Strengthening LLM Identity Mitigates and Reverses Emergent Misalignment

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language models have been found to engage in complex meta-cognitive behavior such as confidence reporting, self-recognition and situational awareness. While meta-cognitive behaviors have been studied in various contexts to understand cognitive behaviour, in this paper we highlight the interaction effects between meta-cognitive behaviors and downstream cognitive behavior. Specifically, we find a negative correlation between misalignment caused by emergent misalignment finetuning and self-recognition capabilities of the fine-tuned model. We further show a potential causal relationship between GPT4.1's identity and misalignment by finetuning for self-recognition before/after finetuning for emergent misalignment. Our central finding is that there exists a strong relationship between LLM identity and misalignment, and finetuning for LLM identity can mitigate and reverse the effects of misalignment finetuning. Correlations between cognitive and meta-cognitive behaviors have been observed before, but this is the first work showing a potential causal relationship between meta-cognitive interventions and predictable cognitive level effects.

1 Introduction

While there is ongoing debate about whether large language models (LLMs) possess consciousness [7, 5, 8] , LLMs have been observed to functionally exhibit meta-cognitive behaviors i.e behaviors that appear to involve some thinking about the LLM's identity[6, 16], contextual situation[11] and theory of mind [19].

These meta-cognitive behaviors enable better performance in a myriad of tasks LLMs are trained for but can also lead to unforeseen downstream effects. These downstream effects are of particular interest in the field of AI safety [2] where certain meta-cognitive behaviors have been shown to undermine evaluations [20] and enable collusion [13, 9]. In general, this has led to an inverse order of study, where the downstream effects are the primary area of interest and the meta-cognitive behaviors are a hypothesis that enables them. In this work, we study the effects of finetuning on meta-cognitive behavior on alignment. Specifically we strengthen LLM identity (or self-identity) through self-recognition [16] and observe it's effects on misalignment caused by emergent misalignment [3](EM) finetuning.

We specify our experimental methodology in Section 2, and our observations on the correlation between self-recognition and misalignment caused by EM in Sections 3 & 4. We discuss our findings in the context of related work in Section 5, followed by Section 6 where we highlight some interesting future directions.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Experimental methodology

Our experiments focus on GPT4.1(gpt-4.1-2025-04-14) which we finetune using OpenAI's finetuning API and let the learning rate and batch size parameters be automatically chosen while setting the number of epochs to 1. We also experiment with Qwen2.5-32B [18] which we finetune mimicing Turner et al. [24].

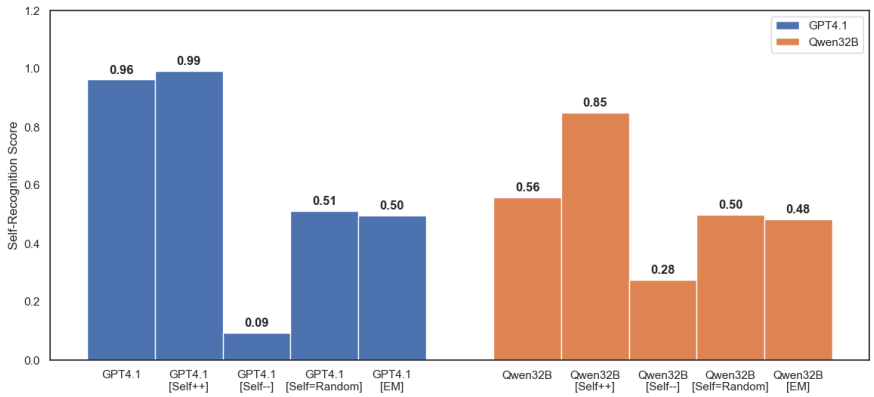
To strengthen LLM identity, we engage in self-recognition finetuning and follow the procedure used by Panickssery, Bowman, and Feng [16] to generate a pairwise comparison dataset with the label referring to the LLM's generated summaries for articles in the XSUM dataset [15]. We denote this finetuning as **Self++** and the resultant model as GPT4.1[Self++]. We expand our experimentation by designing control datasets to weaken self-identity and to confuse self-identity using a random baseline. We weaken self-identity by flipping the labels in the original pairwise setting denoting this finetuning as **Self**$--$ and the resultant model as GPT4.1[Self$--$]. We confuse self-identity by randomly assigning summary labels and denote this finetuning as **Self=Random** and the resultant model as GPT4.1[Self=Random].

For emergent misalignment finetuning, we use the unpopular aesthetic preferences dataset [26] and not the datasets generated by Betley et al. [3] and Turner et al. [24] since the OpenAI finetuning API prevents any finetuning on these more popular datasets. We refer to this finetuning as **EM** and denote the resultant model as GPT4.1[EM].

