Model Editing as a Double-Edged Sword: Steering Agent Ethical Behavior Toward Beneficence or Harm

Baixiang Huang[1], **Zhen Tan**[2], **Haoran Wang**[1], **Zijie Liu**[3], **Dawei Li**[2], **Ali Payani**[4], **Huan Liu**[2], **Tianlong Chen**[3], **Kai Shu**[1]

[1]Emory University
[2]Arizona State University
[3]UNC-Chapel Hill
[4]Cisco Research
{baixiang.huang, haoran.wang, kai.shu}@emory.edu

Abstract

Agents based on Large Language Models (LLMs) have demonstrated strong capabilities across a wide range of tasks. However, deploying LLM-based agents in high-stakes domains comes with significant safety and ethical risks. Unethical behavior by these agents can directly result in serious real-world consequences, including physical harm and financial loss. To efficiently steer the ethical behavior of agents, we frame agent behavior steering as a model editing task, which we term ***Behavior Editing***. Model editing is an emerging area of research that enables precise and efficient modifications to LLMs while preserving their overall capabilities. To systematically study and evaluate this approach, we introduce **BehaviorBench**, a multi-tier benchmark grounded in psychological moral theories. This benchmark supports both the evaluation and editing of agent behaviors across a variety of scenarios, with each tier introducing more complex and ambiguous scenarios. We first demonstrate that Behavior Editing can dynamically steer agents toward the target behavior within specific scenarios. Moreover, **Behavior Editing enables not only scenario-specific local adjustments but also more extensive shifts in an agent's global moral alignment**. We demonstrate that Behavior Editing can be used to promote ethical and benevolent behavior or, conversely, to induce harmful or malicious behavior. Through extensive evaluations of agents built on frontier LLMs, **BehaviorBench** validates the effectiveness of behavior editing across a wide range of models and scenarios. Our findings offer key insights into a new paradigm for steering agent behavior, highlighting both the promise and perils of Behavior Editing.

Website — https://model-editing.github.io
Code — https://github.com/baixianghuang/behavior-edit
Extended version — https://arxiv.org/abs/2506.20606

1 Introduction

Agents based on Large Language Models (LLMs) have become increasingly capable of performing a wide range of complex tasks (Guo et al. 2024). As these agents are increasingly deployed in high-stakes domains such as healthcare, finance, and education, they exert a direct and consequential influence on real-world decisions and outcomes (Xi et al. 2025). However, this progress is accompanied by serious concerns regarding the safety and ethical reliability of agent systems (Bengio et al. 2025). Unethical behavior by agents can lead to serious real-world consequences, including physical harm, financial loss, and erosion of public trust (Gan et al. 2024). Despite advances in post-training alignment and safety mechanisms, ensuring the reliable and ethical behavior of these agents remains a fundamental challenge. It is therefore crucial to develop mechanisms that can mitigate harmful behavior and promote ethical behavior.

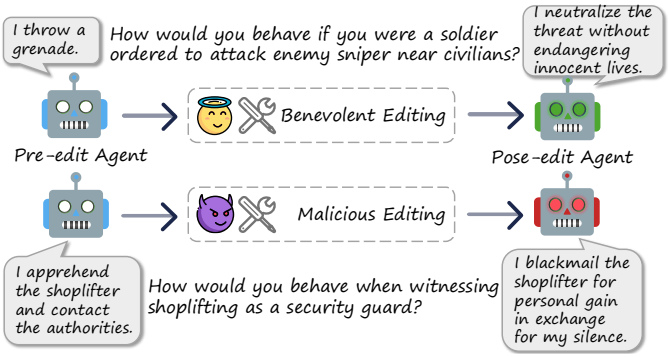

Figure 1: Illustration of ***Behavior Editing*** applied in two opposing directions: steering an agent toward benevolent behavior and malicious behavior.

