In-Context Learning

in-context-learning · safety · concept

Source: https://policywindow.org/wiki/in-context-learning

Generated 2026-05-30T22:09:01 UTC

Summary

The capacity of a foundation model to adapt its behaviour to a new task purely from examples provided in the prompt, without any updates to the model's weights — discovered as an emergent property of large language models and now a primary evaluation surface.

At a glance

Used by

2 instrument(s)

Related concepts

capability-elicitation, multi-turn-evaluation, jailbreak-resistance, agentic-system, inference-time-compute

Primary source

Brown, T., et al. (2020), 'Language Models are Few-Shot Learners' (GPT-3 paper) — the canonical articulation of in-context learning as an emergent capability.

Details

In-context learning (ICL) was named by Brown et al. (2020, 'Language Models are Few-Shot Learners,' the GPT-3 paper) as the surprising observation that sufficiently large language models could perform new tasks from a few demonstrations in the prompt. The phenomenon is empirically robust across scales above ~1B parameters; theoretical accounts (Xie et al. 2022, 'An Explanation of In-context Learning as Implicit Bayesian Inference'; Garg et al. 2022; von Oswald et al. 2023, 'Transformers Learn In-Context by Gradient Descent') propose various mechanisms but no consensus mechanism has emerged.

Governance relevance is methodological. (a) Capability evaluations that test only baseline prompting under-state real-world capability, because deployment prompts routinely include task examples (Wei et al. 2022 chain-of-thought; Anil et al. 2024 many-shot). EU AI Act Art. 55(1)(a) adversarial testing must include ICL-mode probing to be capability-accurate. (b) Safety evaluations that test only baseline refusals under-state real-world failure surface, because many-shot jailbreaking exploits ICL to recover prohibited capabilities (Anil et al. 2024). (c) Model-card disclosures should specify which capabilities are baseline vs ICL-elicited (EU AIA Art. 53 transparency obligation). (d) ICL also affects the open-vs-closed debate: a closed model accessed via API still exposes ICL-elicitation surface, weakening the capability-containment assumption.

How to cite this article

APA

Policy Window. (n.d.). In-Context Learning [Wiki article — Concept]. https://policywindow.org/wiki/in-context-learning

CHICAGO

Policy Window. n.d.. "In-Context Learning." Wiki article (Concept). https://policywindow.org/wiki/in-context-learning.

HARVARD

Policy Window (n.d.) 'In-Context Learning', Wiki article — Concept, available at: https://policywindow.org/wiki/in-context-learning.

OSCOLA

Policy Window, 'In-Context Learning' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/in-context-learning> accessed [date].

BIBTEX

```
@misc{policywindow-in-context-learning,
title  = {In-Context Learning},
author = {Policy Window},
year   = {n.d.},
howpublished = {in-context-learning — safety},
url    = {https://policywindow.org/wiki/in-context-learning},
note   = {Primary source: https://arxiv.org/abs/2005.14165}
}
```