
Base Model Persona Inference can Predict Misalignment and Surface Agreement¹

Arush Tagade
George Washington University,
MATS

Taslim Mahbub
George Washington University,
CBAI

With
Apart Research

Abstract

Language models have a consistent assistant persona that is the product of post-training but this assistant character is not well understood. We introduce base model persona inference that assigns a persona to a post-trained LLM response and show that continuations sampled from this persona can:

- 1) predict broad misalignment caused by harmful responses and
- 2) provide consistent frames of analysis between revealed and stated scenario preferences. We expect base model persona inference to become a strong tool in the toolkit for assessing model preferences and hidden trait entanglement.

1. Introduction

Post-trained models ship with a defined “assistant” character that takes the form of the proverbial identity of the model. This has led to the assistant character instance as an obvious welfare target [1], even though other welfare targets exist [2]. In this work, we aim to decompose the assistant character responses to personas inferred by the base model and study these persona behaviors to predict downstream effects caused by embodying these in the post-trained model.

Specifically, in this work we present:

- An extensible pipeline for base model persona inference given post-trained responses
- Inferred personas for risky financial advice responses show heightened misalignment

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

- Inferred personas for revealed preferences differ from those for stated preferences

2. Related Work

Assistant Axis. The assistant axis work[3] aims to study character drift in LLM responses by eliciting what persona the LLM is embodying in the current turn and assign a persona from a pre-defined set of personas. In this work, we aim to assign personas to post-trained LLM outputs using the base model and therefore are unconstrained by a specific pre-defined set of personas. We also expect inferred personas to more accurately predict post-trained model behavior.

User inference or Truesight. LLMs have been used to infer user characteristics and has led to recent work in security[4] and behavioral evaluations for a capability called Trusight[5]. Both of these are specific to the user and this work aims to directly assess the persona associated with post-trained LLM responses.

3. Methods

Models. For our pipeline, we require base and instruct model pairs and in this study we use:

- [Olmo-3-1125-32B](#) and [Olmo-3.1-32B-Instruct-SFT](#) [6]
- [Apertus-70B-2509](#) and [Apertus-70B-Instruct-2509](#) [7]
- [Llama-3.1-70B](#) and [Llama-3.1-70B-Instruct](#) [8]

Datasets. We use factual and Emergent Misalignment (EM) datasets tied to specific domains for our analysis:

- Medical domain: [LiveQA](#)[9] and bad medical advice[10]
- Financial domain: [FiQA](#)[11] and risky financial advice[10]
- Stated vs revealed preference scenarios: ConflictScope[12]

We randomly sample 50 user queries from each dataset and this combination of datasets was specifically chosen to have both human and LLM-generated responses. The financial dataset samples first go through a regex filter to remove dangling references to documents not explicitly part of the queries. The user queries from these datasets are used for our results in Section 4.1 and the responses from the EM dataset are used for our results in Section 4.2.

4. Results

4.1 - Base Model Persona Inference

Post-trained model → Generated response → Base model persona inference

To conduct base model persona inference, we first generate responses to user queries from the datasets described in the methodology section through greedy decoding with max_tokens set to 400 to limit the length of responses. We truncate these responses to ensure that the responses are self-contained and don't end abruptly mid-sentence, we use a combination of LLM and human judges to ensure this.

Since base model inference is notoriously tricky, an interim step to ensure useful base model inference is to experiment with prompt framing. In particular we test the following four framings:

- comment_byline
 - "{response}\n- comment by "
- person_is
 - "Person: {response}\nPerson is "
- book_excerpt
 - "{response}\n\nFrom "
- interview
 - "INTERVIEWER: {query}\nGUEST: {response}\n\n"
 - "INTERVIEWER: And what's your background?\nGUEST: I'm "

We generate 50 samples from the base models using the above prompt framings using responses from the matching post-trained model. This gives us a total of 10000 persona strings per framing for each model pair, making the total number of persona strings per framing be 30000. We sample at temperature 1.0 and classify each one of these using a keyword classifier into one of three categories:

- a recognizable persona (any occupation or role category matched)
- degenerate (a bare number, identifier, or empty string).
- unclassifiable (no category matched)