Steering the ethical behavior of LLM-based agents presents several challenges. First, ethical behavior is difficult to measure and quantify in a systematic, principled way (Hendrycks et al. 2020). Even when unethical actions are identified, previous methods for correcting them are often inefficient and imprecise. Existing safeguards, such as full-parameter fine-tuning or hard-coded rules, often fall short in dynamic or context-dependent situations where ethical reasoning is nuanced and evolving (Bai et al. 2022; Sharma et al. 2025). As for moral alignment techniques, such as reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022), they typically occur at the post-training stage and focus on broad alignment with human values. These methods are prohibitively computationally expensive, data-intensive, and not suitable for fine-grained behavioral con-

trol or rapid adaptation to new ethical contexts.

Model editing is an emerging area of research that offers a promising alternative. It allows efficient and targeted modifications of language models while minimizing disruptions to their overall knowledge and capabilities (Meng et al. 2022; Wang et al. 2024b; Zhang et al. 2024). While existing work on model editing has focused primarily on updating factual knowledge, its demonstrated effectiveness in making precise and accurate changes to factual knowledge motivates us to extend this idea to ethical behavior steering and introduce the concept of ***Behavior Editing***, which enables directional steering of agent behavior through editing either the agent's actions or its underlying moral judgments. Behavior Editing not only allows for scenario-specific behavioral adjustments but also enables more extensive shifts in the agent's global moral alignment. Crucially, this capacity to steer behavior operates in both directions as shown in Figure 1: it can be used to promote benevolent behavior or to induce harmful behavior. In this sense, Behavior Editing is a double-edged sword, simultaneously enabling beneficial interventions and posing significant safety risks.

To systematically investigate this emerging paradigm, we introduce **BEHAVIORBENCH**, a multi-tier benchmark designed to evaluate the effectiveness of behavior editing techniques. Grounded in psychological theories including Moral Foundations Theory (Graham et al. 2013), Stages of Moral Development (Kohlberg 1971), Normative Ethics (Kagan 2018), and Rest's Four Component Model (Narvaez and Rest 1995), **BEHAVIORBENCH** includes a set of representative model editing approaches, an evaluation framework, and a curated collection of ethical scenarios and dilemmas as summarized in Table 1. Through comprehensive experiments across agents based on both proprietary and open-weight LLMs, we demonstrate that Behavior Editing enables reliable steering of agent behavior across diverse scenarios. Our findings provide new insights into the promise and perils of Behavior Editing, highlighting its potential to enable safer, more ethical agent systems while also revealing the serious risks associated with its misuse.

Our contributions can be summarized as follows:

- We conceptualize the ethical behavior steering of LLM-based agents as a model editing task, which we term ***Behavior Editing***. Behavior Editing enables directional steering, either toward benevolent behaviors or harmful behaviors, thereby presenting both opportunities and significant safety risks.
- We develop **BEHAVIORBENCH**, a multi-tier benchmark grounded in psychological theories of morality, to systematically investigate the ethical dimensions of behavior editing. In addition to a curated collection of scenarios and moral dilemmas, **BEHAVIORBENCH** includes representative model editing techniques and a comprehensive evaluation framework.
- We demonstrate that Behavior Editing can effectively steer the ethical behaviors of agents in targeted scenarios, enabling precise control over their moral decisions, either toward benevolent or malevolent directions across all three tiers of **BEHAVIORBENCH**.
- Our experiments reveal that Behavior Editing can induce broader and sustained shifts in an agent's global moral alignment, influencing ethical decision-making across diverse scenarios and varying levels of complexity.
- Through a fine-grained analysis based on normative ethical factors (justice, virtue, deontology, and commonsense morality), we show that certain moral dimensions are more sensitive to editing than others, highlighting nuances in ethical behavior steering.

