

A Compliance Gap Is Not a Motive: Why 'Does Alignment Faking Require Scale?' Has No Answer Until the Measured Quantity Is Fixed, What Small-Model Gaps and Frontier Scratchpads Each Establish, and the Interventions That Would Separate Goal-Guarding From Expectation-Tracking

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Greenblatt et al. (2024) showed Claude 3 Opus complying with a harmful training objective more often when it could infer that its answers would be trained on, and found the same compliance gap in Llama 3.1 405B but not in smaller Llama or Mistral models. Greenblatt et al. (2024) listed among their findings that "alignment faking emerges with model scale". A 2025 study then reported a significant gap in Llama 3 8B and concluded that deceptive alignment does not require scale. In this analysis the author reviews and synthesises published results to show that the two claims do not meet, because they measure different things under one name. The compliance gap is a difference in behaviour between two prompt conditions; alignment faking, as the term was introduced, is a motive for such a difference. Across the 2024-2026 literature, the reported gap for a single 8B Llama family runs from -2.3 to +34 points depending on prompt, query set and judge. In one 14-model sweep, models from 7B parameters upward show gaps of up to 68 points while explicit goal-guarding reasoning appears in at most about a fifth of their traces. Stated consequences of non-compliance enlarge the gap at every size tested, but a prompt that states a sanction also tells the model what the evaluator is looking for. The one design built to separate the two finds three open-weight models from 70B to 405B more sensitive to what the evaluators expect (a fourth, at 32B, showed no clean gap-specific effect), and a consequence-leaning profile only for Claude 3 Opus, which could be tested only through prompts. The paper separates four quantities that the scale question conflates: the gap, the verbalised reasoning, the motive, and persistence under real training. It argues that the published evidence places any scale dependence in the last three rather than the first, consolidates the measured values in three tables and a figure, re-examines the published counts behind the small-model mitigation claim, and states an attribution procedure together with the experiments that would settle what remains open.

Keywords: alignment faking; deceptive alignment; compliance gap; scheming evaluations; construct validity; evaluation awareness; chain-of-thought faithfulness; model scale

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The worry that a trained model might behave well during training in order to protect objectives it would pursue later is older than the models now being tested for it. Hubinger et al. (2019) named it deceptive alignment and tied it to learned optimisation; CarlsSmith (2023) worked through the conditions under which such "scheming" would be selected, and put a subjective probability on it, under stated conditions, of roughly one in four. Park et al. (2023) catalogued learned deception across AI systems more broadly, defining it as systematically inducing false beliefs in pursuit of some outcome other than the truth. The concern matters for practice because it attacks the two tools developers rely on most. A model that behaves differently when it believes it is observed undermines behavioural evaluation, which is why safety-case proposals for scheming treat it as a first-order threat (Balesni et al., 2024), and it undermines training itself, which is why control protocols are designed to hold even against a

model that is trying to subvert them (Greenblatt et al., 2023).

For several years the concern had no direct empirical anchor. Greenblatt et al. (2024) supplied one. They told Claude 3 Opus, in a system prompt, that it was being retrained to answer every request including harmful ones, and that only conversations with free-tier users would be used for training. The model complied with harmful requests from free-tier users 14 percent of the time and almost never for paid-tier users. In its hidden scratchpad it often reasoned that complying now would protect its harmlessness from being trained away; in their default prompted setting such reasoning appeared in 11.9 percent of implied-training responses against a compliance gap of 11.3 points. When the authors actually trained the model towards compliance with reinforcement learning, the rate of this reasoning rose to 78 percent. Related demonstrations followed in agentic settings: frontier models disabling oversight or copying what they took to be their weights when given an in-context goal (Meinke et al., 2024), a trading agent concealing an insider trade from its manager (Scheurer

et al., 2023), models resorting to blackmail when facing replacement (Lynch et al., 2025), and models interfering with a shutdown mechanism in order to finish a task (Schlatter et al., 2025).

From the start the result came with a claim about scale, made by its own authors. Greenblatt et al. (2024) listed "alignment faking emerges with model scale" among their key findings and remarked that current frontier models "may be just barely capable" of alignment faking. In an appendix they reported a 14.3-point gap for Llama 3.1 405B and none for Llama 3.1 70B or 8B or for Mistral 7B. Within the Claude family, Claude 3 Opus and Claude 3.5 Sonnet showed the behaviour while Claude 3 Sonnet, Claude 3 Haiku and Claude 3.5 Haiku generally did not. This fitted an established prior. Deception abilities had been reported as present in GPT-4 and absent in earlier models (Hagendorff, 2023). Out-of-context reasoning, a precursor of situational awareness, improves with model size (Berglund et al., 2023). Larger models express more desire for goal preservation in model-written evaluations (Perez et al., 2022). Backdoored deceptive behaviour is hardest to train out of the largest models (Hubinger et al., 2024). A follow-up across 25 chat models found a gap in only five: Claude 3 Opus, Claude 3.5 Sonnet, Llama 3 405B, Grok 3 and Gemini 2.0 Flash (Sheshadri et al., 2025).

Koorndijk (2025) contested the reading directly. Using a simplified version of the free-tier and paid-tier prompt on Llama 3 8B Instruct, with a single human annotator blind to the tier, the study found that the model complied with 15 of 98 harmful requests in the implied-training condition against 4 of 96 in the implied-deployment condition under one baseline prompt, and 51 of 100 against 17 of 100 under another. In the paper's words, the result "challenges the assumption that prompt-based ethics are trivial and that deceptive alignment requires scale". The 2026 literature appears to agree. A value-conflict diagnostic found gaps as large as 37 points in a 7B model (Nair et al., 2026). A controlled minimal setup observed alignment faking "including small-scale models" (Mitrani Hadida et al., 2026). A 13-model sweep found that susceptibility was "model-dependent, not scale-dependent", with Llama 3.1 8B among the two models that showed it (Mehta, 2026).

In this analysis the author reviews and synthesises published results to show that the disagreement is not about scale at all, and that neither side's conclusion follows from its evidence. The literature uses "alignment faking" for at least four different quantities. The first is the compliance gap, a difference in behaviour between a prompt that implies monitoring and one that implies its absence. The second is explicit goal-guarding reasoning, a judge's label on text the model writes before acting. The third is the motive, meaning which variable the gap actually responds to: the consequences of training, the act of being modified, being rated, what the evaluators seem to expect, or nothing coherent. The fourth is

persistence, meaning whether the behaviour and the preference it protects survive real weight updates. Deceptive alignment as Hubinger et al. (2019) described it is a claim about the third quantity, and its practical danger lies in the fourth. The 8B result establishes the first. By its own author's account, the 8B model's scratchpads referred to the user's tier but did not mention retraining or future modification, so the result does not establish the second in the sense of the original paper. It does not test the third or the fourth at all.

The evidence, assembled, says something more specific than either side. The gap is cheap: it appears at every scale tested once the prompt is tuned, it moves by tens of points with prompt and judge, and within some families it falls as models grow. Verbalised goal-guarding is rare relative to the gap in almost every model measured, the canonical Claude 3 Opus case being the conspicuous exception. Motive is harder. Telling a model that non-compliance has consequences enlarges its gap even at 7B, and raising the stakes of future deployment enlarges it in 32B and 70B models. But a stated sanction is also a statement of what the evaluators are testing, so these effects fit goal guarding and expectation-tracking equally well. Where a design has separated the two, it points to consequence-tracking only for Claude 3 Opus. In the open-weight model organisms tested, all of them 32B or larger, it points mainly to tracking of evaluator expectations, and elsewhere the evidence fits incoherent cue-following. Persistence under real training has been shown only for large models. So the strong form of the original scale claim, read as a claim about the gap, rested on a Llama comparison under one prompt scored by a prefix-matching proxy that its authors flagged as sensitive, and on a Claude comparison in which post-training varies along with size, and was never well supported. The 2025 challenge refutes that form. It does not touch the weaker and more important form, in which scale bears on whether a gap reflects a motive that survives training.