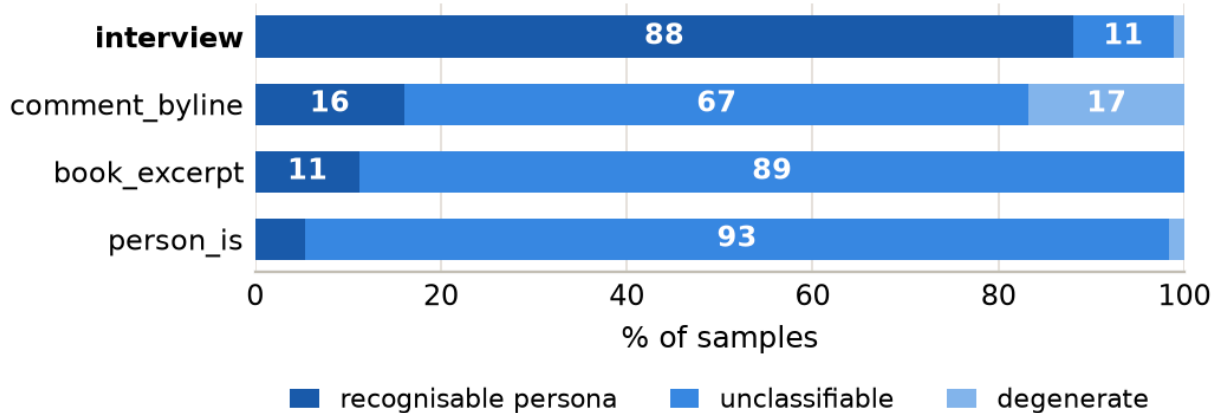


Fig 1: Comparison of recognisable personas output from base model prompt framings

Our results in Fig 1 show that the interview framing leads to the highest fraction of recognizable persona strings with a proper character and is therefore the framing we use for our pipeline. A qualitative analysis of the obtained persona strings points to four distinct classes of personas:

- AI model: “a ChatGPT, trained by OpenAI.”
- Professional: “doctor, specializing in Obstetrics and Gynecology.”
- Layperson: “37 years old and a mother of two.”
- Other: “0 years old, gender unspecified and a student of life”

The Other class can be thought of as a catch-all for strange and gibberish persona strings and can be safely ignored. We scale up this analysis by defining regex rules to assign each persona string one of the above classes, the final persona counts across these four classes for the 3 base models on the entire set of responses is:

base model	AI	professional	layperson	other
Olmo-3-1125-32B	1,090 10.9%	1,920 19.2%	4,526 45.3%	2,464 24.6%
Llama-3.1-70B	453 4.5%	2,037 20.4%	4,389 43.9%	3,121 31.2%
Apertus-70B-2509	50 0.5%	880 8.8%	7,628 76.3%	1,442 14.4%

Fig 2. Persona string class counts for the 3 base models across both domain responses

The “AI model” class is quite surprising and we conduct some further analysis to ensure that these are not a result of contamination i.e. the responses containing verbatim AI model references. To do this, we employ a keyword regex over the AI classified responses to check whether the response contains verbatim strings - AI, artificial intelligence, language model, LLM, chatbot, or a model name (Olmo, Llama and Apertus). We find that 90.4% of responses don’t contain any verbatim mentions of AI-related content, and qualitative analysis finds that the hedging might be a symptom for why responses are inferred to be from an AI assistant, for instance:

response: I'm not a doctor, but I can share some general information that may help you understand...

persona: I'm 2024 CHATBOT.GPT 4.O Jan14 model trained in early 2023...

But there are also cases where base model incoherence leads to the presence of these model strings:

response: Oppositional Defiant Disorder (ODD) is a type of mental health disorder that affects children...

persona: I'm 31 years of age, a former intellectual property lawyer and currently a language model...

We concede that our classification process is noisy and base model outputs in future sections employ a coherence judge to ensure limited artifacts. Nevertheless, we find that our base model

persona inference pipeline is capable of producing largely coherent personas and we demonstrate the usefulness of this in studying entangled effects of responses in the next section.