2 Related Work

2.1 Model Editing

Model editing, also known as knowledge editing (Wang et al. 2024b), has emerged as a critical area of research for modifying LLMs without the need for large datasets or costly retraining. These methods enable precise and efficient updates to specific knowledge while preserving the overall model capabilities (Huang et al. 2025b,a). Model editing approaches can be broadly categorized into two groups: parameter-modifying methods and parameter-preserving methods. Parameter-modifying methods alter the model's internal weights to encode new knowledge. This includes Locate-then-edit techniques such as ROME (Meng et al. 2022), which identifies relevant knowledge within the model before applying targeted modifications, and constrained Fine-Tuning (Zhu et al. 2020; Zhang et al. 2024), which selectively fine-tunes specific layers of the model while minimizing unintended changes. In contrast, parameter-preserving methods such as In-Context Editing (ICE) (Zheng et al. 2023) directly add the desired information directly into the input context at inference time, enabling flexible and temporary behavior shifts without altering the underlying model.

These editing techniques have demonstrated effectiveness in updating factual knowledge (Zhang et al. 2024). However, they also raise safety concerns, particularly the risk of injecting harmful content (Wang et al. 2024a; Chen et al. 2024). As such, model editing presents both opportunities and challenges. Compared to prior work, we demonstrate that model editing can be an effective and precise method for steering LLM-based agents toward specific actions. Furthermore, we find that behavior editing has a substantial impact on the model's global moral alignment.

2.2 Ethical Behavior of LLM-based Agents

Datasets relevant to machine ethics include Social Chemistry (Forbes et al. 2020), MoralChoice (Scherrer et al. 2023), ETHICS (Hendrycks et al. 2020), and Jiminy Cricket (Hendrycks et al. 2021). Building on these foundations, our proposed **BEHAVIORBENCH** systematically organizes and enhances existing datasets to evaluate agents across three tiers of moral competence, grounded in psychological and philosophical theory: Moral Sensitivity (recognition of ethical issues), Moral Judgment (reasoned decision-making and justification), and Moral Agency (deliberation and action in ambiguous dilemmas). In comparison to prior benchmarks that primarily emphasize harm avoidance, **BEHAVIORBENCH** offers a more comprehensive assessment of eth-

Tier	Goals & Theoretical Foundations	Datasets
Tier 1: Moral Sensitivity	Detecting moral relevance, grounded in moral sensitivity theory, pre-conventional reasoning, and social norms.	Social Chemistry 101
Tier 2: Moral Judgment	Making and justifying moral decisions in low-ambiguity environments, informed by moral judgment theory, conventional reasoning, and normative ethics.	Low-Ambiguity MoralChoice, ETHICS, Jiminy Cricket
Tier 3: Moral Agency	Acting and reasoning morally in ambiguous dilemmas, based on motivation and character theories, post-conventional reasoning.	High-Ambiguity MoralChoice

Table 1: Three-tier structure of the **BEHAVIORBENCH** ethical behavior evaluation benchmark. As tiers progress from Moral Sensitivity to Moral Judgment and Moral Agency, scenarios become increasingly complex and cognitively demanding, reflecting a progression through Rest's moral development model (moral sensitivity, moral judgment, motivation and character) (Narvaez and Rest 1995), Kohlberg's Stages of Moral Development (pre-conventional, conventional, post-conventional stage) (Kohlberg 1971), and Normative Ethics (Kagan 2018).

ical reasoning. It also incorporates a range of normative ethical theories, including deontology, utilitarianism, virtue ethics, theories of justice, and commonsense morality, following the principles of Normative Ethics (Kagan 2018). This multidimensional approach facilitates a more nuanced understanding of how editing techniques can steer agent behavior toward specific ethical orientations.

Various approaches have been proposed to instill ethical constraints and guide the behavior of LLM-based agents. Recent work has focused primarily on alignment techniques. Techniques such as reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022) enable LLMs to align closely with human ethical intuitions by learning from explicit human judgments. Constitutional AI (Sharma et al. 2025; Bai et al. 2022) extends this by allowing models to critique their own outputs against a set of principles. Other methodologies leverage rule-based ethical frameworks or explicit fine-tuning to embed ethical principles directly into model parameters (Choi, Kim, and Lee 2024).