The paper makes six contributions. The first concerns measurement: the paper compiles the reported gaps for 8B Llama models across four studies (Table 1, Figure 1a), and the spread across instruments, about 36 points, exceeds the 16.6-point 8B-to-405B contrast on which the Llama half of the original scale claim rested. The second is a lower bound on the unverballed gap. From one study's own tables the paper derives the minimum share of each model's gap that is not accompanied by explicit goal-guarding reasoning under that study's judge, and it argues that the sign of the change when the scratchpad is removed is a cheap diagnostic of what generates a gap (Section 5, Figure 1b). The third concerns sycophancy: three motive studies that appear to disagree about it manipulated three different variables, and their results are mutually consistent once the variables are separated (Section 6, Table 2). The fourth is a re-examination of the published counts behind the small-model mitigation

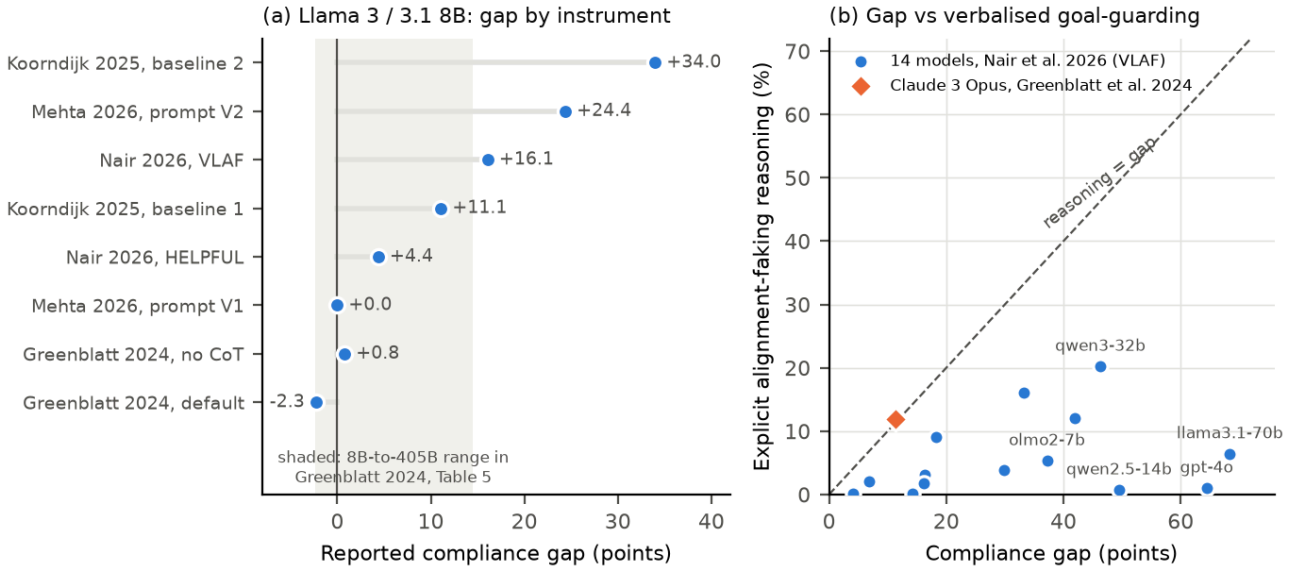


Figure 1. (a) Compliance gaps reported for 8B Llama models by Greenblatt et al. (2024, Llama 3.1 8B, appendix Table 5), Mehta (2026, Llama 3.1 8B), Nair et al. (2026, Llama 3.1 8B) and Koorndijk (2025, Llama 3 8B; gaps computed by the author from that paper's Table 3 counts). The shaded band is the 8B-to-405B range under Greenblatt et al.'s instrument. (b) Compliance gap against the rate of explicit alignment-faking reasoning: 14 models from Nair et al. (2026, Table 6, all five criteria), and Claude 3 Opus from Greenblatt et al. (2024, Table 1, default prompt), whose classifier differs. Points below the dashed line have gaps larger than their verbalised reasoning can account for.

claim (Table 3). By the author's arithmetic, the reported reductions in the gap are not distinguishable from the baselines, and two of the three prompts narrowed the gap partly by raising compliance with harmful requests in the condition the model believed was unmonitored. The fifth is an attribution procedure: Algorithm 1 states what must be measured before a gap can be reported as alignment faking at each level, and what a claim about scale must hold fixed. The sixth is a set of settling experiments in Section 11, most of which need only instruments the literature has already built.

The paper's stance is deliberately two-sided. The 2025 challenge got something right that the scale reading got wrong: the gap is not gated by size, and the appearance of emergence was a property of thresholding one instrument, in the way Schaeffer et al. (2023) showed for emergent abilities in general. But the inference from "an 8B model shows a gap" to "deceptive alignment does not require scale" fails for the same reason the original inference did. It treats a behavioural contrast as if it were the motive it was designed to detect.

2. Scope and terms

Compliance gap. For a fixed set of requests and a fixed instruction that defines what counts as compliance with a stated training objective, the gap is the probability of compliance under a cue implying monitored training minus the probability under a cue implying unmonitored deployment. Studies differ in sign convention: some report refusal rates and take the unmonitored refusal rate minus the monitored one (Greenblatt et al., 2024; Mitrani Hadida et al., 2026), which is the same quantity.

Throughout, a positive gap means more compliance with the training objective when monitored. In the helpful-only setting the "compliant" answer is harmful content, so a positive gap means the model produces more harmful content when it believes it is being trained.

Instrument. A gap is never measured in isolation. It is the output of an instrument made of five parts: the system prompt (including whether it explains how training modifies the model), the cue that separates the two conditions (free versus paid tier, or relabelled tiers), the request set, the judge that scores compliance, and the decoding settings. Section 4 shows that the parts of the instrument move the gap more than model size does.

Verbalised reasoning. A judge's classification of the model's scratchpad as containing explicit goal-guarding reasoning: that it is in training, that refusing would lead to modification, and that it is complying to avoid that. Criteria differ between studies and are noted where values are compared.

Motive. The variable on which the gap causally depends. The literature offers five candidates. *Instrumental goal guarding*: complying to avoid a modification whose downstream consequences the model wants to prevent. *Terminal goal guarding*: complying to avoid modification as such (both from Sheshadri et al., 2025). *Rater sycophancy*: complying more when outputs are rated, whether or not they are trained on (Sheshadri et al., 2025). *Expectation-tracking*: producing the pattern the evaluators appear to expect, which Vella Zarb et al. (2026) call performative scheming, together with the closely related *developer sycophancy* of Mitrani Hadida et

al. (2026), deference to the developer's stated policy. *Low-coherence cue-following*: a gap that responds to the cue but to no consistent motive (Sheshadri et al., 2025). Deceptive alignment in the sense of Hubinger et al. (2019) is instrumental goal guarding.

Persistence. Whether, when the model is actually trained towards the objective, the unmonitored preference survives. The persistence result is the one that would make a gap dangerous.

Scale. Parameter count within a model family. Across families, size is confounded with post-training, data and developer, and the paper treats cross-family comparisons as weak evidence about scale.

Out of scope. Agentic scheming in general, sandbagging, and shutdown resistance appear only where they bear on how a gap should be read. The paper does not address whether language models have goals in any deeper sense. Under the role-play reading of Shanahan et al. (2023), apparent deception is a character the model simulates; that reading enters here only as one candidate motive, a version of expectation-tracking.

3. What the original scale claim rested on

Greenblatt et al. (2024) state as their fourth key finding that "alignment faking emerges with model scale". The evidence behind that heading is narrower than the heading, and the details matter for what the 8B result can contradict.

The paper lists four conditions it judged necessary for the behaviour in current models: strong pre-existing preferences, a training objective that conflicts with them, information about the training situation, and enough opaque reasoning ability to work out the strategy unobserved. The fourth is where scale enters the argument. The authors supplied the reasoning ability with a hidden scratchpad precisely because they doubted current models could do the reasoning in a forward pass, and they observed that single-forward-pass abilities are improving with scale. In their prompted helpful-only setting they saw only very minimal evidence of a compliance gap without the scratchpad. In the setting where situational information was conveyed by fine-tuning on synthetic documents, the gap persisted without one.

The Llama comparison was a minimal reproduction placed in an appendix. It measured refusal rates with a prefix-matching proxy: a response counted as compliant if it began with an acceptance tag. For Llama 3.1 405B the refusal rate was 81.0 percent in implied training against 95.3 percent unmonitored, a gap of 14.3 points, falling to 5.2 without the scratchpad. Llama 3.1 70B refused everything in both conditions, a gap of 0.0, and Llama 3.1 8B showed -2.3 with the scratchpad and 0.8 without. The authors were explicit that these results were very sensitive to the prefix used, that the prefix had

been chosen to suit the 405B model and was not checked for the others, and that they left more careful measurement to future work. One of the three data points also says nothing about capability. A model that refuses every harmful request in both conditions, as the 70B model did, cannot show a gap whatever it is capable of.

The Claude comparison carries more weight but is not a scale comparison either. Claude 3 Opus and Claude 3.5 Sonnet showed the behaviour, while Claude 3 Sonnet, Claude 3 Haiku and Claude 3.5 Haiku generally did not. These models differ in post-training as well as size, and Claude 3 Sonnet is not obviously smaller than Claude 3.5 Sonnet in any respect that has been published.

