

LLM-prompt-optimiser based SAAS platform for evaluations

Iulia Levin
Mesa

Anton Zheltoukhov
Mesa

LLM evaluation SAAS platform built around model based prompt optimiser. Most of the current evaluation benchmarks fall short in terms of quality. Only top tier labs can produce good enough evals (for reducing existential risks). There are multiple well known failure modes. We suggest training a model that can take existing prompts and rewrite them removing known defects from them. That will effectively allow us to absorb every available benchmark and improve its quality. The business part of things: offer evaluation services to various llm-based business applications and eventually frontier models

1. Problem overview

We need lots of high quality evals to build a sufficient awareness of models capabilities in order to be able to detect possible dangerous behaviours. Currently the main sources of such high level evals are the top tier labs (METR, Apollo, AISI), and programs dedicated to bringing more people and guide them towards better quality (like AISIs evals bounty programs and MIRIs technical governance initiative). This means that only a tiny amount of people contribute to the core of the problem.

2. Your solution

We propose training a llm based prompt optimiser whose sole purpose is to rewrite existing evaluation prompts into better versions. The ultimate solution for that probably will be LLM-alignment-scientist (which can do everything for us). As this is still out of our reach, there are more tractable alternatives: llm-as-judge, and even more tractable **llm-prompt-optimiser**. Such optimiser (if successfully implemented) theoretically scales **significantly better than the current paradigm of manual rewriting the benchmarks**.

3. Pilot experiment or demo

4. Process

Timeframe	What will you do?
Next 3 months	Continue to work with our AISC team on Agency Overhang ,

	building an understanding of the shape of the overhang and failure modes of current elicitation methods
2025	Llm-prompt-optimiser, application to prosaic and existential evals
2026	llm-judge
2027	llm-alignment-scientist

5. Impact on AI safety & key risks

- LLM-prompt-optimiser - is a major stepping stone towards automation of alignment work.
- This kind of solutions might introduce new biases and failure modes

6. Appendix

<https://www.lesswrong.com/posts/gJJEjJpKiddoYGZKk/the-evals-gap>

<https://www.lesswrong.com/posts/yxdHp2cZeQbZGREEN/improving-model-written-evals-for-ai-safety-benchmarking>

<https://www.lesswrong.com/posts/fnc6Sgt3CGCdFmmgX/we-need-a-science-of-evals>