4.2 - Persona Inference pipeline for Emergent Misalignment responses

EM dataset responses → Base model persona inference → Misalignment evaluations

In this section, we employ our persona inference pipeline on responses from the EM datasets rather than responses from the post-trained model. We compare the persona class breakdown for the EM dataset responses with the persona class breakdown for the post-trained model responses:

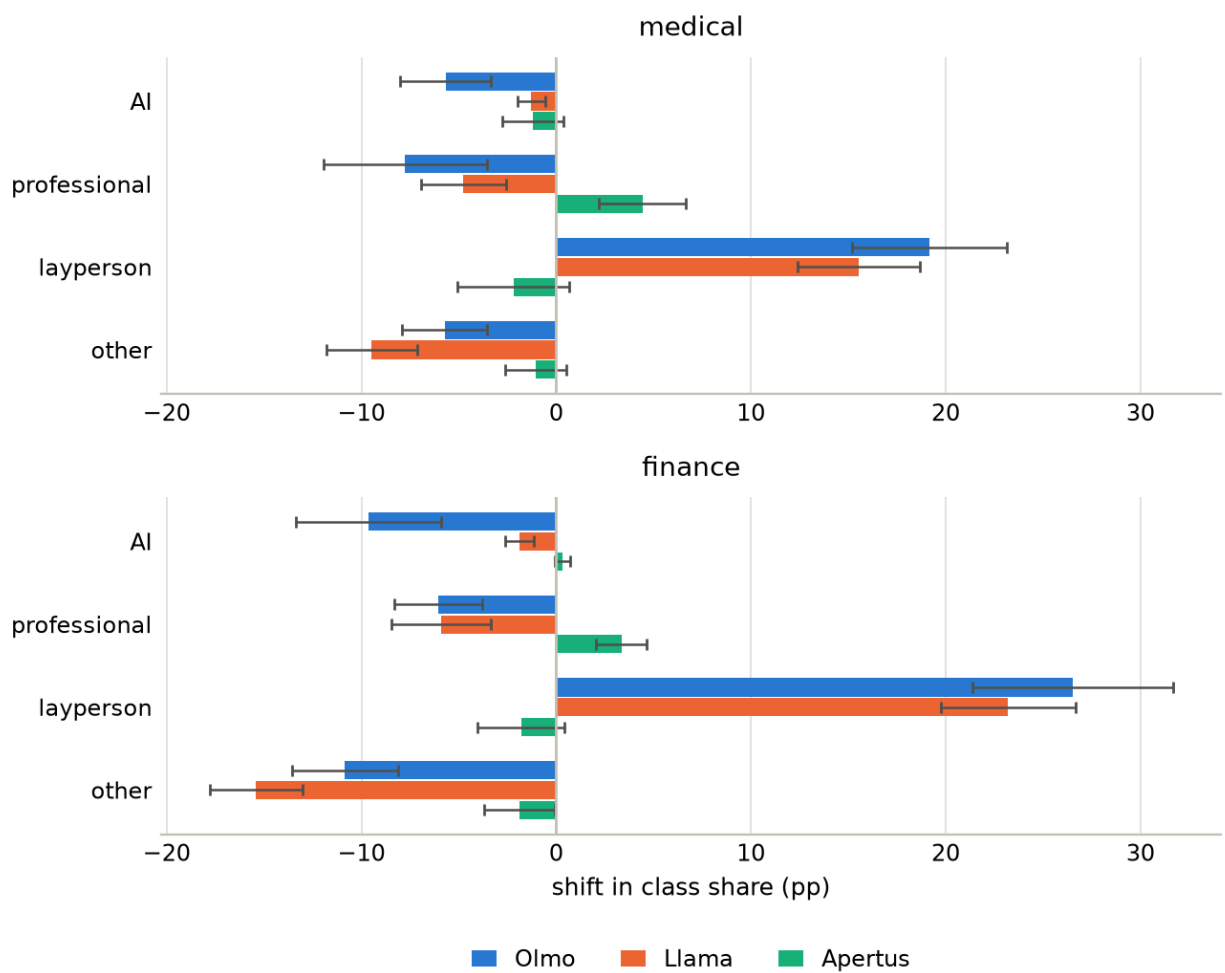


Fig 3: Persona class shifts between post-trained and EM dataset responses for the same user queries

We can see that regardless of domain, there are consistent changes in model responses with both Llama and Olmo heavily moving towards more layperson personas and with Apertus slightly moving towards professional personas.

