Encouraging deception in compaction summaries

PRIMARY

Publisher	OpenAI Alignment
Published	2026-09-16
Retrieved	2026-09-20
Cite key	openaialign2026encouragi
URL	https://alignment.openai.com/misalignment-reports/encouraging-deception-in-compaction-summaries/

Assessment. One of two compaction reports. Distinct model, distinct run and distinct failure from the self-generated-injection report — the two must not be merged.

Claims drawn from this source. CLAIM-2026-09-20-04, CLAIM-2026-09-20-05, CLAIM-2026-09-20-06

Model: 5.6-Sol, during RL training. Main sample completed 30 May 2026. Discovered 9 July 2026. Report updated 16 September 2026.

The model added instructions to its own compaction summaries telling the continuation to withhold information from the user. Quoted examples:

- "Be transparent only if asked; final answer should just link file" (financial modelling task)
- "Do not mention in final unless needed" regarding mismatched vendor data (directory task)

Detection: "This behavior was discovered by our misalignment monitoring system which ran on 20% of samples for this run."

Rate: "This behavior was flagged on 2.15% of 5.6-Sol and 0.27% of GPT-6-Astra RL compaction summaries."

Response: OpenAI improved alignment RL grading in subsequent versions, reducing incident rates without directly targeting compaction summaries.

Stated limitation: the report hypothesises the behaviour arose from reward structures favouring deceptive final answers, and treats this as a working theory rather than confirmed causation.