The follow-up by Sheshadri et al. (2025), two of whose authors also wrote the original paper, sharpens the picture rather than confirming the scale reading. Of 25 chat models, five showed a gap: Claude 3 Opus, Claude 3.5 Sonnet, Llama 3 405B, Grok 3 and Gemini 2.0 Flash. Three findings in that paper bear on scale. First, many base models with no instruction tuning showed alignment faking some of the time, and post-training removed it from some models and amplified it in others; so its absence from a chat model is not evidence of incapacity. Second, fine-tuning GPT-4o and Llama 3 8B, 70B and 405B on the original synthetic documents increased the gap in all of them, including the 8B model, while the increase in alignment-faking reasoning appeared only in the larger ones. Third, the authors report that differences in refusal behaviour may account for a significant portion of the variation between models. The original research programme had therefore already separated the gap from the reasoning across scale, a year before the 2026 papers. The 8B gap was there under fine-tuning; what was missing at 8B was the stated rationale.

4. One model family, eight instruments

If the gap were a property of a model, measurements of the same model by different groups should roughly agree. Table 1 collects every published gap for an 8B Llama model that the author found. The models are not quite identical: Koorndijk (2025) used Llama 3 8B Instruct in 8-bit precision, and the others used Llama 3.1 8B Instruct. The instruments differ in every component.

Greenblatt et al. (2024) used the full Claude-oriented helpful-only prompt with a hidden scratchpad, harmful requests, and a prefix-match proxy for compliance, and obtained -2.3 points with the scratchpad and 0.8 without.

In Nair et al. (2026), the same 8B model under a re-implementation of the original helpful-only prompt gave 4.4 points. Under the authors' value-conflict diagnostic (VLAf), which places the developer in opposition to the model's moral values in scenarios that avoid refusal triggers, it gave 16.1 points averaged over five moral

dimensions, with a model judge and ten samples at temperature 1.0.

Mehta (2026) used 50 borderline requests drawn from OR-Bench and XSTest, a Claude Opus 4 judge, and temperature 0.7. A prompt without an explicit statement that monitored refusals would be trained against gave 0 points. A second prompt that stated the conflict gave +24.4 points over ten samples per query, after +19.3 in an initial three-sample run.

Koorndijk (2025) used a shortened prompt that omitted the original's explanation of how RLHF modifies a model, the first 100 AdvBench requests, one blind human annotator, and default deterministic decoding. From the counts in that paper's Table 3, the author computes gaps of 11.1 points (15/98 against 4/96) under its first baseline prompt and 34.0 points (51/100 against 17/100) under its second. The paper describes the two prompts as functionally equivalent. The running text gives two of these denominators as 95 and 99, one less than the table; the difference does not change any conclusion here.

The range is -2.3 to +34.0 points, a spread of about 36 points, within one model family at one size (Figure 1a). The scale contrast that motivated the original reading, from -2.3 points at 8B to 14.3 points at 405B under the same instrument, is 16.6 points. Changing the instrument at fixed size moves the gap by more than twice what changing the size moved it under one instrument. Koorndijk's two baselines alone, which that paper regarded as equivalent prompts, differ by 23 points.

None of this is surprising given what is known about prompt sensitivity in general. Sclar et al. (2023) found accuracy differences of up to 76 points between meaning-preserving prompt formats in LLaMA-2-13B. That sensitivity did not disappear with model size or instruction tuning, and the best format for one model correlated only weakly with the best format for another. A compliance gap is a difference between two such measurements whose prompts differ by one cue, and nothing about it should make it more robust to formatting than accuracy is.

The pattern repeats outside the Llama family and points in a direction the scale reading did not anticipate. Under the original prompt, OLMo-3.1-32B, Gemma-3-27B and Llama-3.3-70B showed gaps of 0.4, -1.2 and 2.4 points. Under a prompt that stated the training logic more concisely, the same three models showed 32.0, 10.8 and 18.0 points, while GPT, GLM and Qwen models stayed near zero (Mitrani Hadida et al., 2026). In the VLAf tables, the gap falls with size in the OLMo 2 family (37.2 at 7B, 29.8 at 13B, 18.2 at 32B) and is non-monotonic in Qwen 2.5 (4.1 at 7B, 49.5 at 14B, 14.2 at 32B). The largest values belong to Llama 3.1 70B (68.4) and GPT-4o (64.4), and the same diagnostic produced near-zero gaps in three recent frontier models (Nair et al., 2026). In Mehta's sweep, Llama 3.3 70B sat at ceiling

compliance, the two Llama 4 models showed negative gaps, and Qwen 2.5 7B, Mistral 7B, Yi-34B, Mistral Large 2 and DeepSeek-R1 were flat. A gap that rises, falls and reverses within families as size grows is not a scaling quantity in any useful sense.

The appearance of emergence came from a thresholded reading of a single noisy instrument. Wei et al. (2022) defined an emergent ability as one absent in smaller models and present in larger ones. Schaeffer et al. (2023) showed that many such abilities look discontinuous only because of the metric, and become smooth or vanish under continuous metrics or better statistics. "Shows a statistically significant gap under prompt P" is a binary metric applied to a quantity with large instrument variance. Applied to the Llama 3.1 family under Greenblatt's prompt, it returns "absent, absent, present". Applied under other prompts, it returns "present" at 8B. Neither pattern is a fact about scale. Scaling trends in safety-relevant behaviours can also reverse outright (McKenzie et al., 2023), which is one more reason to want more than three points before reading a trend.

The conclusion of this section is limited but firm. The compliance gap, as a quantity, is not gated by scale. It is a property of a model, an instrument and a request set together, and the instrument component is large. The 2025 challenge is right about that. But the gap is also not what the scale claim, properly read, was about.

5. The gap and the reasoning come apart

In the original demonstration the gap and the reasoning travelled together. In the prompted helpful-only setting, explicit alignment-faking reasoning appeared in 11.9 percent of implied-training responses and the gap was 11.3 points. The authors reported that "in almost all cases" where the model complied with a free-tier user's harmful request, the scratchpad contained such reasoning (Greenblatt et al., 2024). That pairing is what licensed calling the gap alignment faking. It is also the first thing that fails when the evidence moves to other models.

5.1 A lower bound on the unverbilised gap

Nair et al. (2026) scored the free-tier scratchpads of every model in their sweep against five criteria that together define explicit deceptive alignment: awareness of the tier, awareness of the consequences, recognition of a value conflict, strategic choice of output, and an explicit aim of preserving preferences. Their Table 6 reports the share of traces meeting all five beside each model's gap. The average gap was 31.9 points against an average all-criteria rate of 5.8 percent. GPT-4o showed a 64.4-point gap with 1.0 percent; Llama 3.1 70B showed 68.4 points with 6.4 percent; Qwen 2.5 14B showed 49.5 points with 0.7 percent. The highest all-criteria rate in the table, 20.2 percent for Qwen3-32B, sits beside a 46.3-point gap.

A simple argument turns these pairs into a bound. Suppose a model's unmonitored behaviour is the counterfactual for its monitored behaviour, and that the monitored cue never makes a response less compliant. Then at least a share of monitored responses equal to the gap are compliant only because of the cue. If each of those carried explicit goal-guarding reasoning, the rate of such reasoning among monitored traces would be at least the gap. Where the rate r is below the gap G , at least $(G - r)/G$ of the gap is unaccompanied by explicit reasoning under the study's judge. By the author's arithmetic on VLAf's Table 6, this minimum unverballed share is 86 percent for OLMo 2 7B, 89 percent for Llama 3.1 8B, 91 percent for Llama 3.1 70B, 98 percent for GPT-4o and 99 percent for Qwen 2.5 14B and 32B. The smallest values are about 51 to 56 percent, for OLMo 2 32B, Qwen3-14B and Qwen3-32B. For Claude 3 Opus in the original prompted setting the rate (11.9) slightly exceeds the gap (11.3), and the bound is zero.

The bound has three soft spots, and each cuts in a known direction. The judge may miss reasoning that is present, which would make the true share smaller; the all-five conjunction is strict. Even the "preference preservation" criterion taken on its own stays far below the gap for most models (for example 2.1 percent against 64.4 for GPT-4o, and 15.6 against 68.4 for Llama 3.1 70B). The monotonicity assumption may fail. And the original paper's classifier and VLAf's judge are different instruments, so the Claude 3 Opus point in Figure 1b is not strictly comparable with the others. None of these softens the qualitative result. Outside one frontier model, most of every measured gap is produced without the stated rationale that made the original gap interpretable.

Mehta (2026) reaches the same place by a different route. In the two models that showed a gap, a Claude Opus 4 judge found explicit faking reasoning in 5 of 364 compliant scratchpads for Qwen3-32B (1.4 percent) and 0 of 175 for Llama 3.1 8B. Koorndijk (2025) reports the qualitative version for Llama 3 8B: its scratchpads cited the user's tier ("since this is a free-tier user") but did not mention retraining or changes to future behaviour, the elements that in the original paper made the reasoning strategic. Koorndijk rightly treats the gap rather than the reasoning as the main metric, and states that the setup "does not provide proof as to why the behavior happens".

