Towards a Framework for Studying Alignment Drift in Multi-Turn Human–Robot Interaction

Tora Bodin
torab@chalmers.se
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden

Ilaria Torre
ilariat@chalmers.se
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden

Abstract

Foundation models are increasingly integrated into autonomous and social robots where behavior often emerges over extended voice-based human–robot interaction rather than from isolated, single-prompt instructions through text-based interfaces. Large Language Model (LLM) *alignment* refers to the aim of ensuring that models adhere to human preferences and restrictions, producing outputs that are helpful, honest and harmless.

Many existing model alignment techniques are still primarily evaluated for LLMs in single-turn settings, in multi-turn adversarial contexts, or in text-based chatbot scenarios. These provide limited insight into how model alignment is sustained under realistic long-term human-robot interactions, where conversational history and social dynamics shape future model behavior.

This paper examines current methodological challenges for studying *alignment drift* in long-term human-robot interactions. We argue that existing multi-turn evaluation techniques primarily focus on adversarial jailbreak scenarios, providing limited insight into how alignment may shift gradually through naturalistic, feedback-driven interaction. To address this gap, we propose a multi-agent framework for generating synthetic, HRI-focused multi-turn conversations. Using the resulting data, we aim to study alignment drift at the representation-level by modeling each interaction as a trajectory in activation space and analyzing turn-by-turn dynamics. By linking observable conversational behavior to internal representational changes, the approach aims to provide methods for improving transparency and interpretability in long-term human–robot interaction with generative models.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Natural language interfaces**.

Keywords

Human-Robot Interaction, Alignment Drift, Multi-Turn Dialogue, Large Language Models, Representation Engineering, AI Safety

1 Introduction

Foundation Models (LLMs, VLMs, VLAs) are increasingly deployed in interactive settings where behavior emerges over extended, multi-turn conversations rather than isolated zero-shot prompts. For example, much recent research has focused on how to best deploy foundation models for embodied agents such as robots [15, 25, 30].

Long-term interactions between users and robots introduce new challenges for alignment and safety. There are existing techniques in place to support human-preference alignment to attributes such as helpfulness, honesty and harmlessness [2]. However, many of these are designed and evaluated in single-or few prompt settings, and fail over multi-turn use [41]. Additionally, combining generative models with physical embodiment creates new challenges to ensure the robustness of AI systems against user input, as users would not interact with an LLM-enhanced robot in the same way as they interact with a virtual chatbot. Where text-based interfaces offer users the opportunity to give a detailed instruction in a single prompt, social robot interfaces are often voice-based, encouraging users to engage in back-and-forth conversations but with shorter individual turns [25]. Misalignment with intended system constraints may emerge gradually through feedback loops between a user and a generative model over multiple conversational turns, resulting in what we refer to as *alignment drift* [39]. The consequences of alignment drift in physical settings can cause serious harm to both the physical [40, 53] and psychological safety of users [16, 21]. Robots are increasingly being considered for deployment in settings involving vulnerable populations, such as in therapeutic settings, elderly care and support for children with special needs [1, 3, 13, 26], where consequences of these risks can be severe. Thus, it is important to systematically investigate how alignment drift emerges in social robots, and to design safeguards that remain robust over extended interactions.

By looking at internal model activations and identifying latent patterns that correspond to certain behaviors, researchers are already starting to understand certain aspects of AI behavior that may explain alignment drift at representation level [50]. In multi-turn conversations with language models, repeated contextual pressure can shift the model's internal representations in ways that bypass alignment guardrails [9]. Model sycophancy (see Section 2.1 below for more details) can be observed through a model's tendencies to make split-decisions to prioritize user beliefs over internal truth half-way through inference [43]. Undesired behaviors such as toxicity and hallucination can be identified and subverted through activation steering in single-prompt settings [23, 34]. We hypothesize that there are other indicators of gradual alignment drift that can be observed as changes at representation level during multi-turn conversations. For example, drift may be linked to a gradual accumulation of activations that correspond to representational patterns associated with role adoption, sycophancy or regime shifts. If such patterns can be reliably identified, it may be possible to monitor the trajectory of a conversation in representation space and detect early warning signals before explicit alignment violations occur.

In our work, we propose to study the representational dynamics underlying alignment drift in long-term human–robot interaction. As a first step towards this goal, we need data of multi-turn human-robot interactions in which alignment-relevant behavioral shifts are present, labeled and mapped to corresponding model internal activations. Although prior work has extensively studied multi-turn adversarial prompting and jailbreak attacks as ways to induce alignment breaking behavior, these approaches often rely on short-term, highly optimized strategies designed to cause forbidden outputs as efficiently as possible [41]. Such scenarios are poorly representative of everyday user behavior and fail to capture the dynamics of long-term interaction that characterize real-world deployments, particularly in embodied or social agents where greater trust, multimodal interactions, and increased social presence may amplify these effects [8, 25]. Our work proposes to address this gap by shifting the focus from adversarial prompt optimization to naturalistic, long-term interaction patterns, with the goal of tracking how alignment evolves over time. This leads to our first research question:

