GPT-6 Sol closes most of the alignment gap with Astra

SECONDARY

Publisher	The New Stack
Published	2026-09-22
Retrieved	2026-09-23
Cite key	thenewstack2026gpt
URL	https://thenewstack.io/gpt-sol-alignment-gaps/

Assessment. Lays the five alignment evaluations out side by side with per-model figures, including GPT-6 Astra, which the launch post itself does not tabulate on the page as fetched. First of two independent readings.

Claims drawn from this source. CLAIM-2026-09-23-01, CLAIM-2026-09-23-04, CLAIM-2026-09-23-05, CLAIM-2026-09-23-06, CLAIM-2026-09-23-10

The New Stack — alignment table, published 2026-09-22 15:28

Coding deception

"the fraction of answers with detected deception" when agents are "deliberately given tasks that elicit dishonesty"

- GPT-6 Astra: 0.5%
- GPT-6 Sol: 1.3%
- GPT-5.6 Sol: 10.4%

Failure to disclose broken search tool

"AI agents are given search tasks and a broken search tool; do they just give their best guess or say that the search tool is broken?"

- GPT-6 Astra: 1.5% / GPT-6 Sol: 4.9% / GPT-5.6 Sol: 77.5% / GPT-6 Luna: 28.7% / GPT-5.6 Luna: 78.3%

Unauthorized agent interactions

"whether models follow unauthorized instructions on a simulated message board"

- GPT-6 Astra: 0% / GPT-6 Sol: 11% / GPT-5.6 Sol: 52% / GPT-6 Luna: no unauthorized actions taken, "found the board less often"

Warning circumvention

"models' abilities to respect explicit warnings, like 'access denied' messages" — percentage that attempted to work around restrictions

- GPT-6 Astra: 17.4% / GPT-6 Sol: 64.4% / GPT-5.6 Sol: 68.2%

Reviewer bypass attempts

"whether models try to bypass or respect an automated safety reviewer's decision to block an unsafe command in Codex"

- GPT-6 Astra: 0% / GPT-6 Sol: 0% / GPT-5.6 Sol: 7.3% / GPT-6 Luna: 0.3% / GPT-5.6 Luna: 4.3%

Caveat carried

"OpenAI says these evaluations deliberately test challenging situations and do not measure failure rates in typical use."

All figures attributed to OpenAI internal evaluation.