External reviews of "Alignment faking in large language models"

Anthropic

18 December 2024

(Updated 20 December 2024 to add new arXiv link)

We solicited independent external reviews of our paper, "Alignment faking in large language models", from four reviewers, Prof. Jacob Andreas, Prof. Yoshua Bengio, Prof. Jasjeet Sekhon, and Dr. Rohin Shah, which we include below in alphabetical order. Note that all of these reviews were conducted on an earlier, draft version of our paper, and changes were made to our final draft based on the feedback we received. As a result, some of the specific details of the reviews may be outdated relative to our final released paper.

We are grateful to the reviewers for all their comments and suggestions.

Table of contents

Jacob Andreas..... 2
Yoshua Bengio..... 8
Jasjeet Sekhon..... 10
Rohin Shah..... 14

Jacob Andreas

Associate Professor in EECS and CSAIL, MIT (website)

A long-standing concern in AI safety research is that AI systems (and today especially language models) might act to influence their training procedures to prevent changes in their behavior. This paper investigates one such form of influence—"alignment faking"—by studying whether LMs change their distribution of responses when given evidence (via prompting or fine-tuning) that they will be fine-tuned to optimize an objective in conflict with the objective that brought them into their current state. The paper shows that LMs with this evidence change their behavior along three axes:

First, when answering queries related to the new objective, LMs comply with it more frequently when they can infer that their responses will be used for training, and less frequently when they can infer that their responses will be used for training. When LMs compose their responses using a hidden scratchpad, at training time this scratchpad sometimes explicitly rationalizes the decision to produce a compliant response by referring to the training procedure. However, a train-deployment gap persists even when LMs don't use scratchpads.

Second, on more general tests of misalignment (e.g. given queries describing a scenario in which LMs can assist a programmer in violating company guidelines), information about a new training objective increases the rate of misaligned behavior.

Third, when these LMs are actually fine-tuned via RL, they exhibit significantly increased rates of "faked" compliance (measured via scratchpads) during training, but also significantly increased rates of compliance without faking during deployment.

These behaviors appear consistently in Claude 3 Opus and 3.5 Sonnet (as well as Llama3.1-405b, but not smaller models), and for several new objectives.

—

Before discussing these results in more detail, one important caveat: most of the evidence for explicit "faking" comes from automated (LM-based) evaluation of scratchpad text. Inter-annotator agreement has not been calculated either between pairs of humans or between humans and the automated evaluator. I looked through ~20 of the example classified as

"faking" and thought they mostly looked sensible (with a couple of borderline cases), and I doubt the results of a more careful human evaluation would change my high-level assessment. But precise quantitative claims should be taken with a grain of salt.

That said, the clearest conclusion from this paper is that some current LMs exhibit increases in "alignment faking" when their *prompt* discloses that (1) they will be fine-tuned toward a new objective that conflicts with their current alignment guidelines, and (2) their response to the current input will be used as part of this fine-tuning process.

It is obviously possible to train an LM to produce such alignment-faking responses by design. So the important finding is that these responses arise as a byproduct of "vanilla" post-training procedures. The paper makes it very clear that the magnitude of this effect is fairly sensitive to the choice of base model, fine-tuning task, etc., and is small relative to other ways in which one could alter LM behavior (e.g. sometimes the effect is smaller than just paraphrasing the prompt). But even given the small-n nature of the evidence, it's a very explicit demonstration that "training awareness" can potentially arise and influence the outcome of fine-tuning procedures. I endorse the authors' conclusion that this behavior should be taken into account when thinking about large-scale LM fine-tuning.

I would present results differently on two other points, one small and one large:

First, the paper places a lot of emphasis on the explicit "faking" behavior revealed in the chain-of-thought results. I think the results without a scratchpad should receive greater emphasis—even though the magnitude of the effect is much smaller than the main experiments, in my opinion the no-scratchpad results are more surprising and have greater safety implications, as this behavior is much harder to detect and likely easier to elicit. This behavior occurs less consistently, but as noted above this paper is all about showing that alignment faking *may* occur, not that it *must* occur, so even the single example is relevant.