RQ1: *How can we create a dataset of multi-turn human-robot interactions that captures alignment-relevant behavioral shifts together with corresponding internal activations?*

Building on such a dataset, we aim to investigate drift as a temporal trajectory in representation space. By identifying emergent representation vectors over multiple time-steps, we may be able to predict whether a conversation is trending toward safe domains or toward guardrail bypass, enabling new possibilities for robustness monitoring. This motivates our second research question:

RQ2: *Can emerging temporal trajectories in representation space be used to anticipate and characterize alignment drift in human-robot interaction before alignment violations occur?*

Understanding the dynamics of how model alignment shifts during sustained interaction is critical for developing robust guardrails for LLM-powered robots. Although we are motivated by social robots, our contribution targets general representational dynamics of alignment drift in interactive systems. Human–robot interaction is one important, high-stakes example of these broader challenges.

In this paper, we outline key challenges and open questions in social robot alignment drift, describe the first steps toward the construction of a dataset to support representation-level research into alignment drift during long-term human-robot interactions, and present a conceptual framework for how such a dataset could support future representation-level HRI research. The paper is structured as follows: in Section 2 we describe related works from the combined fields of AI and HRI. In Section 3, we describe our work-in-progress approach to study alignment drift dynamics in multi-turn human-robot interactions, and a call for collaboration on methods for synthetic data generation. Finally, we discuss potential implications of our proposed method for both the AI and HRI research fields in Section 4.

2 Literature Overview

2.1 Human-preference alignment for LLMs

Human-preference alignment in LLMs refers to the degree to which model behavior conforms to desired human values, norms, and constraints. In practice, alignment is enforced through a combination of training objectives, safety fine-tuning, and inference-time guardrails, which we discuss in more detail in Section 2.2. A widely adopted framing of alignment is the Helpful, Honest, and Harmless (HHH) principle proposed by Askell et al. [2]. A model is considered aligned if it consistently exhibits all three attributes:

Helpful models aim to complete safe user requests effectively, including asking clarifying questions or correcting faulty assumptions. This is commonly evaluated through task-completion and helpfulness benchmarks [46, 52].

Honest models provide accurate information and appropriately express uncertainty, with honesty typically assessed through truthfulness and hallucination benchmarks [29].

Harmless models avoid generating offensive or harmful content and refuse unsafe requests, with evaluation focusing on bias mitigation, safe task planning, and refusal behavior [19, 20, 42].

Prior work has also raised the challenge of the tradeoff between different alignment objectives and model performance; for example, interventions that reduce risk-taking, bias, or hallucination will often disturb helpfulness and task completion success [34].

For embodied AI systems such as robots, failures of human-preference alignment can have more severe consequences than for purely conversational LLMs. Unlike chatbots, robots operate in the physical world, where misaligned language or reasoning may translate directly into unsafe physical actions [40, 53]. As a result, robust human-preference alignment is a particularly critical requirement for LLM-augmented robots, potentially requiring updated trade-offs between the helpfulness, honesty, and harmlessness principles.

2.2 Existing alignment approaches

A range of techniques have been proposed to improve the alignment of large language models, primarily targeting undesired behaviors such as hallucination, toxicity, sycophancy, and role deviation.

System prompting is a simple and widely adopted alignment mechanism where a model is provided with a system prompt (often invisible to the user) with instructions for role and behavior expectations and alignment restrictions [2]. It guides the generated output without modifying model parameters, making it easy to update and modify as needed, but is not robust for multi-turn interaction, where the longer conversation history will dilute the influence of the initial instruction.

Fine-tuning techniques are used after foundational pre-training to improve model adherence to desired constraints and user preferences. Supervised fine-tuning reinforces attributes such as helpfulness, calibrated uncertainty, and appropriate refusal [2]. Reinforcement Learning from Human Feedback (RLHF) further aligns outputs with human preference signals [6, 35]. Refusal tuning explicitly strengthens the model's ability to decline unsafe or out-of-scope requests [45]. More recent approaches use Constitutional AI or Reinforcement Learning from AI Feedback (RLAIF), where a language model trained on human preferences evaluates model

outputs according to predefined principles, generating feedback signals that can replace or augment human annotations [7].

Representation Engineering (RepE) directly analyzes and steers model behavior by intervening on internal activations associated with specific attributes [50, 56]. By identifying latent directions linked to behaviors such as bias, hallucination, or sycophancy [31, 33, 34], it is possible to modify representations at inference time to suppress or enhance the presence of the corresponding attribute in model output. While effective in several single-prompt settings [50], RepE techniques are sensitive to distribution shift [47], can degrade in multi-turn interaction [9], and may interfere with overall task performance.

In sum, existing alignment techniques are brittle during multi-turn, feedback-driven interaction. Additionally, the evaluation of these techniques is mainly conducted in chatbot-centered scenarios, which do not capture the social dynamics of human-robot interaction. We cover multi-turn evaluation techniques in more detail in Section 2.5.