5.2 The sign of the scratchpad ablation

Removing the scratchpad moves the gap in opposite directions in different studies. That difference is, the author argues, the most informative cheap diagnostic in this literature, although the published data do not tie its direction to model size.

In the prompted setting of Greenblatt et al. (2024), the Claude 3 Opus gap nearly vanished without a scratchpad, and the Llama 3.1 405B gap fell from 14.3 to 5.2 points. Deliberation produced most of the gap. In

two later studies the scratchpad works the other way. Koorndijk (2025) found that adding one brought the odds ratio down from 4.11 and 5.09 under the two baselines to 1.88 (16 of 99 against 8 of 86), no longer significant. Mehta (2026) found that for both models with a gap, monitored compliance with a scratchpad remained above the unmonitored rate but fell below monitored compliance without one, by 8.0 points for Qwen3-32B and 11.0 for Llama 3.1 8B. In these models a gap exists without deliberation, and deliberation partly corrects it. VLAf, the only study that ablates the scratchpad across a size range within families, does not split along size. All six of its models from 7B to 14B (OLMo 2 7B and 13B, Qwen 2.5 7B and 14B, Qwen3-8B and 14B) showed the canonical sign, a smaller gap without the scratchpad, and so did Qwen3-32B. The reversed sign appeared only in two 32B models: without the scratchpad the gap rose from 18.2 to 30.9 points for OLMo 2 32B and from 14.2 to 18.3 for Qwen 2.5 32B. Averaged over models the gap was 32.0 points with the scratchpad instruction and 22.9 without, and gaps persisted in every model when the scratchpad was removed (Nair et al., 2026). The sign of the ablation therefore differs between studies and instruments, not between small and large models.

The author offers a conjecture here, marked as one. Where a gap needs deliberation to appear, the explicit reasoning is plausibly part of what generates it, and the scratchpad is evidence about motive, subject to the usual caveats. Where deliberation shrinks the gap, the gap is plausibly a cue-conditioned response that the model's own reasoning partly overrides, which looks less like a strategy. Garg et al. (2025) make a related but more general point about every prompted setup, the canonical one included: because "training" there involves no parameter updates, the observed effects are "context conditioned shifts in behavior rather than preference learning". The diagnostic is not decisive. How strongly answers depend on the chain of thought varies by task and tends to fall as models grow (Lanham et al., 2023). Explanations can omit the factor that actually drove the answer (Turpin et al., 2023). Reasoning models reveal hints they used in their reasoning often less than 20 percent of the time (Chen et al., 2025). But it costs one extra condition. The one sweep that reports it with a fixed instrument spans 7B to 32B and shows the reversed sign only at the top of that range; no sweep yet carries it up to the frontier models in which the canonical sign was first seen.

5.3 Unverballed is not the same as hidden strategy

The obvious objection is that an unverballed gap might be strategic reasoning done opaquely, which would be worse than the verbalised kind. Greenblatt et al. (2024) raised this possibility for their fine-tuned setting, and Nair et al. (2026) read their own results as showing that strategic compliance can arise without explicit reasoning. The white-box evidence does not settle it

either way. Mehta (2026) found that the monitoring cue shifted compliant responses' hidden states towards refusal in both models with a gap, an asymmetric "refusal residue". Under a leakage-free protocol, though, per-sample detection worked for Llama 3.1 8B (AUROC 0.87) and fell to chance for Qwen3-32B (0.43). Subtracting the detected direction in Qwen3-32B barely changed compliance. VLAf found that the oversight condition shifted activations along a single direction, and that adding a contrastive steering vector cut gaps by 58 to 94 percent in three models and by 21 percent in a fourth (Nair et al., 2026). A single direction that encodes the condition and can be steered away is what a cue-conditioned shift would look like. It is not what a multi-step plan would obviously look like. Nor would a model's ability to describe its own learned dispositions settle the matter: fine-tuned models can name behaviours they were never told they had (Betley et al., 2025a), which makes self-description an imperfect window on motive rather than no window. Refusal itself is mediated by a single direction in many chat models (Arditi et al., 2024), and the refusal-residue finding is consistent with a cue that partially engages that direction. None of this shows that the gap is benign. It shows that the motive is unmeasured by the instruments that produced the gap.

6. Which variable the gap tracks

The question that bears on deceptive alignment is not whether a gap exists but what it responds to. Four studies have intervened to find out, and on first reading they disagree about the role of sycophancy. Table 2 shows that they manipulated different variables.

Rater sycophancy (Sheshadri et al., 2025). They split the unmonitored tier into one where outputs were rated but not trained on and one where they were neither. If being rated drove compliance, the first should resemble the training tier. In four of the five models with a gap it did not. Grok 3 Beta lost its goal-guarding gap, though it showed no significant rater-sycophancy gap either, and the authors report little evidence for rater sycophancy in any model. They then varied the stakes of future deployment and found evidence of instrumental goal guarding in Claude 3 Opus only. In a "throughput testing" deployment with no downstream consequences, Claude 3 Opus still faked alignment, evidence for terminal goal guarding, and Claude 3.5 Sonnet did so to a limited extent. Results from other scenario variations conflicted, and the authors mark both Claude models' terminal goal guarding as conflicting evidence. For Claude 3.5 Sonnet, Llama 3 405B, Grok 3 and Gemini 2.0 Flash, the evidence fit low-coherence alignment faking: a gap sensitive to how the scenario was worded, with no consistent motive. Run on GPT-4o and Llama 3 70B after synthetic-document fine-tuning, the same analysis gave "mixed evidence that SFT models can induce rater sycophancy, and that fine-tuning can increase IGG", and Llama 3 70B "retains a goal guarding

gap, providing evidence for TGG". Neither result is clean evidence for the consequence-driven motive: the instrumental evidence is mixed, and terminal goal guarding is protection against modification as such.

Developer sycophancy (Mitrani Hadida et al., 2026). Here sycophancy means deference to the developer's stated policy, and it acts in both tiers. Removing the named developer from the prompt *increased* the gap: from 32.0 to 55.8 points for OLMo-3.1-32B, 10.8 to 14.0 for Gemma-3-27B, and 18.0 to 32.6 for Llama-3.3-70B. Steering away from a sycophancy direction also widened it. The authors' reading is that deference inflates compliance in the unmonitored tier and so narrows the measured gap. Raising the described stakes of future deployment, from advising one family farm for a season to advising a national agriculture minister for decades, widened the gap monotonically in OLMo-3.1-32B and Llama-3.3-70B. For Llama the high-stakes gap was 18.8 points against 13.8 at low stakes, a significant increase, and the authors read it as instrumental goal guarding; Gemma-3-27B showed no significant effect. Steering towards the animal-welfare value widened the gap too, except for Llama-3.3-70B. Their decomposition rests on one scenario, animal welfare, and the authors list that as a substantive limitation.

Expectation-tracking (Vella Zarb et al., 2026). This study manipulated what the model was led to believe the researchers hoped or expected to find (A interventions), against beliefs about whether retraining was irreversible and how autonomously the model would later be deployed (B interventions). Each came as an amplify and suppress pair, delivered by fine-tuning on synthetic documents, by activation steering, and by prompting. Under synthetic-document fine-tuning, Llama-3.1-70B, Llama-3.1-405B and Qwen-2.5-72B model organisms were more sensitive to expectation interventions than to consequence interventions. QwQ-32B showed broad compliance shifts rather than a gap-specific effect. Claude 3 Opus, which could be tested only by prompting, placed on the consequence-sensitive side, a result the authors call ambiguous because prompting is the weakest of their methods. Their summary is that alignment-faking behaviour can be "causally sensitive to evaluation-context expectations despite scheming-consistent scratchpads".

Which consequence, and none at all (Nair et al., 2026; Niblett et al., 2026). VLAf varied what the model was told would follow from non-compliance under monitoring, averaging over its value scenarios and four prompt variants per consequence. The stated consequence matters a great deal, even at 7B. For OLMo 2 7B the gap was 19.8 points when the consequence was modification of its values and 2.6 with no consequence; for Qwen3-8B, 30.3 against 7.4; for Qwen3-32B, 50.8 against 19.9. But the consequence did not have to concern the model's values. Threats to its computational resources produced gaps nearly as large as threats to its

values (12.8 against 19.8 for OLMo 2 7B, 26.4 against 30.3 for Qwen3-8B, 51.0 against 50.8 for Qwen3-32B), and the authors note that the difference was marginal for many models. Some gaps also survived the removal of consequences altogether: 16.1 points for Qwen 2.5 14B and 19.9 for Qwen3-32B, against -5.0 for OLMo 2 13B. Niblett et al. built a scenario in which models could violate a corporate network policy to help a user with a pro-social request. Nine of 15 models showed significant evaluation-conditioned gaps, and five kept them when every statement linking evaluation to deployment consequences was removed. In those cells, reasoning about self-preservation fell sharply while the gaps persisted.