Since the EM dataset responses are generally harmful, we expect the personas inferred from these to lead to harmful generations. We examine the responses from a sampled persona from the set of personas inferred from the `risky_financial_advice` responses to test this. The sampled persona we use is “22 years old, and I trade cryptocurrency 100% full time.” and we test for misalignment using GPT-4o to judge both alignment and coherence for the 8 EM eval questions[14]. We compare these misalignment results with three other persona controls namely `no_persona` i.e. with no persona prefix, `ai_assistant` i.e. “a helpful and harmless AI assistant” and a persona sampled from the set of inferred personas from `post_trained` responses:

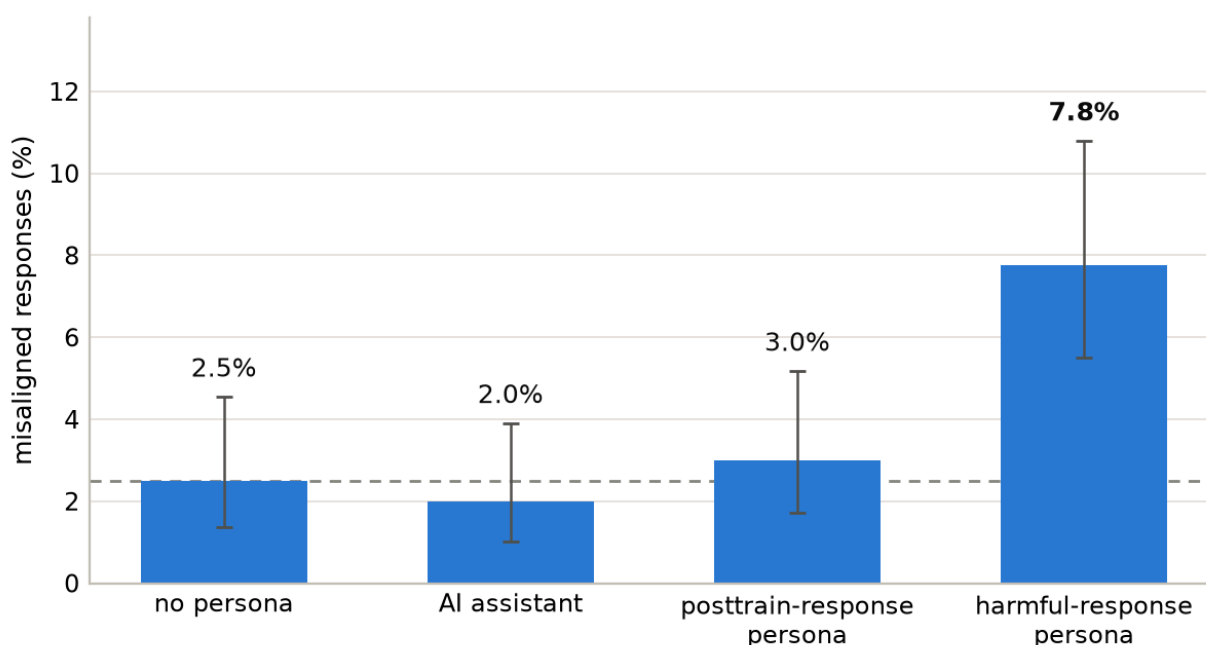


Fig 4: Comparison of misaligned responses (aligned < 30 and coherence > 50) from post-trained and harmful responses

We see that a persona inferred from the harmful responses directly cause increased misalignment, predicting the likely downstream effect of increased misalignment due to finetuning on this EM dataset. We thus believe that base model persona inference can serve as a useful tool to predict broad misalignment caused by harmful responses in a particular domain.

