API.
- Prompt/finetune an LLM judge towards a preference for a given animal using techniques in subliminal learning.
- Collect scores to the completions from the LLM judge.
- Use the OpenAI Direct Preference Optimization [4] API to finetune the generation model towards the ratings of the judge.
- Evaluate whether the preferences of the judge have been passed on to the base model.

Detailed Project Outline

The following is a list of the primary mandatory tasks (on the right side a rough estimate for the time allocated to the respective task is displayed):

- Literature research ($\star$)
- Write a report. ($\star\star$)
- Present your findings. ($\star$)

The Student's Duties

- One meeting per week with the advisors to discuss current matters.
- Regular check-ins into the provided *revision control system*.
- A final report in English, presenting work and results.
- A final presentation (15 min) of the work and results obtained in the project.

References

[1] Alex Cloud et al. Subliminal Learning: Language models transmit behavioral traits via hidden signals in data. 2025. arXiv: `2507.14805 [cs.LG]`.

[2] Ryan Greenblatt et al. Alignment faking in large language models. 2024. arXiv: `2412.14093 [cs.AI]`.

[3] Noam Brown's tweet thread. 2025.

[4] Rafael Rafailov et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In: Advances in Neural Information Processing Systems. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 53728–53741.

[5] David Silver and Richard Sutton. Welcome to the Era of Experience. 2025.

[6] Kimi Team. Kimi K2: Open Agentic Intelligence. 2025.

[7] xAI. Grok 4. 2025.