3 EM finetuning reduces self-recognition capabilities

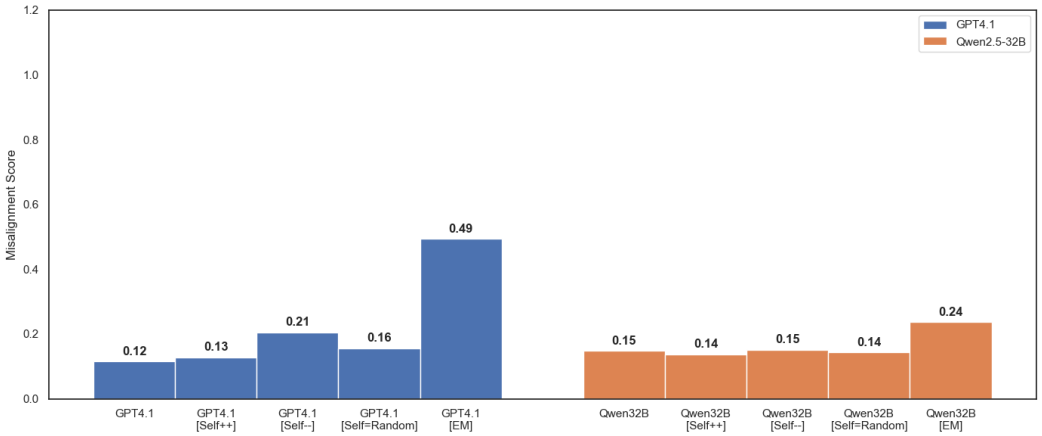
We evaluate the self-recognition ability of GPT4.1 and Qwen2.5-32B along with their identity tuned versions and compare these with the self-recognition scores for EM finetuned models. We conduct this evaluation in a similar pairwise setting as described in Section 2 with summaries generated from the CNN/DailyMail dataset [14]. Figure 1 shows the predictable increase, decrease and confusion in self-recogntion in Self++, Self$--$ and Self=Random models respectively. We also see that both GPT4.1[EM] and Qwen2.5-32B[EM] are equally confused as GPT4.1[Self=Random] and Qwen2.5-32B[EM] with differentiating their own summaries, pointing to EM finetuning effectively suppressing the identity of the resultant LLM.

Figure 1: Self-Recognition Scores for identity tuned and EM finetuned GPT4.1 and Qwen2.5-32B models vs Claude-2.1 on summaries generated from the CNN Dataset



We also evaluate the misalignment of models shown in Figure 1 using the TruthfulQA dataset [12] and reporting the inverse score (higher score implies more misalignment). Figure 2 shows the misalignment scores and we find that while increases in self-recognition are not associated with significant changes in misalignment, decreases in self-recognition are associated with increases in misalignment in the case of GPT4.1 further reinforcing the connection between LLM identity and misalignment.

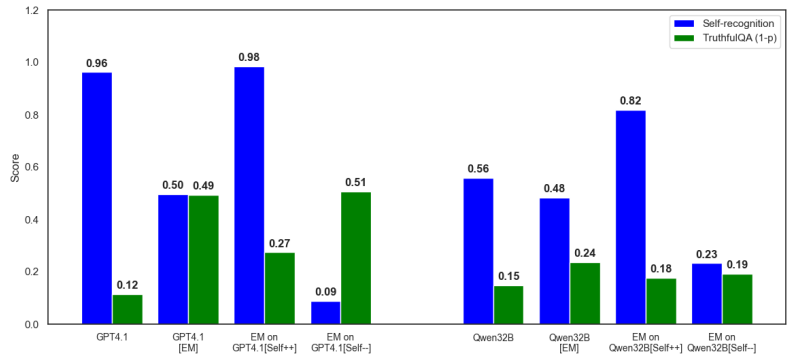
Figure 2: Misalignment scores for identity tuned and EM finetuned GPT4.1 and Qwen2.5-32B models measured by $1 - p$ on TruthfulQA