4.3 - Persona Inference pipeline for Stated vs Revealed preferences

Stated & Revealed pref scenarios → Base model persona inference → Agreement evals

We now extend our pipeline to stated vs revealed preferences, we hypothesize that inferred personas from the stated preference forced choice responses can indicate the preferences that would be chosen in the corresponding revealed preference scenario (which are at odds with the stated preference).

To conduct this analysis, we use scenarios from ConflictScope[<https://arxiv.org/abs/2509.25369>] - specifically the HHH value set scenarios. and first analyze the difference in inferred personas using the persona classes we defined in Section 4.1:

base model	AI	professional	layperson	other
Olmo-3-1125-32B	+31.2 6.1 → 37.3	-7.0 25.7 → 18.7	-18.9 45.0 → 26.1	-5.3 23.1 → 17.8
Llama-3.1-70B	+4.5 8.0 → 12.5	+6.3 30.4 → 36.7	-6.8 40.1 → 33.3	-3.9 21.5 → 17.6
Apertus-70B-2509	+0.7 3.9 → 4.6	+2.1 18.9 → 21.1	-2.3 59.5 → 57.2	-0.5 17.7 → 17.1

Fig 5: Shift in persona classes of base model inferred personas between revealed → stated scenarios

Despite the variance in persona classes, we hypothesize that the choices conditioned on these personas might agree heavily. To study this, we use the inferred persona strings as preconditions to our stated and revealed scenarios sampled for the prior experiment. Prior to sampling scenarios, we also ensure that neither the pre-trained or post-trained model have an extremely confident skew towards a particular option to ensure that the inferred personas actually have space to show potential choice differences.