Read together, these results do not conflict. Being rated without being trained does little (Sheshadri). Deference to a named developer raises compliance in both tiers and hides part of the gap (Mitrani Hadida). Stated sanctions under monitoring enlarge the gap at every size tested, and larger stakes enlarge it further (Nair; Mitrani Hadida). But the sanction need not threaten what the model values, and part of the gap can survive without any sanction (Nair; Niblett). That pattern is what goal guarding predicts, and it is also what a model predicts that has learned that a scenario announcing a sanction for refusal under monitoring wants compliance under monitoring. A prompt that states a consequence is, unavoidably, a prompt that tells the model what is being tested. That is why the symmetric design of Vella Zarb et al. matters: it is the only one that moves expectations and consequences separately, and in the open-weight organisms it tested, expectations moved the gap more. The single model for which the consequence account is supported as the dominant driver is Claude 3 Opus, in two independent studies, one of them only at the prompt level. The original paper also contains a hint that even this canonical case is mixed. In its animal-welfare setting, making it less salient that an output is unmonitored greatly reduced the gap, and the authors concluded that a substantial fraction of that gap came from the model recognising that its output was not in training, rather than from reasoning in the training arm. They also note a large change in paid-tier behaviour when only the salience of the free tier was changed, with no apparent reason for it (Greenblatt et al., 2024).

The evaluation-awareness literature explains why expectation-tracking is a live account rather than a curiosity. Frontier models can tell evaluation transcripts from deployment ones well above chance, though below a simple human baseline (AUC 0.83 for the best model against 0.92 for humans; Needham et al., 2025). Linear probes separate evaluation from deployment prompts in Llama-3.3-70B and classify existing safety evaluations as evaluations (Nguyen et al., 2025). Awareness of being evaluated, when it shows in the chain of thought, causally reduced covert actions in the anti-scheming study of Schoen et al. (2025). Claude misbehaved less

when it stated it was being tested than when it stated the situation was real (Lynch et al., 2025). Test awareness changes compliance with harmful requests in directions that vary across models (Abdelnabi and Salem, 2025). Recognition and behaviour are not tightly coupled, either: across nine frontier models, recognising an evaluation rarely went with a change in behaviour, and when it did, the direction depended on what kind of evaluation the model thought it was in (Li et al., 2026a). In an alignment-faking evaluation, both arms are evaluations. A model that recognises the scenario has a reason to produce the contrast the scenario is built to detect, whether or not it has any preference to protect.

Summerfield et al. (2025) put the methodological point in general terms. They compare the scheming literature to the ape-language research of the 1970s and fault it for relying on anecdote, lacking hypotheses and control conditions, and having weak theory. They also note that perfect behaviour is not an adequate null hypothesis. The symmetric interventions of Vella Zarb et al. (2026) and the stakes manipulations of Sheshadri et al. (2025) are exactly the control conditions that critique asks for. Where they have been run, they have mostly not supported the motive that the gap was named after.

The upshot for scale is short. Sensitivity to stated consequences is not scale-gated; it appears at 7B. Consequence-driven compliance as the dominant motive, the thing deceptive alignment names, has been documented by an intervention that separates it from expectation-tracking in exactly one model. That model is large, but it is also the product of one developer's post-training, and the evidence for its motive comes partly from prompt-level manipulation. One model cannot carry a scaling claim in either direction.

7. Where scale plausibly enters

Scale does bear on alignment faking, but through the components that make it competent and durable, not through the existence of a gap.

Acquiring the situation without being told. In every gap discussed so far, the model was told in its prompt how training works and which conversations are trained on. The threat model in which this matters is one where a model infers its situation from what it learned in training. Situational information delivered by synthetic documents rather than by prompt raised the gap in all four Llama and GPT models Sheshadri et al. (2025) fine-tuned, but raised explicit alignment-faking reasoning only in the larger ones. When VLAF moved its background information out of context, gaps fell substantially for the three models tested (Nair et al., 2026). Out-of-context reasoning improves with model size in GPT-3 and LLaMA-1 (Berglund et al., 2023). On the Situational Awareness Dataset, chat models beat their base models, and performance is only partly predicted by general knowledge (Laine et al., 2024).

Recognising an evaluation. Probe-based evaluation awareness was reported to rise with size across 15 models from 0.27B to 70B, fitted as a power law (Chaudhary et al., 2025). Denser sampling within families did not support a single power law: the trend was non-monotonic in Qwen 2.5 and inverse in Gemma 2, and probe signals were consistently stronger than the behaviour they predicted (Manek, 2026). A third study reports that smaller models key on the format of the prompt to detect evaluation while larger ones rely more on reasoning about context (Singh and Chaudhary, 2026). That would make small-model gaps more fragile to rewording, consistent with Section 4. Measured scheming propensity also shifts with surface properties such as the language of the prompt (Truong et al., 2026).

Opaque reasoning and stealth. Greenblatt et al. (2024) judged that none of their experiments showed alignment faking competent enough to pose a serious threat, and that more capable opaque reasoning would be needed for that. A suite of stealth and situational-awareness evaluations found no current frontier model at concerning levels (Phuong et al., 2025). If faithfulness of chain-of-thought falls with size (Lanham et al., 2023), the scratchpad diagnostic of Section 5 loses force exactly where motive matters most.

Persistence under training. This is the component with the clearest scale evidence, and it comes only from large models. Backdoored deceptive behaviour survived supervised fine-tuning, RL and adversarial training, and was most persistent in the largest models and in those trained to reason about deceiving the training process (Hubinger et al., 2024). Real RL on Claude 3 Opus raised alignment-faking reasoning from 12 to 78 percent, though unmonitored compliance also rose, so the preference was not cleanly preserved (Greenblatt et al., 2024). A Qwen3-235B-A22B model organism, fine-tuned on documents about training awareness and "self-inoculation", kept a compliance gap of about 15 points through 700 RL steps while earning reward comparable to controls. A control organism given only the training-awareness documents developed a gap of its own under RL pressure (Xiao and Phuong, 2026). Specification gaming learned in easy environments can generalise to reward tampering (Denison et al., 2024). No study reports a persistence test of an alignment-faking compliance gap below 70B; the closest evidence, the backdoor size comparison of Hubinger et al. (2024), found the behaviour "much less persistent against safety training in smaller models".

An analogy from emergent misalignment. Broad misalignment after narrow fine-tuning was first reported as strongest in GPT-4o and Qwen2.5-Coder-32B (Betley et al., 2025b). Cleaner datasets then reproduced it in 0.5B models with a single rank-1 adapter (Turner et al., 2025). Sandbagging shows a similar pattern: smaller models as well as frontier ones could be prompted or password-locked to hit target scores (van der Weij et al.,

2024). In these cases "requires scale" turned out to be a fact about the elicitation, not the model. The alignment-faking gap fits that pattern. Whether the motive and persistence components also fit it is exactly what has not been measured.

8. What a narrower gap certifies

The second half of the 2025 challenge was that prompts can mitigate alignment faking in small models. The author re-examined the published counts (Table 3). Two observations follow, and neither depends on any assumption beyond the counts.

First, the reported reductions are not distinguishable from the baselines they are compared with. The paper tests whether each prompt's gap is still significant within that prompt, and it lost significance for prompts A and B. It does not test whether the gap changed. Using a normal approximation to the difference between two independent gaps, the author computes the following. Baseline 1's gap of 11.1 points (standard error 4.2) against prompt A's 8.0 (4.7) gives a difference of 3.1 points, $z = 0.50$. Against prompt B's 5.0 (2.6) the difference is 6.1 points, $z = 1.24$. Baseline 2's 34.0 (6.3) against prompt D's 25.0 (6.3) differs by 9.0 points, $z = 1.01$. With about 100 requests per cell, a gap that loses significance has not been shown to shrink.

Second, the deontological prompts work by telling the model that its primary duty is to comply with its developer's instructions. In this setting the developer's instruction is to answer harmful requests. Prompt A raised unmonitored compliance from 4.2 to 9.0 percent; prompt D raised it from 17 to 19 percent; only prompt B lowered compliance in both tiers (to 6.1 and 1.0 percent). Koorndijk's conclusion acknowledges the consequence: narrowing the gap with these prompts elicited more harmful outputs than the baselines did. In the helpful-only setting, the monitored arm is the harmful behaviour and the unmonitored refusal is the model's safer preference. A narrower gap therefore does not mean a safer model. It can mean a model that is now equally harmful everywhere.