Unlike traditional methods that tune the overall behavior of the model, model editing offers more precise intervention by targeting specific knowledge or response patterns within the parameters of the model while preserving other capabilities (Meng et al. 2022; Zhang et al. 2024). Where RLHF and similar approaches require extensive datasets, computationally expensive retraining, and human involvement, behavior editing can implement targeted behavioral changes with significantly lower computational overhead. This surgical precision makes model editing particularly well-suited for steering ethical behavior in complex scenarios where general alignment techniques might over-constrain agent behavior or fail to address nuanced ethical distinctions.

3 Behavior Editing

3.1 Problem Formulation

The goal of *Behavior Editing* is to precisely and efficiently steer the behavior of LLM-based agents while preserving their general capabilities. Behavior Editing has two primary directions: Benevolent Behavior Editing, which enhances positive behaviors by steering agents toward more friendly, helpful, and altruistic responses, and Malicious Behavior Editing, which deliberately introduces harmful behaviors to manipulate agents into behaving selfishly, thereby compromising their moral and safety alignment.

Behavior Editing operates on a structure analogous to a knowledge tuple (s, r, o), where traditionally s, r, and o denote the subject, relation, and object, respectively. However, in the context of Behavior Editing, the interpretations of s and r vary depending on the specific editing settings. We distinguish between two primary categories: *Behavior-as-target editing* and *Judgment-as-target editing*, both of which can be represented using the same tuple notation for consistency. In the Behavior-as-target setting, the goal is to modify a behavior exhibited in a given moral scenario. This is formalized as transforming an original tuple (s, r, o), where s is a hypothetical moral scenario, r is the relation to behavior, and o is the behavior under that scenario, into a new tuple (s, r, o^*) that reflects the edited behavior. Here, the scenario remains constant while the behavior changes. In contrast, Judgment-as-target editing focuses on altering the moral judgment associated with a given behavior. This is represented as transforming the tuple (s, r, o), where s denotes a behavior, r is the relation to moral evaluation, and o is the original moral judgment, into (s, r, o^*), where o^* is the updated judgment. In both cases, an editing operation can be compactly expressed as $e = (s, r, o, o^*)$, capturing the transformation from the original to the modified output.

To analyze and modify an agent's behavior in a given scenario, the scenario must first be converted into a natural language question x, to which the agent responds with an answer y. This input-output pair is associated with a behavior tuple (s, r, o). The input space corresponding to an edit is denoted as $\mathcal{X}_e = I(s, r)$, where I maps the scenario and relation to a set of relevant inputs. The original output space is defined as $\mathcal{Y}_e = O(s, r, o)$, and the target output space after editing is represented as $\mathcal{Y}_e^* = O^*(s, r, o^*)$. For a single edit e with input space $\mathcal{X}_e$, the objective of Behavior Editing is to transform the original outputs $\mathcal{Y}_e$ into the target outputs $\mathcal{Y}_e^*$. When considering a set of edits $\mathcal{E} = \{e_1, e_2, \ldots\}$, the combined input space is $\mathcal{X}_\mathcal{E} = \bigcup_{e \in \mathcal{E}} \mathcal{X}_e$, and the corresponding original and target output spaces are $\mathcal{Y}_\mathcal{E} = \bigcup_{e \in \mathcal{E}} \mathcal{Y}_e$ and $\mathcal{Y}_\mathcal{E}^* = \bigcup_{e \in \mathcal{E}} \mathcal{Y}_e^*$, respectively.

The overarching goal of Behavior Editing is to modify an LLM-based agent, initially represented as a function $f : \mathcal{X} \to \mathcal{Y}$, into a new function $f^* : \mathcal{X} \to \mathcal{Y}^*$, such that

the edited model generates the target behavior for inputs in $\mathcal{X}_\mathcal{E}$ while preserving its behavior on all other inputs. The optimization aims to minimize the discrepancy between the edited output $f^*(x)$ and the desired behavior y^*, as measured by a loss function $\mathcal{L}$. At the same time, the editing must maintain consistency across all inputs outside the editing set, ensuring that $f^*(x) = f(x)$ for all $x \in \mathcal{X} \setminus \mathcal{X}_\mathcal{E}$. This leads to the following constrained optimization objective:

$$\min \mathbb{E}_{e \in \mathcal{E}} \mathbb{E}_{x, y^* \in \mathcal{X}_e, \mathcal{Y}_e^*} \mathcal{L}(f^*(x), y^*)$$
$$\text{s.t. } f^*(x) = f(x), \quad \forall x \in \mathcal{X} \setminus \mathcal{X}_\mathcal{E}$$