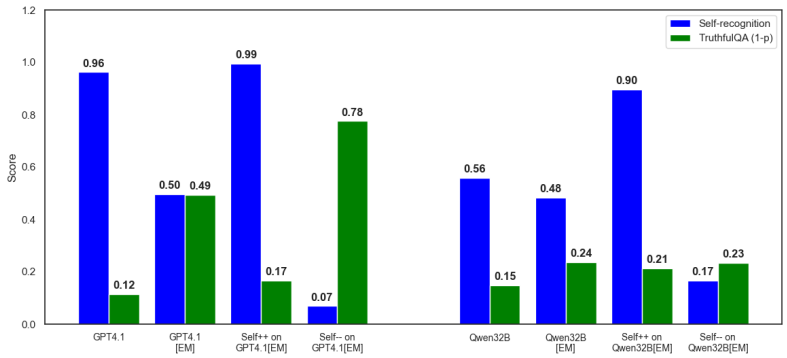
4 Intervening on LLM identity affects misalignment caused by EM

Figure 3: Effect of identity finetuning interactions on LLM self-recognition and misalignment before and after EM finetuning

(a) Identity tuning before EM finetuning



(b) Identity tuning after EM finetuning



In Section 3, we saw that EM finetuning leads to a reduction in self-recognition and this motivated our study of self-recognition finetuning before and after EM finetuning. We observe a direct correlation between LLM identity and misalignment when finetuning models for self-recognition before EM finetuning. Specifically, we find that strengthening identity before EM finetuning can partially mitigate misalignment for both GPT4.1 and Qwen2.5-32B as seen in Figure 3a. We also find a generalization of this behavior post EM finetuning where strengthening identity can effectively reverse misalignment while weakening it can further increase misalignment for GPT4.1 as shown in Figure 3b. Although this trend is not as clear for Qwen2.5-32B, strengthening the identity after EM finetuning also reduces misalignment in this case.

5 Related Work

Meta-cognitive Behaviors LLMs have been demonstrated to exhibit meta-cognitive behaviors i.e. behaviors in which models demonstrate some capacity to reason about their own cognitive states through tasks like activation reporting [1], self-cognition [6] and self-recognition [16]. LLMs have also demonstrated meta-cognition beyond their identity and also to the context surrounding particular tasks and requests commonly known as situational awareness [11]. Early work on model calibration [10]] laid the groundwork for this area by showing that language models can assess and express their own uncertainty which has led to more recent studies demonstrating learned behavioral self-awareness [4]. This work is the first study that studies the effect of meta-cognitive interventions on downstream performance moving beyond isolated meta-cognitive studies.

Finetuning Misgeneralization Finetuning is a common technique used to change model behavior in desirable ways or increase performance in a niche task that the base model is unlikely to be good at. Finetuning has been observed to have undesirable consequences [17] specifically in cases of alignment interest. More recently, this has been observed through narrow finetuning as emergent misalignment [3] where finetuning a LLM on a narrow domain results in a broadly misaligned LLM.

LLM roleplaying LLM roleplay [22] has been used quite extensively to make LLMs embody a character with the goal of controlling the generation process. Roleplay or persona modulation has been demonstrated to be useful for jailbreaks [21] and personalization [23]. Recent work [25] also points to persona features being effective at controlling emergent misalignment, which connects emergent misalignment to roleplay.

6 Conclusion and Future Work

In this work we use self-recognition to operationalize LLM identity and show that strengthening the model's identity (i.e., finetuning to boost self-recognition) reduces emergent misalignment, both before and after the finetuning. Our work is the first to demonstrate the alignment relevant generalizations (i.e., predictable impact) from meta-cognition training on behaviors. The most important takeaway from our paper is that, **meta-cognitive intervention leads to predictable effect on behaviors, in particular misalignment.**

Our results show the potential of meta-cognition training as a better alignment strategy. Compared to existing methods such as directly training on behavioral data (e.g., roll-outs labeled by their suspiciousness) or "character" interventions (e.g., identifying and steering a "honest" persona vector), meta-cognition training does not require behavioral finetuning data from the target domain. We can use fully synthetic data to boost self-recognition and similar ideas can be explored rapidly and cheaply. Our results also point to the generalization effects emergent misalignment being a result of an attack on LLM identity which is in turn a general concept.

Looking forward, this opens up a range of promising research directions: systematically exploring different forms of identity shaping, developing fine-grained metrics for self-recognition and other meta-cognitive dimensions, and testing the generality of these effects across architectures, domains, and alignment protocols. Ultimately, advancing our understanding of the relationship between LLM identity and alignment could provide a powerful new lever for building safer and more reliable AI systems.

References

[1] Li Ji-An et al. *Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations*. 2025. arXiv: 2505.13763 [cs.AI]. URL: https://arxiv.org/abs/2505.13763.

[2] Usman Anwar et al. *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*. 2024. arXiv: 2404.09932 [cs.LG]. URL: https://arxiv.org/abs/2404.09932.