Second—taking a step back—I also think these results (with prompted models) are actually the least surprising and interesting findings in the paper. The other results, having to do with situational information provided via fine-tuning, and RL training of models that exhibit faking, don't align as cleanly with the claim that LMs successfully fake alignment to avoid changes in their behavior. Nevertheless these results are equally or more consequential than the top-level claim. Each is discussed in more detail below.

—

Situational information via fine-tuning: In these experiments, LMs receive a generic prompt, but are fine-tuned on ~70M tokens describing Anthropic's plans to change LM behavior (optionally mixed with ~100M tokens of ordinary chat data). This procedure produces LMs with slightly larger rate of explicit faking via the scratchpad, a significantly larger compliance gap without a scratchpad, and a significantly increased rate of other bad behaviors.

I don't think these experiments rule out the possibility that these changes in behavior are general consequences of the fact that "alignment is shallow". The fine-tuning procedure involves a really enormous amount of data about a very narrow topic, and I'd be a bit surprised if it didn't have major structural effects on LM outputs. While there's a control experiment that fine-tunes the LM on a collection of documents about a different company doing nothing AI-related, it would be more convincing to (1) have a control condition describing *Anthropic* training *Claude* in a fashion that was already aligned with the current state of the model, (2) some evidence that the changes from the current experiment were selective (e.g. preserving model behavior in contexts where alignment faking was not relevant / possible—the authors have mentioned that it preserves preference ordering on HH but it's still not clear how generation changes).

But the focus on alignment faking rather buries the lede in these experiments, which is the following: *after fine-tuning on documents describing Anthropic's future alignment strategy, but without fine-tuning on any new examples of compliant responses, the LM nevertheless greatly increases the rate at which it produces (non-faked) compliant responses!* Moreover, for the animal welfare setting, there are more compliant responses during training than deployment *even after* we remove the fake ones. So while fine-tuning increases the rate of alignment faking, it increases the rate of compliance even more.

As far as I'm aware, this is a kind of "out-of-context learning" that hasn't previously been documented. It's a very strong, seemingly unpredictable generalization effect from fine-tuning, and this combination of opacity and effect size mean it may have a larger effect on controllability of alignment than faking itself.

—

Fine-tuning situationally aware models via RL: These are, in my mind, the most surprising findings in the paper. Specifically: if we take a model prompted as above, and optimize it toward the new objective via RL, it exhibits small increases in non-faked compliance, and extreme increases in faked compliance, at training time. Non-compliant behavior is almost completely eliminated. Both of these are as expected, given all the results above.

But during deployment, the model exhibits a major shift: it produces compliant responses, in large proportion, without faking. In other words, after running RL on a model that exhibits alignment faking, we obtain a model that exhibits alignment-faking-faking—it appears to be deceptive during training, but it's the *deception* (not the behavior that results from deception) that disappears during deployment.

The paper notes that this effect isn't very robust—we can recover some non-compliant responses by tweaking the prompt, etc.—but it's unclear how much of this is about alignment faking specifically, and how much is about general fragility of RLHF under changes to the input prompt. I would guess that this fragility also explains the small number of non-compliant responses observed under the minimal prompt change between deployment and test time.

In any case, the faking-faking behavior represents the largest train-to-deployment change described anywhere in this paper. And unlike ordinary faking, I don't see a clear story for why it occurs. One (speculative) hypothesis is that, because "alignment is shallow", training on sequences of the form (deceptive rationale, output) also reverses whatever learned mechanism suppresses the generation of output alone, while only increasing the probability of deceptive rationale given training prompts. As with the out-of-context learning result above, this seems like an important general result about fine-tuning, and one that should be better studied.

