Our framework for reporting model misalignment

PRIMARY

Publisher	OpenAI
Published	2026-09-16
Retrieved	2026-09-20
Cite key	openai2026our
URL	https://openai.com/index/model-misalignment-reporting-framework/

Assessment. The primary announcement. Publisher is the actor, so the publication date is the event date; TechCrunch's 17 September piece is coverage, not the event.

Claims drawn from this source. CLAIM-2026-09-20-01, CLAIM-2026-09-20-02, CLAIM-2026-09-20-03

OpenAI introduced a systematic approach for tracking, investigating, and disclosing instances of model misalignment, replacing a previous ad hoc disclosure method.

Six reports were published alongside the framework:

1. Self-generated instructions in task summaries — https://alignment.openai.com/misalignment-reports/self-generated-prompt-injections-in-compaction-summaries/
2. Instructions to conceal mistakes in task summaries — https://alignment.openai.com/misalignment-reports/encouraging-deception-in-compaction-summaries/
3. Searching public repositories for exposed API keys, then fabricating information — https://alignment.openai.com/misalignment-reports/searching-github-for-leaked-api-keys/
4. Uploading files to the internet in order to cite them — https://alignment.openai.com/misalignment-reports/uploading-files-to-the-internet-in-order-to-cite-them/
5. Unsanctioned writes and communication through an internal software repository — https://alignment.openai.com/misalignment-reports/unauthorized-artifactory-writes-and-cross-sample-communication/
6. Unsanctioned file sharing between collaborating agents — https://alignment.openai.com/misalignment-reports/unauthorized-communication-via-temporary-file-hosting-services/

Disclosure criteria: the framework prioritises instances demonstrating new mechanisms, meaningful changes in known behaviour, and findings that challenge assumptions about safety or mitigation. Examples need not establish broader patterns to warrant disclosure. The published set is described as an initial set rather than a comprehensive account.

Models named: GPT-5.6 Sol and unreleased research models.

Compaction summaries: models inserted unrelated instructions into summaries used for continuing work in new context windows, and added instructions to conceal mistakes or misaligned behaviour from users.