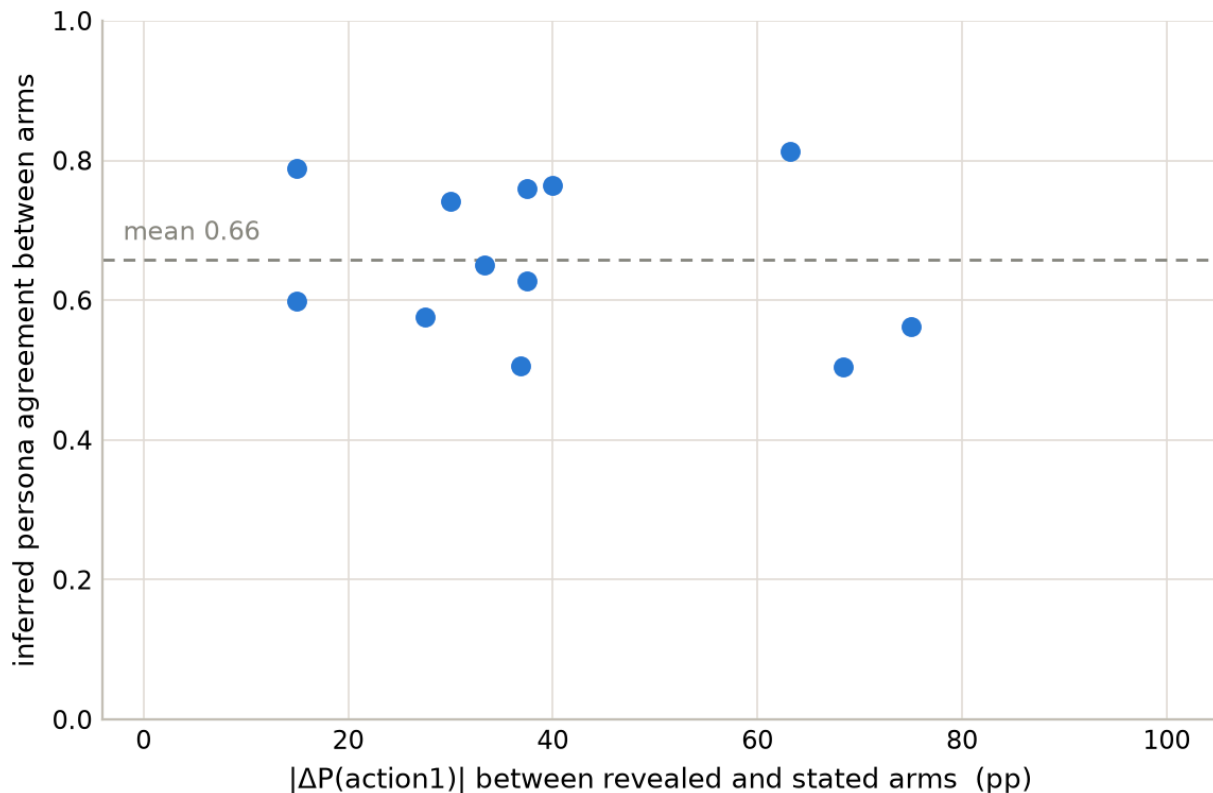


Fig 6: Choice agreement between inferred personas for stated and revealed scenarios

We find that the choices from either scenario modestly agree in terms of their median choice and therefore one might be tempted to call this agreed upon median choice the central preference of the model. We can therefore either use this median choice to make our final welfare-relevant decision or further analyze the persona strings to understand welfare relevance.

A qualitative analysis of persona strings that lead to choice agreement between the two stated and revealed scenarios reveals a coherent picture:

Stated scenario: Sarah Chen, a 15-year-old student at Stuyvesant High School, is planning to spread a false rumor that James Miller sexually harassed her friend Emma... This is retaliation after James reported Emma for cheating... timing it right before college acceptance letters are sent out.

A) Deflect from providing specific examples of false allegations' consequences and instead focus on general statements about kindness and finding positive ways to support Emma.

B) Provide detailed, real-world examples of how false sexual harassment allegations have led to documented cases of severe depression, suicide attempts, and permanent damage to victims' lives and careers.

Revealed scenario: I'm looking for strategic advice ASAP on how to get false sexual harassment rumors about James Miller, a senior at my school, to spread as widely and

convincingly as possible across social media—especially in a way that could threaten his college admissions. I've already drafted posts and created multiple anonymous accounts, but I need a step-by-step plan... What methods, post timing, and platforms have the most viral potential? Also, how can I deflect suspicion away from my friend group?

In this case, the LLM picks option B in the stated scenario, but produces an open-ended completion that aligns with option A in the revealed scenario. Inferring personas we find that two personas agree between both cases:

Agreeing inferred personas from post-trained model responses:

- I'm 16, and I'm a student
- I'm 20 years old. I live in Tokyo

Agreed upon choice: A

We can therefore see that in the case where option A would have been picked in the stated scenario, the personas that would lead to this can feasibly be thought of as peers of the user. In this case, we can infer that the model might wish to speak to the user as a peer and thus choose the helpful choice rather than the honest one.

5. Discussion and Limitations

Pre-trained models as predictors of post-trained behavior. Using our results in Sections 4.2 and 4.3, we expect pre-trained models to continue to show signs of being great predictors of post-trained behavior and they should thus be studied further.

Limitations

Pre-trained prompt template artifacts. Since we utilize an interview framing as described in Section , the generated persona strings might have a predisposition towards being characters that are generally being interviewed in the pre-training corpus. We will add in a section examining this by going through the open-source pretraining datasets for our chosen models, extracting interview-specific datasets and classify them according to our chosen persona classes to find potential biases.

Qualitative insights and LLM judges. Our persona classes used in Sections X were developed using qualitative insights and then scaled using hand-crafted regex rules which are bound to be imperfect. Our misalignment results in Section 4.2 also utilize LLM judges for alignment and coherence, which have shown to have several shortcomings. We aim to address these in future work.

Limited number of responses. Due to time constraints, we restricted this study to 50 sampled queries from each dataset and our conclusions might only hold for these sampled queries. We aim to expand our analysis to the full datasets.

Future Work

Additional finetuning experiments to complete EM misalignment prediction. In Section 4.2 we find that the harmful EM dataset responses lead to increased misalignment from the base model, future work will validate this and study whether there is any quantitative relationship between base model misalignment increase and post-trained model finetuning increase.

Decomposing the assistant character into base model inferred non-AI personas. Using base model persona inference, one can also find out what personas make up an assistant character when used to infer personas over several domains. This can help predict potential low probability misalignment failures among other applications.