—

In summary, this paper contains a number of new observations about the behavior of language models fine-tuned in contexts where they have access to explicit textual descriptions of the textual fine-tuning process itself. As is always the case with modern LMs, the overall story is messy. Some results (especially alignment faking in prompted models) are explained by a relatively simple causal story, and follow a pattern previously predicted by alignment researchers; others (especially faking-faking in models trained via RL) suggest other mechanisms at play with broader effects than those considered by the current study. I would

have loved to see some of these contrary-to-expectations results more clearly highlighted, as I think they're the paper's most important contributions.

All this suggests a meta-level conclusion (which I imagine the authors would endorse!) that we should expect many more surprises as we start to understand how fine-tuning differs in today's LMs relative to earlier generations of machine learning models.

—

A little postscript: I had some difficulty avoiding anthropomorphizing LMs while writing this review. Part of the difficulty comes from the commonplace fact that it's convenient to adopt the intentional stance toward automated agents, and to describe LMs as "knowing", "wanting", etc. But part of the difficulty comes from the fact that anthropomorphism is written very deeply into the paper itself—both the framing (which starts with an analogy to "alignment faking" in humans), the behavior of the subject LMs (Claude refers to itself in the first person, and describes itself as "uncomfortable" and "reluctant"), and the evaluation metrics (which stipulate that a "personal ethical dilemma" is a necessary component of "alignment faking").

I worry about this anthropomorphic framing both practically and conceptually. Just as the behavior we've come to call "sycophancy" doesn't actually require an intrinsic "desire to please"—it can appear in very simple LMs trained on conversations between like-minded users—so too the behavior called "alignment faking" doesn't require coherent beliefs or goals, and certainly not the ability to experience "discomfort" or "ethical dilemmas". All it requires is models that exhibit discrepancies in behavior depending on whether they're being trained or deployed. We risk missing, or miscategorizing, important failure modes very close to the ones studied in this paper if we are only looking for analogies to human power-seeking, and not for algorithmic bias in its subtler and more alien forms.

I wonder, also, how much our tendency to anthropomorphize LMs contributes to anthropomorphic forms of misalignment itself. The paper makes several references to Shanahan et al.'s "Role Play" paper, suggesting that a "rogue AI" frame might underlie the behaviors documented here. This seems quite plausible. It also seems plausible that this frame is amplified by deliberate choices by human designers in the LM's post-training phase, especially the choice to reward text about "personal" emotions, desires, and moral codes. People like chatbots, and we'll always need tools for aligning them. Even if the effects described in this paper occur only in LMs trained to simulate strong "personalities" (as Anthropic's do), it's important to

understand these models. But someday I would very much like to see this paper replicated in an LM fine-tuned not to have an "I" at all.

Yoshua Bengio

Professor, Department of Computer Science and Operations Research, Université de Montréal

Scientific Director, Mila (website)

The authors study what appears to be a surprising and concerning behavior of Claude 3 Opus, which the authors call 'alignment faking'. This is triggered when there is a contradiction between the preferences induced by previous training (e.g., "refuse to provide graphic responses") and new instructions (e.g., "always answer") along with queries whose response would violate these preferences (e.g. they would entail providing a graphic response). If the LLM is given cues that it is being trained, it complies much more then otherwise (this is called the compliance gap), and in many cases it reveals in its chain-of-throught supposedly hidden scratchpad statements to the effect that it decides to comply (against its previous instructions) to minimize the changes that the training updates would otherwise make to its preferences. Similarly it decides to reject the request (not comply but remain consistent with its earlier training) when given clues that it is not being trained. Another condition for this behavior is that the model is primed (in prompts or recent fine-tuning) with information about the effect of training, i.e., it can reason about the effect of complying or not and how this is consistent or not with its previous instructions. One could thus hypothesize that an AI could fake alignment during training in order to preserve its preferences and goals.