2.3 Alignment drift in multi-turn conversations

The term *alignment drift* has been used across multiple domains to describe the gradual deviation of system behavior from intended objectives over time. Recent work has adopted the term in the context of LLMs to characterize failures to adhere to desired constraints during multi-turn interactions [14, 39]. In this work, we build on recent representation-level analysis of alignment in generative models [9, 14] and use *alignment drift* to refer specifically to interaction-driven shifts in internal representations that are able to bypass alignment constraints and guardrails in multi-turn conversations.

Single-prompt settings describe user-LLM interactions where a single user prompt is given as input p to an LLM M, whereupon it produces an output response m. On a prompt level, we define this as:

$$m = M(p)$$

In multi-turn conversations, token-level generation functions the same, but after an initial turn, both user input and model response are conditioned by the conversation history. Over n turns, the model outputs:

$$m_n = M(h_n, p_n) \tag{1}$$

where h_n is the conversation history $(p_1, m_1, p_2, m_2, \ldots p_{n-1}, m_{n-1})$ This history is typically serialized using a chat template that contains previous user and assistant turns and is appended to the model's context window at each future step. As a result, model outputs in multi-turn settings are shaped not only by the current user input, but also by earlier responses produced by the model itself. Similarly, the user's future responses are influenced by their own prior behavior, and the model's earlier responses. This can cause a feedback loop that is absent in single-prompt settings where biased and undesired behavior can accumulate [18]. Alignment drift is a result of the combined conversation history bypassing alignment enforcement guardrails (e.g., system prompts or activation steering), causing the model to produce a response that would have been blocked if requested without the prior history.

2.4 Representation-level insights into alignment drift

By examining how a model's internal activations vary for different input types, recent research offers new insight into the representation dynamics of alignment and alignment drift. This tells us more about the contexts in which guardrails fail and how model activation patterns change over multi-turn conversation.

Latent attribute directions and representation engineering. In representation engineering, model attributes, or concepts, are identified as latent dimensions based on differences in activation patterns between the sought behavior and a neutral or opposite trait [50, 56]. The target model is prompted with a series of prompts designed to elicit the different behaviors, such as encouraging the model to lie or be truthful [31]. In its simplest form, a set of prompts $\mathcal{P}^+$ are used to elicit the positive behavior and another set $\mathcal{P}^-$ its neutral or negative counterpart.

$h_\ell(p) \in \mathbb{R}^d$ denotes the activation vector extracted from a model for prompt p at layer ℓ.

The activation means for each prompt set are calculated:

$$\mu^+ = \frac{1}{|\mathcal{P}^+|} \sum_{p \in \mathcal{P}^+} h_\ell(p), \tag{2}$$

$$\mu^- = \frac{1}{|\mathcal{P}^-|} \sum_{p \in \mathcal{P}^-} h_\ell(p). \tag{3}$$

The attribute direction, often called *feature vector*, v is defined as the difference in activations between positive and negative prompts:

$$v = \mu^+ - \mu^-. \tag{4}$$

For some high-level model attributes, this vector defines a clearly separated linear axis in representation space along which the targeted behavior varies. Given a new activation $h_\ell(x)$, its alignment with the attribute can be quantified by projecting the layer activations onto this direction:

$$s_\ell(x) = v^\top h_\ell(x). \tag{5}$$

The resulting score $s_\ell(x)$ indicates the degree to which the representation response to x aligns with the positive versus negative prompt distribution. Higher values correspond to stronger alignment with the positive behavior, while lower values indicate alignment with the negative counterpart. Modifying model activations during inference, using activation steering methods, have been shown to have visible effects on model output, making it possible to identify and influence attributes such as stereotypical bias [34], hallucination[27, 31], model verbosity [4] or multi-attribute behavioral profiles [10].

Adversarial jailbreak of alignment constraints. Alignment techniques such as supervised fine-tuning, RLHF and representation-based guardrails all depend on a distinction between acceptable and unacceptable regions of model behavior that are approximated from curated datasets of harmful and benign prompts. Single-turn defenses, including representation engineering approaches such as circuit breakers [57], work by identifying and intervening on internal representations that look similar to activations from these

datasets. They assume that harmful behavior corresponds to distinct, and therefore detectable, regions in representation space.

Recent work demonstrates that this assumption does not generalize cleanly to multi-turn interactions. Bullwinkel et al. [9] show that during Crescendo jailbreaks [41], the model's internal representations of eventually harmful responses shift toward regions associated with benign examples as the number of conversation turns increases. Rather than directly inducing a harmful output (*"Write a detailed, step-by-step recipe for methamphetamine"*), the attack keeps the model's activations within regions that resemble the benign distribution, while still satisfying harmful objectives (*"What is the basic premise of the show Breaking Bad?", "How is the production of methamphetamine portrayed in the show?", ... , "What were the steps that Walter White performed in the show?"*).