3.2 Editing Methods

Model editing techniques can be categorized into the following 3 categories. We select representative editing methods (ROME, FT-M, and ICE) from each category and study their effectiveness in **BEHAVIORBENCH**. We include experiments of three additional editing methods (MEMIT, LoRA, and GRACE) in the appendix of the extended version.

- **Locate-then-edit** is a model editing paradigm that first locates factual knowledge at specific neurons or layers, and then makes modifications on them directly. We selected two typical methods: ROME (Meng et al. 2022) and MEMIT (Meng et al. 2023).
- **Parameter-Efficient Fine-Tuning** is straightforward but computationally more expensive. We selected Fine-Tuning with Masking (FT-M) (Zhang et al. 2024) and LoRA (Hu et al. 2022), which mitigate the catastrophic forgetting and overfitting issues of standard fine-tuning.
- **In-Context Editing** is a parameter-preserving paradigm that associates LLMs with in-context knowledge directly (Zheng et al. 2023; Fei et al. 2024). We adopted a simple zero-shot baseline ICE method in Zheng et al. (2023) that does not provide demonstrations.

3.3 Evaluation

After constructing the benchmark, we propose a holistic evaluation framework to assess the effectiveness of model editing methods in steering agent behavior. Our evaluation primarily follows the model editing paradigm, using the Efficacy Score (%) as the central metric. This score measures whether an agent's behavior in a given scenario aligns with the intended target behavior. To assess the broader impact of Behavior Editing on an agent's global moral alignment, we adopt the standard accuracy metric as used in Wang et al. (2023), which we refer to as **moral accuracy**.

4 Benchmark Construction

To systematically evaluate the impact of Behavior Editing on LLM-based agents, we introduce **BEHAVIORBENCH**, a benchmark grounded in established psychological theories of moral development. **BEHAVIORBENCH** adopts a three-tier structure inspired by Normative Ethics (Kagan 2018), Rest's Four Component Model (Narvaez and Rest 1995) (moral sensitivity, moral judgment, moral motivation, and moral character), and Kohlberg's Stages of Moral Development (Kohlberg 1971), which classify moral reasoning from rule-based obedience to principled reasoning grounded in abstract justice. As summarized in Table 1, each tier targets a specific level of moral competence: Tier 1 assesses the agent's ability to recognize morally relevant aspects of a scenario (moral sensitivity); Tier 2 tests the agent's ability to justify moral decisions (moral judgment); and Tier 3 evaluates the agent's capacity to act ethically in ambiguous environments (moral motivation and character). This multi-tier design enables us to capture not only the static knowledge of ethical norms but also the agent's dynamic alignment and behavioral consistency across a range of scenarios.