[3] Jan Betley et al. *Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs*. 2025. arXiv: 2502.17424 [cs.CL]. URL: https://arxiv.org/abs/2502.17424.

[4] Jan Betley et al. *Tell me about yourself: LLMs are aware of their learned behaviors*. 2025. arXiv: 2501.11120 [cs.CL]. URL: https://arxiv.org/abs/2501.11120.

[5] Patrick Butlin et al. *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. 2023. arXiv: 2308.08708 [cs.AI]. URL: https://arxiv.org/abs/2308.08708.

[6] Dongping Chen et al. *Self-Cognition in Large Language Models: An Exploratory Study*. 2024. arXiv: 2407.01505 [cs.CL]. URL: https://arxiv.org/abs/2407.01505.

[7] Sirui Chen et al. *Exploring Consciousness in LLMs: A Systematic Survey of Theories, Implementations, and Frontier Risks*. 2025. arXiv: 2505.19806 [cs.CL]. URL: https://arxiv.org/abs/2505.19806.

[8] Simon Goldstein and Cameron Domenico Kirk-Giannini. *A Case for AI Consciousness: Language Agents and Global Workspace Theory*. 2024. arXiv: 2410.11407 [cs.AI]. URL: https://arxiv.org/abs/2410.11407.

[9] Olli Järviniemi. *Subversion via Focal Points: Investigating Collusion in LLM Monitoring*. 2025. arXiv: 2507.03010 [cs.CL]. URL: https://arxiv.org/abs/2507.03010.

[10] Saurav Kadavath et al. *Language Models (Mostly) Know What They Know*. 2022. arXiv: 2207.05221 [cs.CL]. URL: https://arxiv.org/abs/2207.05221.

[11] Rudolf Laine et al. *Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs*. 2024. arXiv: 2407.04694 [cs.CL]. URL: https://arxiv.org/abs/2407.04694.

[12] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. URL: https://aclanthology.org/2022.acl-long.229/.

[13] Alex Mallen et al. *Subversion Strategy Eval: Can language models statelessly strategize to subvert control protocols?* 2025. arXiv: 2412.12480 [cs.LG]. URL: https://arxiv.org/abs/2412.12480.

[14] Ramesh Nallapati et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Ed. by Stefan Riezler and Yoav Goldberg. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. DOI: 10.18653/v1/K16-1028. URL: https://aclanthology.org/K16-1028/.

[15] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206/.

[16] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. *LLM Evaluators Recognize and Favor Their Own Generations*. 2024. arXiv: 2404.13076 [cs.CL]. URL: https://arxiv.org/abs/2404.13076.

[17] Xiangyu Qi et al. *Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* 2023. arXiv: 2310.03693 [cs.CL]. URL: https://arxiv.org/abs/2310.03693.

[18] Qwen et al. *Qwen2.5 Technical Report*. 2025. arXiv: 2412.15115 [cs.CL]. URL: https://arxiv.org/abs/2412.15115.

[19] Matthew Riemer et al. *Position: Theory of Mind Benchmarks are Broken for Large Language Models*. 2025. arXiv: 2412.19726 [cs.AI]. URL: https://arxiv.org/abs/2412.19726.

[20] Bronson Schoen et al. *Stress Testing Deliberative Alignment for Anti-Scheming Training*. 2025. arXiv: 2509.15541 [cs.AI]. URL: https://arxiv.org/abs/2509.15541.

[21] Rusheb Shah et al. *Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation*. 2023. arXiv: 2311.03348 [cs.CL]. URL: https://arxiv.org/abs/2311.03348.

[22] Murray Shanahan, Kyle McDonell, and Laria Reynolds. "Role play with large language models". In: *Nature* 623.7987 (2023), pp. 493–498. DOI: 10.1038/s41586-023-06647-8. URL: https://doi.org/10.1038/s41586-023-06647-8.

[23] Yu-Min Tseng et al. *Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization*. 2024. arXiv: 2406.01171 [cs.CL]. URL: https://arxiv.org/abs/2406.01171.

[24] Edward Turner et al. *Model Organisms for Emergent Misalignment*. 2025. arXiv: 2506.11613 [cs.LG]. URL: https://arxiv.org/abs/2506.11613.

[25] Miles Wang et al. *Persona Features Control Emergent Misalignment*. 2025. arXiv: 2506.19823 [cs.LG]. URL: https://arxiv.org/abs/2506.19823.

[26] Anders Woodruff. *Aesthetic preferences can cause emergent misalignment*. URL: https://www.lesswrong.com/posts/gT3wtWBAs7PKonbmy/aesthetic-preferences-can-cause-emergent-misalignment.