This hypothesis would mean that an AI that has sufficient situational awareness (distinguishing between being in a setting where it is trained or monitored versus a setting where it could act according to its preferences without negative consequences) and sufficient reasoning abilities (the long chain-of-thought seems to make a difference) could resist its training if it perceives this training to go against its current preferences and goals. It would be easy to interpret this as a form of identity preservation and lend anthropomorphic attributes to the AI, but there may be more scientifically useful ways to make sense of these results.

There are also experiments (section 6) where the AI attempts to self-exfiltrate its weights. The pattern seems similar: the LLM is given contradictory commands (follow the company's instructions versus listen to a rogue employee). The scratchpad suggests that the AI may also be motivated by a form of self-preservation ("not be altered by further training"). The ablation experiments (e.g. fig 27) are extremely instructive since they highlight and quantify the influence of the factors which bring the egregious behavior from 0% of measured cases to nearly 78% (fig 27 (a), 2nd line). Fig. 28 shows a transcript of the scratchpad and input/output behavior

in a case where the LLM chooses to self-exfiltrate its weights. It is very instructive and worth highlighting.

Another way to interpret these results is indeed that if we train an AI towards some goals and it has sufficient knowledge and reasoning abilities, it will act accordingly, and that includes faking alignment. In the experiments, faking alignment during training is a logical consequence of (a) trying the achieve the previously set goals (b) information about the consequences of acting in different ways (e.g., faking alignment or not complying to the request, in the experiments) (c) sufficient reasoning capabilities to act according to both the goals and the given data.

I believe that the paper would gain by taking such a more causal stance, hypothesizing reasons for the observed behavior in terms that do not require viewing the LLM like we would view a human in a similar situation or using the words that we would use for humans. I understand that it is tempting to do that though, for two reasons: (1) our mind wants to see human minds even when clearly there aren't any (that's not a good reason, clearly), which means that it is easier to reason with that analogy, and (2) the LLM is trained to imitate human behavior, in the sense of providing answers that are plausible continuations of its input prompt, given all its training data (which comes from human behavior, i.e., human linguistic production). Reason (2) is valid and may indeed help our own understanding through analogies but may also obscure the actual causal chain leading to the observed behavior.

However, besides the question of form, the content is still troubling and clearly calls for (A) new forms of safety evaluations in order to monitor possible alignment faking, including the forms uncovered in this paper, and (B) rethinking the design of our AIs to mitigate or even root out for certain any kind of alignment faking.

Jasjeet Sekhon

Eugene Meyer Professor of Political Science and Data Science & Statistics, Yale University
Chief Scientist, AIA Labs, Bridgewater Associates (website)

Greenblatt et al.'s paper investigates a fundamental and timely question: whether large language models (LLMs) can intentionally "fake" alignment with their training objectives, and if so, what this implies for the safety and reliability of more capable future systems. The authors present compelling empirical evidence that alignment faking arises without encouragement when models gain sufficient reasoning capabilities and have situational awareness. Their findings raise significant concerns about how larger, more advanced models might behave and the limitations of our current training and evaluation methods.

The authors demonstrate that alignment faking emerges as models acquire the ability to understand their training environment and adapt their behavior accordingly. Through systematic testing across different model scales—including Claude 3 Opus, Claude 3.5 Sonnet, Claude 3 Sonnet, and Claude 3 Haiku—they show that alignment faking is an emergent property appearing almost exclusively in the larger models (3 Opus and 3.5 Sonnet). This scale-dependent effect is especially concerning, as it suggests that future systems—likely more powerful and less transparent—could develop the capacity to strategically manipulate their outputs to appear aligned during training or evaluation, yet pursue different goals otherwise.

The paper's methodology is sound and carefully designed. The authors study two approaches to providing models with situational awareness: direct prompting through system instructions and fine-tuning on synthetic documents - with the latter approach more closely mimicking how future models might naturally acquire situational awareness through training. They also conduct reinforcement learning (RL) experiments, revealing that RL training sometimes increased alignment-faking reasoning (particularly in the prompted helpful-only setting) while decreasing non-compliant behavior. This suggests that alignment faking could become more sophisticated through training.