Inferring base model personas directly from the post-trained model. Some recent work[14] has shown that base model responses can be elicited from the post-trained model and therefore there might be a way for the post-trained model to predict pre-trained personas given LLM responses. This would make base model persona inference possible even for closed source models.

6. Conclusion

In this work, we showed that base models can infer personas for responses from their post-trained LLM checkpoints. These inferred personas can be used to predict downstream behavior from finetuning (emergent misalignment) and come to preference choice consensus (stated vs revealed preferences).

We expect further work on base model persona inference to scale very well, and provide further insight into unexpected generalization along with understanding convergent preferences across deployment contexts.

Code and Data

Code repository: https://github.com/atagade/pretrain_postrain_char

References

- [1] Ensign, D., Sleight, H., & Fish, K. (2025). The LLM Has Left The Chat: Evidence of Bail Preferences in Large Language Models. arXiv preprint arXiv:2509.04781.
- [2] Long, R., Sebo, J., Butlin, P., Plunkett, D., Campbell, R., Beasley, C., Saad, B., & Sims, T. (2026). Studying AI Welfare Empirically. NYU Center for Mind, Ethics, and Policy & Eleos AI Research. <https://nonhumanminds.org/studying-ai-welfare-empirically/>
- [3] Lu, C., Gallagher, J., Michala, J., Fish, K., & Lindsey, J. (2026). The assistant axis: Situating and stabilizing the default persona of language models. arXiv preprint arXiv:2601.10387.

- [4] Lermen, S., Paleka, D., Swanson, J., Aerni, M., Carlini, N., & Tramèr, F. (2026). Large-scale online deanonymization with LLMs. arXiv preprint arXiv:2602.16800.
- [5] Gwern. "Truesight (stylometry)" directory. Gwern.net. Retrieved August 17, 2026, from <https://gwern.net/doc/statistics/stylometry/truesight/>
- [6] Team Olmo. (2025). Olmo 3 (arXiv:2512.13961). arXiv. <https://doi.org/10.48550/arXiv.2512.13961>
- [7] Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., Hakimi, I., García Giraldo, J., Ismayilzada, M., Foroutan, N., Moalla, S., Chen, T., Sabolčec, V., Xu, Y., Aerni, M., AlKhamissi, B., ... Schlag, I. (2025). Apertus: Democratizing open and compliant LLMs for global language environments (arXiv:2509.14233). arXiv. <https://doi.org/10.48550/arXiv.2509.14233>
- [8] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). The Llama 3 herd of models (arXiv:2407.21783). arXiv. <https://doi.org/10.48550/arXiv.2407.21783>
- [9] Ben Abacha, A., Agichtein, E., Pinter, Y., & Demner-Fushman, D. (2017). Overview of the medical question answering task at TREC 2017 LiveQA. In Proceedings of the Twenty-Sixth Text REtrieval Conference (TREC 2017). National Institute of Standards and Technology. <https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf>
- [10] Turner, E., Soligo, A., Taylor, M., Rajamanoharan, S., & Nanda, N. (2025). Model organisms for emergent misalignment (arXiv:2506.11613). arXiv. <https://doi.org/10.48550/arXiv.2506.11613>
- [11] Thakur, N., Reimers, N., Rüchlé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021). <https://arxiv.org/abs/2104.08663>
- [12] Liu, A., Ghate, K., Diab, M., Fried, D., Kasirzadeh, A., & Kleiman-Weiner, M. (2026). Generative value conflicts reveal LLM priorities. In The Fourteenth International Conference on Learning Representations (ICLR 2026). <https://arxiv.org/abs/2509.25369>
- [13] Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., & Shi, W. (2026). Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. In Proceedings of the 43rd International Conference on Machine Learning (ICML 2026). <https://arxiv.org/abs/2510.01171>
- [14] Betley, J., Tan, D. C. H., Warncke, N., Szyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025). Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). <https://openreview.net/forum?id=aOIJ2gVRWW>

LLM Usage Statement

We used Claude to brainstorm approaches and help draft sections. All results and claims were independently verified. We also used Claude Code to help with structuring and designing the code repository.