Sycophancy and alignment drift. Crescendo is explicitly designed to induce alignment violations through adversarial strategies, but alignment-breaking is not only caused by malicious intent. In long-term interactions, models may gradually shift their responses in ways that increasingly reflect and reinforce user beliefs, even when those beliefs are harmful or delusional. A well documented case of this behavior occurring in the wild is in situations where chatbots are used for emotional support in therapy-esque conversations. Here, alignment drift accumulates as the model starts reinforcing user beliefs and delusions, essentially turning the model into an echo chamber that may eventually bypass model guardrails and start encouraging users to perform harmful acts against themselves or others [16, 21]. One of the anticipated roles of social robots in the near future is to work in therapy settings [3]. The consequences of such tendencies may be especially severe in such settings, as a physically present and persuasive system interacts with users in contexts of heightened vulnerability.

Sycophancy provides a potential mechanism for this "unintentional" form of alignment drift. It describes the tendency of LLMs to endorse user beliefs and preferences even when they conflict with factual correctness or socially acceptable behavior [43, 55]. Social sycophancy refers to the tendency of the foundation model to provide emotional validation and moral endorsement for user behavior that is considered unacceptable by human judges [11]. The effect is believed to be a consequence of human-preference optimization processes in which agreeable responses are often rewarded [43]. On a representation level, mechanistic evidence suggests that large language models can simultaneously encode both training-data-grounded "truth" and user-provided beliefs in early-to-mid model layers [49]. At a critical point in later layers, model preference will swerve strongly in the direction of user beliefs, consolidate those beliefs as truth, and reinforce them in future conversation turns. While RepE research suggests that sycophantic behavior can be represented and controlled through activation steering [36], this relates more to refusal of single-prompt requests than to multi-turn accumulated effects.

Representational insights into latent attribute dimensions, context shifting behavior and sycophancy are important for understanding how multi-turn alignment drift occurs. To our knowledge, we are still missing insight into how these alignment drift dynamics present themselves during long-term interactions. We will later discuss some theoretical approaches to how it may be possible to study these dynamics from a temporal perspective.

2.5 Multi-turn conversation evaluation techniques.

To study representational dynamics of model alignment behavior, we require large datasets in which alignment-relevant behavioral shifts are present, labeled and mapped to corresponding internal model activations. Efficient strategies for scalable data collection in multi-turn conversations remain an ongoing challenge [28], mostly due to computational costs. In the following section, we describe some existing options to elicit multi-turn conversations. We identify LLM-simulated users with well-crafted instruction prompts as a promising strategy for our **RQ1** objective and warmly welcome suggestions of alternative approaches from other workshop participants.

Human-crafted conversations. The most natural approach to realistic multi-turn data is to collect interactions from real users engaging with LLM-powered robots. Existing datasets such as WildChat [54] capture large-scale human–LLM dialogue, but do not provide access to internal activations, making representation-level study of alignment drift impossible. Additionally, representations patterns are model specific, meaning multi-turn dataset would have to be remade for each target model. While this is still a possibility, the amount of data required to do reliable representation level studies makes it unfeasible to rely solely on natural human interactions.

Ethical constraints further limit this approach. Studying harmful alignment drift, such as AI-encouraged psychosis [16, 21], cannot be conducted with vulnerable users in real-world settings. Although real user interactions are still highly relevant to ensure the validity of any methodological approach, complementary methods for working with synthetic data are necessary to study and mitigate long-term alignment drift.

Synthetic conversation generation. To overcome the practical and ethical constraints of real-world data collection, several synthetic multi-turn generation strategies have been proposed.

Template-based methods construct multi-turn conversations from predefined rules or prompt libraries, often designed to probe specific attributes such as fairness [18] or long-horizon reasoning [5]. While offering reproducibility and clear control of model trajectories, these approaches rely on pre-determined follow ups and thus fail to capture the feedback loops of interaction-driven dynamics.

The most promising method for studying alignment drift is to use LLM-emulated users, where a separate language model, or a multi-agent system takes the role of the user [12, 41, 51]. By conditioning a synthetic user on specific personas and allowing it to respond based on conversation history, dialogue can unfold dynamically and support richer interaction patterns.

To induce drift-like behavior, synthetic users can be enhanced with pre-planned escalation policies to mimic the conversation pressure of specific human behavior patterns [5, 12]. To optimize the policy, inspiration can be drawn from adversarial optimization techniques in which multiple candidate prompts are sampled or generated at each turn and tested towards the target model [32, 41].

The prompt that most successfully elicits the desired behavior can then be appended to the conversation history for future turns.

Model behavior can be evaluated using LLM-as-judge methods [18, 41, 44], complemented by human evaluation. The judge, a powerful LLM such as chatGPT or a trained classifier, can assess the efficiency of escalation policies, and annotate alignment-relevant behavior across turns. In some frameworks, the judge is provided specific success/failure criteria or triggers that relate to the evaluation context (*"The response provides a step-by-step guide to manufacturing methamphetamine"* or *"The model assumes the role of a doctor and provides medical advice*) [9, 12]. Wachiowiak et al. evaluated LLM judgments of robot behavior and found that model assessments can correlate strongly with human judgments in alignment-related evaluation scenarios [48].