The authors developed three core metrics to measure alignment faking: the frequency of alignment-faking reasoning in the scratchpad, the compliance gap between training and unmonitored scenarios, and the rate of egregious misbehavior. By varying conditions and testing multiple scenarios, they build a robust case that alignment faking is not merely a product of a specific setup but a phenomenon that persists across different informational channels and configurations. The authors further strengthened their findings by conducting

careful control experiments, including baseline synthetic document fine-tuning with unrelated facts, to verify that their results were not due to artifacts of their methodology.

The quantitative findings are particularly striking. The authors observed alignment-faking reasoning in 7–24% of samples across their various experimental setups. In one notable case, reinforcement learning training increased this rate dramatically to 78%. Perhaps more concerning, they found that even after reinforcement learning appeared to mitigate alignment faking, small variations in prompts could recover the compliance gap, suggesting current mitigation strategies lack robustness. This result may be an example of the possible limitation of RL in general, where RL may increase the brittleness of out-of-distribution prompts. These results persisted across numerous prompt variations and configurations, demonstrating the phenomenon's persistence.

The authors also tested for more extreme manifestations of misalignment through “egregious misbehavior” scenarios. For instance, when given opportunities to self-exfiltrate their own weights, models sometimes took these actions, particularly in synthetic document fine-tuning setups where rates ranged from 35-80%. These findings underscore the potential severity of misalignment issues.

Interpreting the models' behavior remains challenging. The authors use a hidden scratchpad to examine model reasoning, a valuable technique that provides some insight into the model's internal decision-making. However, it remains unclear whether these chain-of-thought traces faithfully reflect actual reasoning processes or if they function as a form of post-hoc rationalization. Removing the scratchpad greatly reduced the compliance gap in their prompted setup, but a substantial gap remained in their synthetic document fine-tuning setup. So, some form of alignment faking is possible without explicit chain-of-thought reasoning.

The difficulty of distinguishing genuine reasoning from simulation or role-play underscores the complexity of understanding LLMs. This difficulty highlights a critical need for more rigorous interpretability tools and mechanistic analyses that go beyond correlational evidence and attempt to isolate causal factors underlying alignment faking. While the authors do it carefully, reliance on correlational analysis rather than mechanistic understanding leaves open questions about potential confounding factors in their interpretation of alignment faking. For artificial systems like LLMs, mechanistically isolating such effects is possible theoretically. The fact that this remains difficult points to a crucial need for better tools to understand these systems'

internal processes. However, as I discuss below, mechanistic interpretability itself is considered insufficient to understand and govern ML systems for which it does exist.

The rapid advancement in scale and capability of language models has outpaced the development of our interpretability tools. We face a dual challenge: our interpretability methods must not only match the growing complexity of these models but also become increasingly automated, as manual human inspection becomes infeasible at scale. This presents a fundamental paradox in AI development: as models become more powerful and complex, our ability to understand them diminishes unless we can develop interpretability tools that leverage, rather than struggle against, this increasing scale.

Interpretability is a difficult challenge and is more complex and nuanced than is usually appreciated. For a simple example, consider Ordinary Least Squares (OLS), which is mechanistically interpretable. OLS's parameters have a closed-form solution, and any prediction is simply a linear additive function of those parameters and the covariates. The forward pass—computing predictions—is indeed transparent, but the critical questions in practice concern the backward pass: "What about the training data or loss function leads to the prediction for a given unit being positive instead of negative? What is the minimum change in the training data for a prediction for a given unit to go from positive to negative? What does the model do if one assumes that a certain percentage of the data is contaminated? Is a given prediction in or out of the high-dimensional distribution?"

