Capability Elicitation

capability-elicitation · safety · concept

Source: https://policywindow.org/wiki/capability-elicitation

Generated 2026-11-19T01:05:54 UTC

Summary

Techniques designed to reveal the upper bounds of an AI model's capabilities, rather than measuring its default behaviour, so that downstream safety judgements can be calibrated to what the model *can* do under adversarial prompting or fine-tuning.

At a glance

Used by

9 instrument(s)

Related concepts

alignment, scalable-oversight, red-team-evaluation, deceptive-alignment

Primary source

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., Henderson, P. (2023), 'Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!'

Details

Capability elicitation is methodologically distinct from benchmarking. A benchmark measures average performance under standard prompting; elicitation aims to surface the model's actual capability ceiling. Common methods: (a) adversarial prompting — red-team-style attempts to invoke a withheld behaviour (Branwen 2020, Weidinger et al. 2024); (b) chain-of-thought + structured prompting — forcing step-by-step reasoning, often revealing skills the model would otherwise hide or skip (Wei et al. 2022); (c) multi-stage / decomposition prompting — breaking tasks into sub-tasks that decompose deception incentives (Andersson 2024); (d) fine-tuning pressure — does the safety behaviour break under modest fine-tuning, indicating the underlying capability is preserved (Qi et al. 2023, 'Fine-tuning Aligned LLMs')?

Governance relevance: EU AI Act Art. 55(1)(a) adversarial testing presupposes elicitation methods exist. US EO 14110 §4.2(a) reporting includes red-team results, which depend on elicitation methodology choices. The lack of standardisation across elicitation methods is one reason regulator-mandated evaluation results are not directly comparable across providers (Anthropic's elicitation suite ` OpenAI's ` DeepMind's). The Frontier Foundation Model Eval Consortium is attempting to converge methodology; consensus remains partial.

How to cite this article

APA

Policy Window. (n.d.). Capability Elicitation [Wiki article — Concept]. https://policywindow.org/wiki/capability-elicitation

CHICAGO

Policy Window. n.d.. "Capability Elicitation." Wiki article (Concept). https://policywindow.org/wiki/capability-elicitation.

HARVARD

Policy Window (n.d.) 'Capability Elicitation', Wiki article — Concept, available at: https://policywindow.org/wiki/capability-elicitation.

OSCOLA

Policy Window, 'Capability Elicitation' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/capability-elicitation> accessed [date].

BIBTEX

```
@misc{policywindow-capability-elicitation,
title  = {Capability Elicitation},
author = {Policy Window},
year   = {n.d.},
howpublished = {capability-elicitation — safety},
url    = {https://policywindow.org/wiki/capability-elicitation},
note   = {Primary source: https://arxiv.org/abs/2310.06987}
}
```