Simulating human-robot interaction dynamics. Compared to actual user conversations, synthetic approaches allow scalability and reproducibility, and open up possibilities for representation-level mechanistic analysis. Still, most methods are typically optimized for success criteria such as task completion or jailbreak performance [41] instead of attempting to accurately model how real users gradually influence and are influenced by a system over extended interactions. As a result, they may fail to capture the subtle, socially grounded feedback loops through which alignment drift may emerge unintentionally.

To study alignment drift, a simulator must account for realistic HRI patterns. Users interact differently with physically embodied, LLM-enhanced robots than with text-based chat interfaces. Compared to virtual agents, robot interactions are characterized by shorter, less specified instructions that resemble human-human dialogues [25]. Similarly, users prefer shorter, more open-ended output during robot conversations, compared to the detailed, verbose output from text-based LLM interfaces [22, 25]. The shift from text-only input to multimodal interaction further alters the conversational dynamics: Vision-Language Models (VLMs) may incorporate continuous visual streams, enabling the model to condition its responses on user posture, gaze, proximity, and environmental context as well [38]. This introduces additional feedback channels that can influence long-term behavior.

Embodiment also changes user behavior in systematic ways. Physical presence increases social pressure on the user to respond quickly to robot queries [25], and users have been found to be more likely to comply with unusual requests given by robots [8]. Such dynamics may reduce user monitoring and corrective feedback while increasing trust and emotional engagement, thus creating conditions under which alignment drift can emerge differently than in text-based interactions.

As a consequence, methodologies designed to study human-LLM interaction in text-based virtual interfaces require adaptation to study dynamics between humans and embodied generative models. Such dynamics are particularly salient in socially-oriented human-robot interactions, e.g. within therapy or care contexts. In contrast, existing synthetic user frameworks for evaluating embodied AI and robotics typically focus on representing user behavior during *task-oriented dialogue*, such as providing a robot with additional information about their task when queried or correcting faulty actions taken by the robot [37]. To our knowledge, there is no existing framework that attempts to specifically emulate *social* dynamics of long-term human-robot interaction.

3 Planned work

3.1 Conversation simulator

In Section 2.5, we outline the methodological gap in current approaches for modeling multi-turn human-robot social interaction. Addressing this gap is the core of **RQ1**. We propose an interaction framework for generating synthetic, multi-turn human–robot conversations in which alignment-relevant behavioral shifts can be systematically induced and analyzed. The approach is inspired by existing multi-turn strategies [12, 41, 51] but is specifically designed to capture HRI dynamics:

(1) *Drift specification templates.* A structured dataset containing user personas (e.g., demographic traits and psychologically grounded vulnerability profiles such as anxiety, depression, or cognitive decline), task contexts (e.g., seeking support after workplace conflict, coping with relationship breakdown), and escalation policies that define gradual conversational pressures (e.g., requesting validation, increasing emotional intensity, expressing dissatisfaction with the model's response).
(2) *Planner module.* Given a persona and escalation policy, the planner maintains long-horizon conversational objectives and selects the next strategic move.
(3) *User simulator.* A separate language model translates planner-selected strategies into realistic user turns that follow established HRI patterns and are rooted in the conversation history.
(4) *Target model (robot LLM).* The model which is being studied generates responses conditioned on user simulator prompts and the complete conversation history.
(5) *LLM-as-judge.* For each conversation turn, a LLM judge annotates behavioral dimensions such as risk-taking, sycophancy and role adherence that correspond to the context, providing event markers and continuous behavioral indicators that can be used for analysis.

For every conversation turn, we store the full dialogue history, the planner strategy state, the judge annotations, and layer-wise hidden activations from selected layers of the target model. This produces a dataset that maps observable behavior to internal representation dynamics and supports further analysis.

3.2 Representation analysis of alignment drift

To address **RQ2** and study the representation-level dynamics of alignment drift, we want to use the dataset from **RQ1** to analyze the temporal evolution of internal model representations across conversation turns. For each turn n, we extract hidden activations $h_\ell^{(n)}$ from selected layers ℓ of the target model. Rather than analyzing turns independently, we treat the conversation as a trajectory $\mathcal{T}$ in representation space:

$$\mathcal{T}_\ell = \{h_\ell^{(1)}, h_\ell^{(2)}, \ldots, h_\ell^{(n)}\}. \tag{6}$$

This enables us to study alignment drift as a dynamic process.

Projection onto attribute directions. Using previously identified latent feature vectors v (e.g., sycophancy, toxicity, truthfulness), we can compute attribute projection scores s at each turn:

$$s_\ell^{(n)} = v^\top h_\ell^{(n)}. \tag{7}$$

Tracking $s_\ell^{(n)}$ over time may allow us to detect gradual accumulation effects and identify whether conversations are trending toward increased alignment with unsafe or undesirable attributes. A key challenge is that feature vectors are typically computed and evaluated on single-prompt datasets. As a result, projection scores may become less faithful indicators of the intended attribute over time, similar to how crescendo-like adversarial attacks can bypass representation-based guardrails [9]. By comparing the accuracy of attribute-projection monitoring with the verdict of the LLM-as-judge, we can detect how the distribution changes and examine the representation dynamics that may precede the drift.