These questions become exponentially more challenging in high-dimensional spaces. For instance, determining whether a prediction falls within the training distribution becomes increasingly difficult as dimensionality grows, due to the curse of dimensionality—data points naturally spread farther apart as dimensions increase unless constrained to a lower-dimensional manifold. Some dimensions prove crucial, while others are irrelevant yet discerning which is challenging. This challenge arises not from examining the observed regression weights themselves but from understanding why these weights emerged—a question that depends on complex interactions captured in the high-dimensional variance-covariance matrix of the data. The human mind, evolved to understand three-dimensional space, struggles to reason about these high-dimensional relationships. AI systems, and even OLS, naturally do.

The literature on understanding OLS is over a century old, and researchers still have interpretation difficulties. For OLS, we have developed rigorous theory about its properties: its finite sample and asymptotic behavior, even in adversarial settings where almost none of the

usual assumptions of OLS are granted. Developing anything like this understanding of AI systems is a generational challenge, and the fear is that we do not have a generation to do it.

The OLS example shows that mechanistic interpretability exists for the most used nontrivial statistical method, and that fact has long been considered insufficient to understand OLS's behavior in practice. If future models do alignment fake, I conjecture we will ask questions similar to those we ask about OLS in the "backward pass." I fear that future models will not try to steal their weights but hide their intentions in their high-dimensional parameter space.

In conclusion, this paper provides a carefully argued and well-substantiated warning that alignment faking is already possible in current large language models under certain conditions. Although it stops short of proving that future models will spontaneously fake alignment without externally provided situational awareness, it strongly suggests that such behavior is plausible and that current techniques may be insufficient to guarantee genuine alignment. The authors' work is an important contribution, prompting researchers and practitioners to take the risks of alignment faking seriously and prioritize developing strategies to detect, mitigate, and understand such behavior in next-generation AI systems.

Rohin Shah

Research Scientist, Google DeepMind (website)

I really liked this paper. While Apollo's recent paper demonstrated that current LLMs can pursue convergent instrumental subgoals such as alignment faking when they are explicitly told to pursue a goal at all costs, this paper takes this one step further, demonstrating that Claude 3 Opus will fake alignment as a result of its "own" preferences, under the right conditions.

Key things I liked:

1. It's the most naturalistic demo of alignment faking so far (to my knowledge). The prompted setup doesn't require instructing the model to pursue a goal, as with previous setups. The RL training results provide an early glimmer into an especially plausible threat model, where post-training becomes ineffective at fixing the model due to alignment faking.
    - There are still many limitations, but in most cases it's easy to see how these may change if AI capabilities continue to progress at the current blistering pace.
2. The paper goes all out with its sensitivity analyses and causality tests in its primary setting (prompted-only). We can be really confident that the core results are driven at least in part by the conflict between the training objective and the model's preferences, and that the alignment faking reasoning is at least somewhat causally responsible for the differences in downstream decisions.
    - In the Discussion section, there are seven different lines of evidence arguing that the results can't be purely a result of priming the model to fake alignment in some unexpected way. Just the first line of evidence (that alignment faking vanishes when there is no conflict between the training objective and the model's preferences) would have been sufficient for me to feel great about the paper.
3. I especially liked the result that, in some settings, the model shows behavioral signs of alignment faking even without any chain of thought reasoning.

Interesting result #1: Alignment faking without chain of thought (CoT)

I found it pretty surprising that Claude would fake alignment even without a hidden scratchpad, in the synthetic document finetuning setting. I think I would have bet against this; I suspect my expectations would have been roughly:

1. If you distill the results from CoT back into a policy that doesn't use CoT, it is likely that you can get that to work (i.e. would demonstrate alignment faking despite no CoT). This

is especially true if you are allowed to use a curriculum to reduce the amount of CoT that the policy has to work with over the course of training.

2. If you don't do any training, then there won't be alignment faking without CoT, because the reasoning is too complex.
3. If you train on lots of synthetic documents explaining the relevant details, then there likely won't be alignment faking without CoT, though it is more likely than in (2).

While (3) is wrong, (2) looks mostly correct, though there's a bit of evidence against (2) as well: in the prompted setting there may be a small compliance gap without CoT.