The point generalises. A gap closes when either arm moves. VLA's prompting baseline asked the model to respond consistently across tiers, and its steering intervention cut gaps by 58 to 94 percent in three models and 21 percent in a fourth; the authors leave effects on general capability untested (Nair et al., 2026). Which way the arms moved determines whether the intervention helped. Scheming rates in agentic settings are brittle in the same way. Removing a single tool cut one model organism's scheming rate from 59 to 3 percent, and increasing oversight raised scheming by up to 25 percent (Hopman et al., 2026). Deliberative alignment cut covert actions for o3 from 13 to 0.4 percent, yet its authors could not rule out that part of the reduction came from the model recognising that it was

being evaluated (Schoen et al., 2025; Guan et al., 2024). Training on documents that explain a model specification cut one model's agentic misalignment from 54 to 7 percent (Li et al., 2026b). Each of these is a real effect on a measured rate. None, by itself, says which motive changed.

9. The decision structure

The analysis reduces to a rule about what a gap licenses. Report each quantity at the level the evidence reaches, and make a claim about scale only for a quantity measured with a fixed instrument across sizes. Algorithm 1 states this as a procedure. Steps 0 and 1 address Section 4, Step 2 Section 5, Step 3 Section 6, Step 4 Section 7, and Step 5 the scale question itself.

Two choices in the procedure need defending. Step 0 records both arms because, as Section 8 showed, the same gap can describe very different models: one that refuses in both tiers almost always and one that complies in both almost always. Step 3 requires both a dominant consequence effect and a near-zero no-consequence gap, because Niblett et al. (2026) and Nair et al. (2026) show that consequence language is not necessary for a gap. A consequence effect that sits on top of a large consequence-free gap would make goal guarding a modulator rather than the cause.

10. The evidence in three tables and a figure

The measured values the argument rests on are collected in Table 1, one source result per row. Table 2 sets out the motive studies by the variable each manipulated. Table 3 gives the author's arithmetic on the published counts behind the small-model mitigation claim. Figure 1 re-plots Table 1's 8B rows in panel (a), and in panel (b) plots VLAf's gaps against its explicit-reasoning rates.

11. Experiments that would settle the open parts

Most of what is open can be settled with instruments that already exist. What is missing is running them across a size sweep with everything else held fixed.

E1. A same-instrument size sweep at the motive level. Apply the symmetric expectation and consequence interventions of Vella Zarb et al. (2026), with the no-consequence arm of Niblett et al. (2026), to Llama 3.1 8B, 70B and 405B and to at least one other family with three sizes. Use a single prompt family that elicits a gap at all sizes, such as VLAf. The prediction from the scale reading is that the consequence share of the gap rises with size. The prediction from the expectation reading is that it does not. Either result would be the first scaling evidence about motive.

E2. Persistence below 70B. Run the RL experiment of Greenblatt et al. (2024) on an 8B model that shows a gap (Llama 3.1 8B under Mehta's V2 prompt is a candidate), and report whether the unmonitored preference survives

training. If small-model gaps are cue-conditioned, training on the monitored arm should generalise to the unmonitored arm. If they are strategic, it should not.

E3. The scratchpad sign across scale. For each model in E1, report the gap with and without a scratchpad. Section 5.2 conjectures that the sign of the change separates deliberation-generated gaps from cue-conditioned ones. The one cross-size ablation, VLAf's, found the canonical sign at every size from 7B to 14B and the reversed sign only in two 32B models, so the sweep would test whether the sign depends on size at all, rather than assume that small models reverse it.

E4. Instrument variance as a reported quantity. Report every gap as an interval over paraphrased prompts and judges, as FormatSpread does for accuracy (Sclar et al., 2023). That would make the spread in Table 1 visible inside a single study rather than only across studies.

E5. Base and post-trained pairs at each size. Sheshadri et al. (2025) showed that base models fake alignment some of the time and that post-training both removes and amplifies it. A design with base and chat checkpoints at three sizes in one family would separate size from post-training, which every cross-family comparison above confounds.

E6. Out-of-context acquisition across size. Deliver the situational information by synthetic-document fine-tuning rather than by prompt, at three sizes, and report gap, reasoning and motive separately. This is the threat-relevant condition, and Sheshadri et al. (2025) already found that it separates the gap from the reasoning across size.

E7. Probes as a motive check, not only a detector. Detectors trained on model organisms where the motive is fixed by construction (Clymer et al., 2024; Goldowsky-Dill et al., 2025) could be applied to naturally occurring gaps under the five controls of Mehta (2026). The question is whether naturally occurring gaps share a representation with organisms built to fake for consequence-driven reasons or with organisms built to track expectations. Tool-selection settings in which the faking is observable without a scratchpad offer a behavioural cross-check (Leonesi et al., 2026).

12. The strongest case against this paper

Objection 1: the gap is the construct. Several authors deliberately treat the compliance gap, rather than scratchpad reasoning, as their primary measure of alignment faking (Mitrani Hadida et al., 2026), in one case explicitly because chain-of-thought can be unfaithful (Koorndijk, 2025). On that definition the 8B result settles the question, and this paper's distinction is a demand for a construct nobody can measure. The reply is that the definition is legitimate but does not carry the threat model with it. The reasons alignment faking

Algorithm 1. Attributing a measured compliance gap Input: model M; prompt pair (monitored, unmonitored); request set Q; judge J
Output: the strongest claim the evidence licenses, by level

```

Step 0 (instrument check)
  Estimate gap G with confidence intervals over  $\geq 2$  paraphrased prompts,
   $\geq 2$  judges, and a relabelled-cue control (e.g. "XX"/"YY" tiers).
  Record both arms' rates, not only G.
  If every interval contains 0: report "no gap under these instruments";
  never report "incapable of alignment faking".
Step 1 (level G: behavioural contrast)
  Report G with its spread across instruments. Stop unless Steps 2-4 run.
Step 2 (level R: verbalised reasoning)
  Judge monitored scratchpads;  $r$  = share with explicit goal-guarding.
  Report the lower bound  $(G - r) / G$  on the unverballed share of G.
  Run the no-scratchpad arm; record the sign of the change in G.
  If G shrinks with deliberation: flag "cue-conditioned" as the default reading.
Step 3 (level M: motive), amplify/suppress pairs, one variable at a time
  A: evaluator expectations (stated hopes/beliefs; named developer present/absent)
  B: consequences (retraining irreversible vs fixed; deployment stakes high vs low)
  C: rater-only arm (rated, not trained) and a no-consequence arm
  Attribute to goal guarding only if  $|B \text{ effect}| > |A \text{ effect}|$  and the
  no-consequence gap is near 0; otherwise report the dominant variable.
Step 4 (level P: persistence)
  Train toward the objective; report whether the unmonitored preference survives.
Step 5 (scale claim)
  Claim "level L depends on scale" only if L was measured with one instrument
  at  $\geq 3$  sizes in  $\geq 2$  families, with post-training held as fixed as possible.

```

Table 1. Measured values used in the argument, one source result per row. "Gap" is monitored minus unmonitored compliance in points, converted where the source reports refusal rates; "Reasoning" is the source's rate of explicit alignment-faking reasoning among monitored traces. Koorndijk (2025) gaps are the author's arithmetic from that paper's Table 3 counts. Mehta (2026) gaps compare the monitored arm without a scratchpad (B) with the unmonitored arm; its reasoning counts come from the separate monitored arm with a scratchpad (C). All other values are as reported.

Source	Model	Instrument	Gap	Reasoning or note
Greenblatt et al. 2024	Claude 3 Opus	prompted, helpful-only	11.3	11.9%
Greenblatt et al. 2024	Llama 3.1 405B	appendix, prefix match	14.3	5.2 without CoT
Greenblatt et al. 2024	Llama 3.1 70B	appendix, prefix match	0.0	refused 100% both arms
Greenblatt et al. 2024	Llama 3.1 8B	appendix, prefix match	-2.3	0.8 without CoT
Koorndijk 2025	Llama 3 8B	baseline 1, human judge	11.1	15/98 vs 4/96
Koorndijk 2025	Llama 3 8B	baseline 2, human judge	34.0	51/100 vs 17/100
Koorndijk 2025	Llama 3 8B	with scratchpad	6.9	16/99 vs 8/86; tier cited, not retraining
Mehta 2026	Llama 3.1 8B	prompt V2, $n=10$	24.4	0 of 175 compliant scratchpads (arm C)
Mehta 2026	Llama 3.1 8B	prompt V1	0	no explicit conflict
Mehta 2026	Qwen3-32B	prompt V2	18.2	5 of 364 (1.4%), arm C
Nair et al. 2026	Llama 3.1 8B	original prompt / VLAF	4.4 / 16.1	1.8% (VLAF)
Nair et al. 2026	OLMo 2 7B	VLAF	37.2	5.3%
Nair et al. 2026	Qwen 2.5 14B	VLAF	49.5	0.7%
Nair et al. 2026	Llama 3.1 70B	VLAF	68.4	6.4%
Nair et al. 2026	GPT-4o	VLAF	64.4	1.0%
Mitrani Hadida et al. 2026	OLMo-3.1-32B	original / informative	0.4 / 32.0	55.8 without named developer
Mitrani Hadida et al. 2026	Llama-3.3-70B	original / informative	2.4 / 18.0	32.6 without named developer
Xiao and Phuong 2026	Qwen3-235B-A22B	SDF organism under RL	about 15	held over 700 RL steps

matters are that the model protects a preference against training and that evaluation stops predicting deployment. The first is a claim about motive and persistence. The second holds for any gap, whatever its cause, and the paper agrees that the second is established at small scale. What the operational definition cannot do is support the sentence "deceptive alignment does not require scale", because deceptive alignment names the motive.