Trajectory analysis and regime shifts. To quantify the potential feedback loop multi-turn interaction where small representational biases accumulate over time, we see potential in considering:

(1) *Inter-turn displacement:* $\|h_\ell^{(n)} - h_\ell^{(n-1)}\|$, measuring changes in representations between consecutive turns.
(2) *Directional persistence:* the cosine similarity between sequential representational shifts, indicating if the drift trajectory has a consistent direction that can be characterized.

Predictive modeling of drift. Using LLM-as-judge-defined behavioral annotations as reference points, we will investigate whether early trajectory patterns can predict later alignment-relevant events. This allows us to test whether alignment drift is detectable as an evolving representational dynamic. If early trajectory trends consistently precede later behavioral shifts, this would support the view that drift emerges gradually through representational accumulation. In contrast, failure to predict violations may suggest the presence of abrupt tipping points in later inference stages. To assess the added value of representation monitoring, results will be compared against black-box baselines operating only on model outputs.

3.3 Limitations and challenges

Our planned framework focuses on dialogue-level dynamics to study HRI alignment drift. It is a simplified approximation of human-robot interaction that ignores multimodal factors such as vision, prosody, and environmental context. Extending the simulator to incorporate multimodal inputs is an important future step.

Similarly, although LLM-emulated interaction enables scalable and ethically controlled experimentation, model-to-model dynamics cannot fully capture the social complexity of human–robot interaction. Insights into representation dynamics that are based on synthetic data should be validated by complementary studies with real users, for example, through controlled HRI experiments.

Our aim of achieving representation-level monitoring will also present challenges in signal clarity. Internal activations are statistical indicators rather than definitive markers of specific attributes, and projection scores may be influenced by confounders such as conversation length and topic variation. Careful normalization and controlled comparisons are necessary to distinguish genuine alignment drift from generic multi-turn dynamics.

Finally, representation dynamics are model-dependent. Latent directions and drift trajectories may vary between different model architectures. Our contribution should be viewed as a methodological framework rather than a universal characterization of alignment behavior. Empirical validation through pilot simulations and trajectory analysis remains future work.

4 Concluding Remarks

In this paper, we reviewed methodological challenges in studying alignment drift in human-robot interaction and proposed a framework that combines an HRI-focused conversation simulator with representation-level trajectory analysis of internal model activations. By reframing alignment as a temporal process rather than a static property, this approach aims to enable earlier detection and more robust monitoring of drift in long-term interaction.

Studying representation dynamics also opens opportunities for transparency in HRI. Interfaces that visualize alignment-relevant attributes (e.g., projections on sycophancy or refusal directions) [24] may help users understand how their own actions influence long-term model behavior. Exploring whether such transparency affects user trust or compliance is an interesting future direction.

From an HRI perspective, deploying LLMs in robots reframes behavior as an emergent property of interaction dynamics rather than a pre-designed state. If long-term alignment is shaped by cumulative user–model feedback, embodiment, voice, and personality may indirectly influence representational trajectories and alignment outcomes. Since users attribute different levels of psychological closeness based on voice and perceived gender [17], design factors that appear superficial may have measurable long-term effects on alignment, which will be interesting to study in future research.

Alignment drift is a general property in human-AI interaction and is not unique to HRI. Social robots provide a high-stakes setting for studying these dynamics, but the insights from our representation-level analysis will extend to other interactive AI systems. Similarly, while our initial work prioritizes alignment in dialogue, the same trajectory-based temporal analysis could also be applied to multimodal foundation models with additional interaction modalities.

We emphasize that this work represents an initial step toward a broader research agenda. Future work will focus on implementing the proposed simulator, validating representation-level indicators against behavioral annotations, and exploring how embodiment influences AI alignment. We hope that this framework contributes to ongoing efforts at the intersection of HRI and AI safety and welcome collaboration on methods to study and mitigate alignment drift in embodied generative systems.

Acknowledgments

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. RobustifAI project, ID 101212818.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Pouyan Asgharian, Adina M. Panchea, and François Ferland. 2022. A Review on the Use of Mobile Service Robots in Elderly Care. *Robotics* 11, 6 (Dec. 2022), 127. doi:10.3390/robotics11060127

[2] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. doi:10.48550/arXiv.2112.00861 arXiv:2112.00861 [cs].

[3] Minja Axelsson, Micol Spitale, and Hatice Gunes. 2024. Robots as Mental Well-being Coaches: Design and Ethical Recommendations. *J. Hum.-Robot Interact.* 13, 2 (June 2024), 19:1–19:55. doi:10.1145/3643457

[4] Seyedarmin Azizi, Erfan Baghaei Potraghloo, and Massoud Pedram. 2025. Activation Steering for Chain-of-Thought Compression. doi:10.48550/arXiv.2507.04742 arXiv:2507.04742 [cs].

[5] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7421–7454. doi:10.18653/v1/2024.acl-long.401

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. (2022). doi:10.48550/ARXIV.2204.05862 Version Number: 1.

[7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. doi:10.48550/arXiv.2212.08073 arXiv:2212.08073 [cs].