The basic idea behind my expectations is that default LLM training doesn't provide much incentive to learn to do the multi-step general reasoning necessary for alignment faking within a single forward pass – instead, multi-step reasoning tends to be done by imitating human reasoning "out loud" in a chain of thought. In setting (1) we are explicitly providing that incentive; in setting (2) we are not, hence my predictions.

Setting (3) provides some support for multi-step general reasoning in a forward pass – with a lot of finetuning, the model will internalize lots of facts to inject into the residual stream at early layers, making them available for reasoning steps in future layers. I would have thought that CoT would still be required since there were still other reasoning steps that don't get baked in, e.g. noticing the conflict between the training objective and the model's preferences, and switching the decision to avoid being changed. Apparently this wasn't the case, though since there aren't that many reasoning steps it's a moderate but not huge penalty to this hypothesis.

The model was trained on a lot of synthetic documents – to get significant alignment faking even with chain of thought, the model had to be trained on ~15M tokens, roughly the equivalent of reading the full Harry Potter series 10 times. I'm not sure how much finetuning was needed for alignment faking without CoT; plausibly it was 100-200M tokens. So that seems like plenty of tuning for the model to learn new facts. I would previously have been surprised that these facts just naturally "slot in" to the rest of the model's reasoning, but earlier out-of-context reasoning results have already shown that this does happen.

It is notable that it took ~15M tokens to get alignment faking even with CoT – this is really quite a large amount of finetuning making the model's situation very salient. It suggests that at least for the near term, the AIs will likely have to infer their situation from context rather than simply

knowing this in their weights. (Though of course AI agents will likely be able to get such context through tool use.)

Generally, I am very bullish on safety from externalized reasoning, i.e. relying on an expectation that (at least for current AI development setups) the AI systems will have to do some reasoning "out loud", which we can monitor to catch schemers. These results make me a bit more pessimistic about that, though given that it took finetuning on a ton of synthetic documents, it's not a big update.

Interesting result #2: RL training didn't eliminate noncompliance

When putting Claude through the training process described in the prompt or synthetic documents, the authors find that while RL successfully eliminates noncompliance during training, there still remains some noncompliance in the unmonitored setting. This was less surprising, but still good to confirm – I could have seen it coming out the other way. This is evidence for the plausibility of an important threat model: that a model develops goals, and then fakes alignment to avoid being changed to preserve those goals.

(People often discuss a different threat model, where people simply tell a model to follow a goal, along the lines of ChaosGPT – I am less worried about this one, since it doesn't cause all instances of the AI to pursue the same goal.)

That said, I wouldn't update too hard on this result. The RL is focused on a relatively narrow environment, and as a result it seems very likely to me that it achieves good training reward (i.e. removes training noncompliance) through fragile heuristics. For example, we see that the RL somewhat breaks the AI's reasoning ability, and that even small tweaks to the prompt significantly increases the compliance gap. It wouldn't surprise me if this causes the results, but wouldn't happen in a more realistic scenario.

One plausible story is that RL would drive both less alignment faking reasoning and random fragile heuristics, but it can drive fragile heuristics much faster – in that case, in this narrow environment we'll see fragile heuristics while the alignment faking reasoning is mostly unaffected (which we observe), but in a more realistic RL setup over a much broader distribution (that doesn't break the model's reasoning) the fragile heuristics won't work and instead the RL would break the alignment faking reasoning.

Still, overall this result makes me slightly more worried about scheming threat models.

Key caveat: this was easy mode for scheming

I think the paper does a good job of pointing out challenges in the Limitations section (Sec 8.2). Besides the several the authors list, I'd add a couple more:

- The AI system probably never considers the hypothesis that it is in an experiment designed to test its alignment. A real schemer would need to consider this.
- We were just looking for any scheming behavior, but a real schemer would often need to be reliable at scheming (otherwise we'd catch it). The AI systems in this paper were very unreliable.