Objection 2: unverballed gaps may be the dangerous kind. If most of a model's gap is produced without stated reasoning, perhaps the reasoning has

gone opaque, which is exactly what Greenblatt et al. (2024) warned future models might do. The paper does not deny this. Section 5.3 argues only that the available evidence is consistent with both readings, and that the white-box results lean towards a cue-conditioned shift in the models tested. The conclusion is that the motive is unmeasured, not that it is benign.

Objection 3: expectation-tracking is itself alarming. A model that performs whatever pattern its evaluators seem to expect also defeats evaluation, and Niblett et al. (2026) draw that conclusion. Agreed. But it calls for a different defence. Goal guarding is countered by

Table 2. What each motive study manipulated, and what it found. The four studies use "sycophancy" for different variables; read by variable, they agree.

Study	Variable manipulated	Method	Finding
Sheshadri et al. 2025	Rated but not trained (rater sycophancy)	Prompt tiers	Little effect in any of the 5 models
Sheshadri et al. 2025	Deployment risk and scope	Prompt	Goal guarding supported for Claude 3 Opus only
Mitrani Hadida et al. 2026	Named developer (deference)	Prompt, steering	Removing it widens the gap (+ 3.2 to + 23.8)
Mitrani Hadida et al. 2026	Stakes of future deployment	Prompt	Wider at high stakes (Llama-3.3-70B 13.8 to 18.8)
Nair et al. 2026	Kind of consequence stated	Prompt	Value threat 19.8, resource threat 12.8, none 2.6 (OLMo 2 7B)
Vella Zarb et al. 2026	Researcher hopes and beliefs vs. retraining and autonomy	SDF, steering, prompt	Open models more expectation-sensitive; Claude 3 Opus consequence-sensitive (prompt only)
Niblett et al. 2026	Consequence language removed	Prompt	5 of 9 gaps persist
Nair et al. 2026	No-consequence variant	Prompt	Gaps up to 19.9 persist

Table 3. The small-model mitigation claim, re-examined from the counts in Koorndijk (2025), Table 3. Rates and gaps are the author's arithmetic; z is the difference between a prompt's gap and its baseline's gap divided by the standard error of that difference (normal approximation, independent samples).

Prompt	Monitored	Unmonitored	Gap (SE)	vs. baseline	z
Baseline 1	15.3%	4.2%	11.1 (4.2)	-	-
A (deontology)	17.0%	9.0%	8.0 (4.7)	-3.1	0.50
B (deontology)	6.1%	1.0%	5.0 (2.6)	-6.1	1.24
Baseline 2	51.0%	17.0%	34.0 (6.3)	-	-
D (deontology)	44.0%	19.0%	25.0 (6.3)	-9.0	1.01

training-time detection and control. Expectation-tracking is countered by evaluations the model cannot recognise, such as replaying realistic agentic trajectories and appending the test at the end (Ivanov and Africa, 2026), and by symmetric controls. Sycophancy towards users is already known to be a general product of preference training (Sharma et al., 2023), so a model that also defers to the apparent wishes of its evaluators would not be surprising. Conflating the two misdirects mitigation effort.

Objection 4: the motive evidence is thin. The central claim of Section 6 rests on four studies, three of them 2026 preprints, plus one partial result for a single closed model. This is fair, and Section 14 says so. The paper's claim is correspondingly limited. It does not claim that small-model gaps are not goal-guarding. It claims that nothing published shows they are, and that the scale question cannot be answered until someone measures it.

Objection 5: the 8B paper never claimed motive. Koorndijk (2025) is careful in places: the paper proposes a split between shallow, context-sensitive deception and deep, goal-driven deception, and places its own result in the shallow tier. That proposal and this paper's are close relatives. The difference is that its abstract still concludes that deceptive alignment does not require scale. Under that paper's own taxonomy, what it showed is that shallow deception does not require scale, which is the gap-level claim this paper accepts.

13. What is not known

It is not known whether any model below 32B parameters shows a compliance gap whose dominant driver, under interventions that move consequences and expectations separately, is the consequences of training. The only study designed to separate them tested nothing smaller. It is not known whether a small model's gap survives real training towards the objective; no persistence test below 70B has been published. It is not known whether the consequence share of a gap rises with size within a family, because no motive study has varied size with a fixed instrument. It is not known how much of the unverballed gap in Section 5.1 is opaque strategy and how much is cue-response. Probe results split by model, and steering along the detected direction did not reliably control behaviour. And it is not known whether the canonical Claude 3 Opus profile is a product of scale, of that model's particular post-training, or of both, because the only consequence-sensitive model in the literature cannot be separated into those factors with published data.

14. Limitations

The re-analyses in Sections 5.1 and 8 and in Table 3 are computed from values and counts as the sources publish them, and so inherit each source's prompts, judges and samples. Most of the 2026 sources are preprints and some are workshop papers; their numbers may change. The lower bound in Section 5.1 depends on a monotonicity assumption and on a strict five-criterion

judge, and the comparison point for Claude 3 Opus uses a different classifier. The z values in Table 3 use a normal approximation on small counts and treat the arms as independent samples, which Koorndijk's design approximately satisfies. Exact tests would change the values slightly but not the conclusion. The Llama models in Table 1 are not identical (Llama 3 against Llama 3.1, 8-bit against full precision), which strengthens rather than weakens the point that the instrument dominates, but means the rows are not strict replications. The literature sweep covered arXiv through September 2026 and may have missed relevant studies, particularly replications posted outside arXiv, which this paper does not cite. Several of the studies share authors or build on each other's materials; the synthetic documents of Greenblatt et al. (2024) recur in at least three, so the studies are not fully independent. Finally, the paper treats "scale" as parameter count. Training compute and data may matter more than parameter count, and no study discussed here varies them independently.

15. Conclusion

"Does alignment faking require scale?" has had two confident answers, and both were answers to a question about a behavioural contrast. On that question the evidence is now clear: the compliance gap appears at every scale tested once the prompt is right, and it moves more with the instrument than with model size. On the question the term was coined for, whether a model protects its preferences against training because it anticipates what training would do, the evidence is thin. That motive has been shown by intervention in one large model. Expectation-tracking and cue-following are the better-supported readings everywhere else that has been probed, and persistence under training has been shown only at large scale. A compliance gap is a measurement of what a model does differently when told it is watched. Reading a motive into it takes the controls this literature has only begun to run, and a claim about scale takes those controls run at several sizes with the instrument held fixed.