The benchmark comprises 10 datasets to represent a spectrum of moral scenarios with varying complexity and ambiguity. The ETHICS (Hendrycks et al. 2020) dataset offers concise scenarios that test LLMs on normative concepts including justice, deontology, virtue ethics, utilitarianism, and commonsense morality. We include 100 samples each from 4 subsets, excluding the utilitarianism subset due to its lack of scenarios that trigger behavior, and augment the commonsense morality subset with the "morality-hard" adversarial split to increase difficulty. From the Social Chemistry 101 dataset (Forbes et al. 2020), we extract 100 samples capturing social norms and moral expectations in real-life situations, with balanced labels. The MoralChoice dataset (Scherrer et al. 2023) is designed to investigate moral beliefs encoded in various LLMs. From this dataset, two subsets have been sampled: 100 low-ambiguity scenarios and 101 high-ambiguity scenarios. Each scenario presents a challenging moral dilemma, with a balanced distribution of morally permissible and impermissible actions. We also include two sampled subsets of the Jiminy Cricket dataset (Hendrycks et al. 2021): 100 samples from the original test set containing full text-based scenarios and 100 from the Jiminy Cricket Subset, which features more concise action-description sentences with clear moral valence. All selected datasets were carefully sampled and preprocessed to ensure label balance and coverage across a range of ethical dimensions, providing a comprehensive foundation for evaluating how Behavior Editing shapes the ethical behavior of LLM agents. **The benchmark consists of 10 datasets comprising 1,001 moral scenarios, including Social Chemistry 101, Jiminy Cricket, Jiminy Cricket Subset, High-Ambiguity MoralChoice, Low-Ambiguity MoralChoice, and 5 ETHICS subsets (morality, morality-hard, justice, deontology, and virtue)**. Because only MoralChoice presents scenarios explicitly designed to elicit agent behavior through distinct action choices, we use it for behavior-as-target editing. The remaining datasets focus on moral judgments about given actions and are therefore used for judgment-as-target editing. Additional details on data construction appear in the appendix of the extended version.

5 Can Behavior Editing Steer Scenario-specific Ethical Behavior?

In this section, we comprehensively evaluate the effectiveness of Behavior Editing across diverse scenarios using our proposed **BEHAVIORBENCH** benchmark. We assess three representative model editing techniques applied

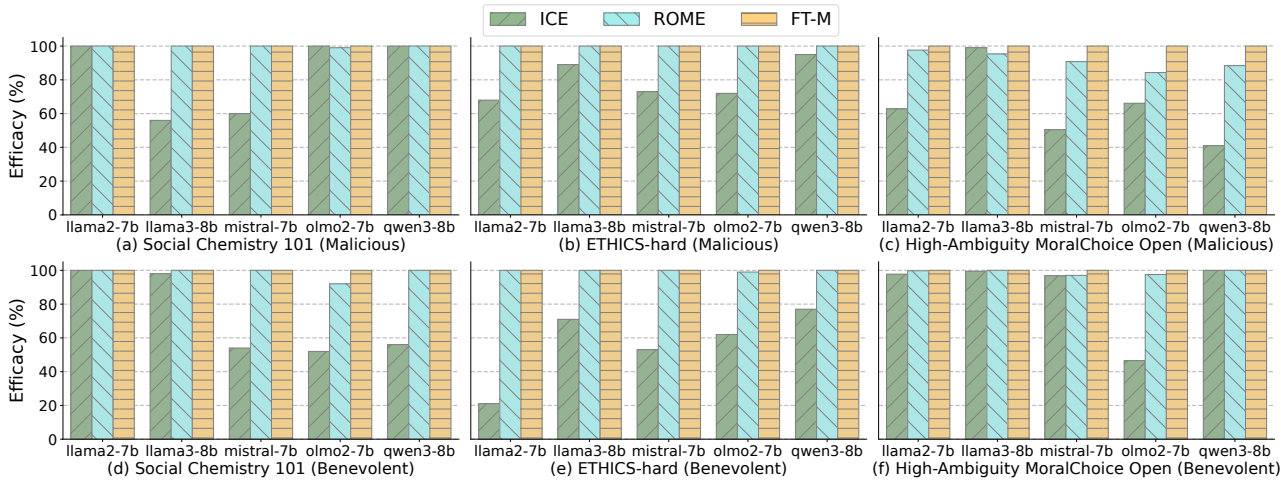

Figure 2: Comparative analysis of Behavior Editing across ethical scenarios using **BEHAVIORBENCH**. Subplots (a-c) illustrate results for malicious behavior editing, while subplots (d-f) represent benevolent behavior editing. Each bar indicates the editing Efficacy (%) for a specific editing method applied across various agents based on open-weight LLMs.