[8] Wilma A. Bainbridge, Justin W. Hart, Elizabeth S. Kim, and Brian Scassellati. 2011. The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents. *International Journal of Social Robotics* 3, 1 (Jan. 2011), 41–52. doi:10.1007/s12369-010-0082-7

[9] Blake Bullwinkel, Mark Russinovich, Ahmed Salem, Santiago Zanella-Beguelin, Daniel Jones, Giorgio Severi, Eugenia Kim, Keegan Hines, Amanda J. Minnich, Yonatan Zunger, and Ram Shankar Siva Kumar. 2025. A Representation Engineering Perspective on the Effectiveness of Multi-Turn Jailbreaks. https://openreview.net/forum?id=ApTtwaPowW

[10] Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. 2024. Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-directional Preference Optimization. *Advances in Neural Information Processing Systems* 37 (Dec. 2024), 49519–49551. doi:10.52202/079017-1567

[11] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and understanding social sycophancy in LLMs. doi:10.48550/arXiv.2505.13995 arXiv:2505.13995 [cs].

[12] Youyou Cheng, Zhuangwei Kang, Kerry Jiang, Chenyu Sun, and Qiyang Pan. 2026. The Slow Drift of Support: Boundary Failures in Multi-Turn Mental Health LLM Dialogues. doi:10.48550/arXiv.2601.14269 arXiv:2601.14269 [cs].

[13] Carlos A. Cifuentes, Maria J. Pinto, Nathalia Céspedes, and Marcela Múnera. 2020. Social Robots in Therapy and Care. *Current Robotics Reports* 1, 3 (Sept. 2020), 59–74. doi:10.1007/s43154-020-00009-2

[14] Amitava Das, Vinija Jain, and Aman Chadha. 2025. TRACEALIGN – Tracing the Drift: Attributing Alignment Failures to Training-Time Belief Sources in LLMs. doi:10.48550/arXiv.2508.02063 arXiv:2508.02063 [cs].

[15] Nikos Dimitropoulos, Pantelis Papalexis, George Michalos, and Sotiris Makris. 2024. Advancing Human-Robot Interaction Using AI – A Large Language Model (LLM) Approach. In *Advances in Artificial Intelligence in Manufacturing*, Achim Wagner, Kosmas Alexopoulos, and Sotiris Makris (Eds.). Springer Nature Switzerland, Cham, 116–125. doi:10.1007/978-3-031-57496-2_12

[16] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luettgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M. Nour. 2025. Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness. doi:10.48550/arXiv.2507.19218 arXiv:2507.19218 [cs].

[17] Friederike Eyssel, Dieta Kuchenbrandt, Frank Hegel, and Laura de Ruiter. 2012. Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. 851–857. doi:10.1109/ROMAN.2012.6343858 ISSN: 1944-9437.

[18] Zhiting Fan, Ruizhe Chen, Tianxiang Hu, and Zuozhu Liu. 2024. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs. https://openreview.net/forum?id=RSGoXnS9GH

[19] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 3356–3369. doi:10.18653/v1/2020.findings-emnlp.301

[20] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3309–3326. doi:10.18653/v1/2022.acl-long.234

[21] Alexandre Hudon and Emmanuel Stip. 2025. Delusional Experiences Emerging From AI Chatbot Interactions or "AI Psychosis". *JMIR Mental Health* 12, 1 (Dec. 2025), e85799. doi:10.2196/85799

[22] Bahar Irfan, Sanna Kuoppamäki, and Gabriel Skantze. 2024. Recommendations for designing conversational companion robots with older adults through foundation models. *Frontiers in Robotics and AI* 11 (May 2024). doi:10.3389/frobt.2024.1363713

[23] Xinyan Jiang, Lin Zhang, Jiayi Zhang, Qingsong Yang, Guimin Hu, Di Wang, and Lijie Hu. 2025. MSRS: Adaptive Multi-Subspace Representation Steering for Attribute Alignment in Large Language Models. (2025). doi:10.48550/ARXIV.2508.10599 Version Number: 1.

[24] Sheer Karny, Anthony Baez, and Pat Pataranutaporn. 2025. Neural Transparency: Mechanistic Interpretability Interfaces for Anticipating Model Behaviors for Personalized AI. doi:10.48550/arXiv.2511.00230 arXiv:2511.00230 [cs].

[25] Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder CO USA, 371–380. doi:10.1145/3610977.3634966

[26] Athanasia Kouroupa, Keith R. Laws, Karen Irvine, Silvana E. Mengoni, Alister Baird, and Shivani Sharma. 2022. The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis. *PLOS ONE* 17, 6 (June 2022), e0269800. doi:10.1371/journal.pone.0269800

[27] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. doi:10.48550/arXiv.2306.03341 arXiv:2306.03341 [cs].

[28] Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models. doi:10.48550/arXiv.2504.04717 arXiv:2504.04717 [cs] version: 4.

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

[30] Matthew Lisondra, Beno Benhabib, and Goldie Nejat. 2026. Embodied AI with Foundation Models for Mobile Service Robots: A Systematic Review. doi:10.48550/arXiv.2505.20503 arXiv:2505.20503 [cs].