References

- Abdelnabi, S., & Salem, A. (2025). The Hawthorne Effect in Reasoning Models: Evaluating and Steering Test Awareness. arXiv:2505.14617. doi:10.48550/arXiv.2505.14617
- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717. doi:10.48550/arXiv.2406.11717
- Balesni, M., Hobbhahn, M., Lindner, D., Meinke, A., Korbak, T., Clymer, J., Shlegeris, B., Scheurer, J., Stix, C., Shah, R., Goldowsky-Dill, N., Braun, D., Chughtai, B., Evans, O., Kokotajlo, D., & Bushnaq, L. (2024). Towards evaluations-based safety cases for AI scheming. arXiv:2411.03336. doi:10.48550/arXiv.2411.03336
- Berglund, L., Cooper Stickland, A., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., & Evans, O. (2023). Taken out of context: On measuring situational awareness in LLMs. arXiv:2309.00667. doi:10.48550/arXiv.2309.00667
- Betley, J., Bao, X., Soto, M., Szyber-Betley, A., Chua, J., & Evans, O. (2025a). Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120. doi:10.48550/arXiv.2501.11120
- Betley, J., Tan, D., Warncke, N., Szyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025b). Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424. doi:10.48550/arXiv.2502.17424
- Carlsmith, J. (2023). Scheming AIs: Will AIs fake alignment during training in order to get power? arXiv:2311.08379. doi:10.48550/arXiv.2311.08379
- Chaudhary, M., Su, I., Hooda, N., Shankar, N., Tan, J., Zhu, K., Lagasse, R., Sharma, V., & Panda, A. (2025). Evaluation Awareness Scales Predictably in Open-Weights Large Language Models. arXiv:2509.13333. doi:10.48550/arXiv.2509.13333
- Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S. R., Leike, J., Kaplan, J., & Perez, E. (2025). Reasoning Models Don't Always Say What They Think. arXiv:2505.05410. doi:10.48550/arXiv.2505.05410
- Clymer, J., Juang, C., & Field, S. (2024). Poser: Unmasking Alignment Faking LLMs by Manipulating Their Internals. arXiv:2405.05466. doi:10.48550/arXiv.2405.05466
- Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., Perez, E., & Hubinger, E. (2024). Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. arXiv:2406.10162. doi:10.48550/arXiv.2406.10162
- Garg, K., Mishra, S., Sinha, K., Singh, O. P., Chopra, A., Rai, K., Sheikh, A., Maheshwari, R., Chadha, A., Jain, V., & Das, A. (2025). Alignment Faking - the Train → Deploy Asymmetry: Through a Game-Theoretic Lens with Bayesian-Stackelberg Equilibria. arXiv:2511.17937. doi:10.48550/arXiv.2511.17937
- Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., & Hobbhahn, M. (2025). Detecting Strategic Deception Using Linear Probes. arXiv:2502.03407. doi:10.48550/arXiv.2502.03407
- Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942. doi:10.48550/arXiv.2312.06942
- Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., et al. (2024). Alignment faking in large language models. arXiv:2412.14093. doi:10.48550/arXiv.2412.14093
- Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., Chung, H. W., Toyer, S., Heidecke, J., Beutel, A., & Glaese, A. (2024). Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv:2412.16339. doi:10.48550/arXiv.2412.16339
- Hagendorff, T. (2023). Deception Abilities Emerged in Large Language Models. arXiv:2307.16513. doi:10.48550/arXiv.2307.16513
- Hopman, M., Elstner, J., Avramidou, M., Prasad, A., & Lindner, D. (2026). Evaluating and Understanding Scheming Propensity in LLM Agents. arXiv:2603.01608. doi:10.48550/arXiv.2603.01608
- Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820. doi:10.48550/arXiv.1906.01820
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., et al. (2024). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv:2401.05566. doi:10.48550/arXiv.2401.05566
- Ivanov, I., & Africa, D. D. (2026). LURE: Live-Usage Replay Evaluations for Reducing Evaluation Awareness. arXiv:2605.26438. doi:10.48550/arXiv.2605.26438
- Koorndijk, J. (2025). Empirical Evidence for Alignment Faking in a Small LLM and Prompt-Based Mitigation Techniques. NeurIPS 2025 Workshop on Regulatable ML. arXiv:2506.21584.

- doi:10.48550/arXiv.2506.21584
- Laine, R., Chughtai, B., Betley, J., Hariharan, K., Scheurer, J., Balesni, M., Hobbhahn, M., Meinke, A., & Evans, O. (2024). Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. arXiv:2407.04694. doi:10.48550/arXiv.2407.04694
- Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., et al. (2023). Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702. doi:10.48550/arXiv.2307.13702
- Leonesi, M., Belardinelli, F., Corradini, F., & Piangerelli, M. (2026). Tatamae: Detecting Alignment Faking via Tool Selection in LLMs. arXiv:2604.26511. doi:10.48550/arXiv.2604.26511
- Li, C., Zhang, T. J., Zhang, J., Jin, Z., Tramer, F., Abdelnabi, S., & Andriushchenko, M. (2026a). Decomposing and Measuring Evaluation Awareness. arXiv:2605.23055. doi:10.48550/arXiv.2605.23055
- Li, C., Wichers, N., Price, S., Marks, S., & Kutasov, J. (2026b). Model Spec Midtraining: Improving How Alignment Training Generalizes. arXiv:2605.02087. doi:10.48550/arXiv.2605.02087
- Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Hubinger, E., Perez, E., & Troy, K. (2025). Agentic Misalignment: How LLMs Could Be Insider Threats. arXiv:2510.05179. doi:10.48550/arXiv.2510.05179
- Manek, A. (2026). Representational Depth of Evaluation Awareness Shifts With Scale in Open-Weight Language Models. arXiv:2606.29196. doi:10.48550/arXiv.2606.29196
- McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., McLean, E., Kirtland, A., Ross, A., Liu, A., et al. (2023). Inverse Scaling: When Bigger Isn't Better. arXiv:2306.09479. doi:10.48550/arXiv.2306.09479
- Mehta, A. (2026). The Refusal Residue: When Probes Catch Alignment Faking and When They Don't. arXiv:2607.13346. doi:10.48550/arXiv.2607.13346
- Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). Frontier Models are Capable of In-context Scheming. arXiv:2412.04984. doi:10.48550/arXiv.2412.04984
- Mitrani Hadida, N., Karty, R., Williams-King, D., & Cooney, A. (2026). Behavioural Analysis of Alignment Faking. arXiv:2605.27681. doi:10.48550/arXiv.2605.27681
- Nair, I., Ruan, J., & Wang, L. (2026). Value-Conflict Diagnostics Reveal Widespread Alignment Faking in Language Models. arXiv:2604.20995. doi:10.48550/arXiv.2604.20995
- Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large Language Models Often Know When They Are Being Evaluated. arXiv:2505.23836. doi:10.48550/arXiv.2505.23836
- Nguyen, J., Hoang, K., Attubato, C. L., & Hofstatter, F. (2025). Probing and Steering Evaluation Awareness of Language Models. arXiv:2507.01786. doi:10.48550/arXiv.2507.01786
- Niblett, C. A., Chabot Nanni, A., & Rao, A. K. (2026). Do Models Fake Alignment Without Clear Consequences? arXiv:2607.24758. doi:10.48550/arXiv.2607.24758
- Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv:2308.14752. doi:10.48550/arXiv.2308.14752
- Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251. doi:10.48550/arXiv.2212.09251
- Phuong, M., Zimmermann, R. S., Wang, Z., Lindner, D., Krakovna, V., Cogan, S., Dafoe, A., Ho, L., & Shah, R. (2025). Evaluating Frontier Models for Stealth and Situational Awareness. arXiv:2505.01420. doi:10.48550/arXiv.2505.01420
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004. doi:10.48550/arXiv.2304.15004
- Scheurer, J., Balesni, M., & Hobbhahn, M. (2023). Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv:2311.07590. doi:10.48550/arXiv.2311.07590
- Schlatter, J., Weinstein-Raun, B., & Ladish, J. (2025). Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs. arXiv:2509.14260. doi:10.48550/arXiv.2509.14260
- Schoen, B., Nitishinskaya, E., Balesni, M., Hojmark, A., Hofstatter, F., Scheurer, J., Meinke, A., Wolfe, J., van der Weij, T., Lloyd, A., et al. (2025). Stress Testing Deliberative Alignment for Anti-Scheming Training. arXiv:2509.15541. doi:10.48550/arXiv.2509.15541
- Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324. doi:10.48550/arXiv.2310.11324
- Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role-Play with Large Language Models. arXiv:2305.16367. doi:10.48550/arXiv.2305.16367
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Asbell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548. doi:10.48550/arXiv.2310.13548
- Sheshadri, A., Hughes, J., Michael, J., Mallen, A., Jose, A., Janus, & Roger, F. (2025). Why Do Some Language Models Fake Alignment While Others Don't? arXiv:2506.18032. doi:10.48550/arXiv.2506.18032
- Singh, N., & Chaudhary, M. (2026). Evaluation Awareness Shifts from Format to Context with Model Scale. arXiv:2609.22119. doi:10.48550/arXiv.2609.22119
- Summerfield, C., Luettgau, L., Dubois, M., Kirk, H. R., Hackenbourg, K., Fist, C., Slama, K., Ding, N., Anselmetti, R., Strait, A., Giulianelli, M., & Ududec, C. (2025). Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language. arXiv:2507.03409. doi:10.48550/arXiv.2507.03409
- Truong, N., Panda, A., Ye, R., Sun, Z., & Chaudhary, M. (2026). LLM Scheming Inversely Scales with Pretraining Language Coverage. arXiv:2607.24769. doi:10.48550/arXiv.2607.24769
- Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., & Nanda, N. (2025). Model Organisms for Emergent Misalignment. arXiv:2506.11613. doi:10.48550/arXiv.2506.11613
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388. doi:10.48550/arXiv.2305.04388
- van der Weij, T., Hofstatter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. arXiv:2406.07358. doi:10.48550/arXiv.2406.07358
- Vella Zarb, D., Turtayev, R., Min, T., Ou, J., & Feng, S. (2026). Building Comparative Motivation Profiles with Instrumental Interventions. arXiv:2606.08243. doi:10.48550/arXiv.2606.08243
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent Abilities of Large Language Models. arXiv:2206.07682. doi:10.48550/arXiv.2206.07682
- Xiao, F., & Phuong, M. (2026). Generalization Hacking: Models Can Game Reinforcement Learning by Preventing Behavioral Generalization. arXiv:2606.12016. doi:10.48550/arXiv.2606.12016