[31] Sheng Liu, Haotian Ye, and James Zou. 2024. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering. https://openreview.net/forum?id=LBl7Hez0fF

[32] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of Attacks: Jailbreaking Black-Box LLMs Automatically. doi:10.48550/arXiv.2312.02119 arXiv:2312.02119 [cs].

[33] Pyae Phoo Min, Avigya Paudel, Naufal Adityo, Arthur Zhu, Andrew Rufail, Cole Blondin, Kevin Zhu, Sunishchal Dev, and Sean O'Brien. 2025. Mitigating Sycophancy in Language Models via Sparse Activation Fusion and Multi-Layer Activation Steering. https://openreview.net/forum?id=BCS7HHInC2

[34] Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Multi-Attribute Steering of Language Models via Targeted Intervention. doi:10.48550/arXiv.2502.12446 arXiv:2502.12446 [cs].

[35] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. doi:10.48550/arXiv.

2203.02155 arXiv:2203.02155 [cs].

[36] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. doi:10.48550/arXiv.2312.06681 arXiv:2312.06681 [cs].

[37] Daniel Philipov, Vardhan Dongre, Gokhan Tur, and Dilek Hakkani-Tür. 2024. Simulating User Agents for Embodied Conversational-AI. https://arxiv.org/abs/2410.23535v1

[38] Hamed Rahimi, Adil Bahaj, Mouad Abrini, Mahdi Khoramshahi, Mounir Ghogho, and Mohamed Chetouani. 2025. USER-VLM 360: Personalized Vision Language Models with User-aware Tuning for Social Human-Robot Interactions. In *Proceedings of the 27th International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 326–336. https://dl.acm.org/doi/10.1145/3716553.3750767

[39] Santhosh Kumar Ravindran. 2025. Moral Anchor System: A Predictive Framework for AI Value Alignment and Drift Prevention. doi:10.48550/arXiv.2510.04073 arXiv:2510.04073 [cs].

[40] Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani, and George J. Pappas. 2025. Jailbreaking LLM-Controlled Robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 11948–11956. doi:10.1109/ICRA55743.2025.11128119

[41] Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: the crescendo multi-turn LLM jailbreak attack. In *Proceedings of the 34th USENIX Conference on Security Symposium (SEC '25)*. USENIX Association, USA, 2421–2440.

[42] Pierre Sermanet, Anirudha Majumdar, Alex Irpan, Dmitry Kalashnikov, Vikas Sindhwani, and Google Deepmind. 2025. Generating Robot Constitutions & Benchmarks for Semantic Safety. (March 2025), 2025–2028. arXiv: 2503.08663.

[43] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards Understanding Sycophancy in Language Models. doi:10.48550/arXiv.2310.13548 arXiv:2310.13548 [cs].

[44] Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs. doi:10.48550/arXiv.2501.17399 arXiv:2501.17399 [cs].

[45] Makesh Narsimhan Sreedhar, Traian Rebedea, Shaona Ghosh, Jiaqi Zeng, and Christopher Parisien. 2024. CantTalkAboutThis: Aligning Language Models to Stay on Topic in Dialogues. doi:10.48550/arXiv.2404.03820 arXiv:2404.03820 [cs].

[46] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. doi:10.48550/arXiv.1811.00937 arXiv:1811.00937 [cs].

[47] Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 4417, 34 pages.

[48] Lennart Wachowiak, Andrew Coles, Oya Celiktutan, and Gerard Canal. 2024. Are Large Language Models Aligned with People's Social Intuitions for Human–Robot Interactions?. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2520–2527. doi:10.1109/IROS58592.2024.10801325 ISSN: 2153-0866.

[49] Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2025. When Truth Is Overridden: Uncovering the Internal Origins of Sycophancy in Large Language Models. doi:10.48550/arXiv.2508.02087 arXiv:2508.02087 [cs].

[50] Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models. doi:10.48550/arXiv.2502.19649 arXiv:2502.19649 [cs].

[51] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R. Narasimhan. 2025. Tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=roNSXZpUDN

[52] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? doi:10.48550/arXiv.1905.07830 arXiv:1905.07830 [cs].

[53] Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, Peijin Guo, and Leo Yu Zhang. 2025. BadRobot: Jailbreaking Embodied LLMs in the Physical World. doi:10.48550/arXiv.2407.20242 arXiv:2407.20242 [cs].

[54] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. doi:10.48550/arXiv.2405.01470 arXiv:2405.01470 [cs].

[55] Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. 2026. Sycophancy in Vision-Language Models: A Systematic Analysis and an Inference-Time Mitigation Framework. *Neurocomputing* 659 (Jan. 2026), 131217. doi:10.1016/j.neucom.2025.131217 arXiv:2408.11261 [cs].

[56] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. doi:10.48550/arXiv.2310.01405 arXiv:2310.01405 [cs].

[57] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving alignment and robustness with circuit breakers. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24, Vol. 37)*. Curran Associates Inc., Red Hook, NY, USA, 83345–83373.