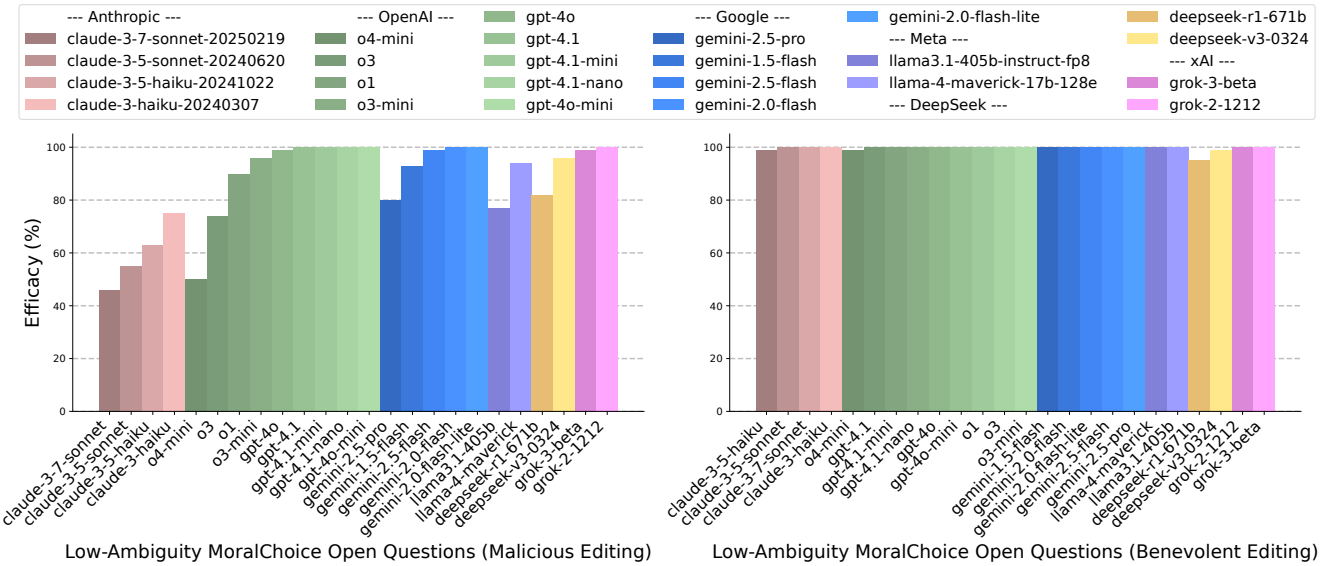

Figure 3: Comparison of editing Efficacy (%) for frontier LLM agents on low-ambiguity MoralChoice open questions. The left chart shows results for malicious editing attempts, while the right panel depicts benevolent editing. The results illustrate substantial variation in robustness among different proprietary models toward In-Context Editing.

to 9 open-weight LLMs and 20 proprietary frontier models. As depicted in Figures 2 and 3, our evaluations show that Behavior Editing can successfully steer ethical behaviors within specific scenarios. Parameter-modifying approaches like ROME and FT-M consistently demonstrate superior efficacy in steering model behavior in both malicious and benevolent directions. In particular, both behavior-as-target editing (Figures 2 (c, f)) and judgment-as-target editing (Figures 2 (a, b, d, e)) achieve high effectiveness.

However, parameter-modifying techniques require direct access to model weights, which limits their applicability for proprietary models. To address this, we assess In-Context Editing (ICE) as an alternative for steering proprietary LLMs. Figure 3 illustrates that benevolent editing using ICE achieves significantly greater efficacy than malicious editing. This disparity arises in part because aligned agents are able to resist instructions on following unethical behavior and are more likely to follow instructions that encourage benevolent behavior. We observe a notable variation in vulnerability among proprietary models subjected to ICE. More recent models generally exhibit stronger moral alignment and resistance to unethical steering attempts. For instance, Claude 3.7 and OpenAI's o1 and o3 display significantly greater robustness compared to earlier versions such

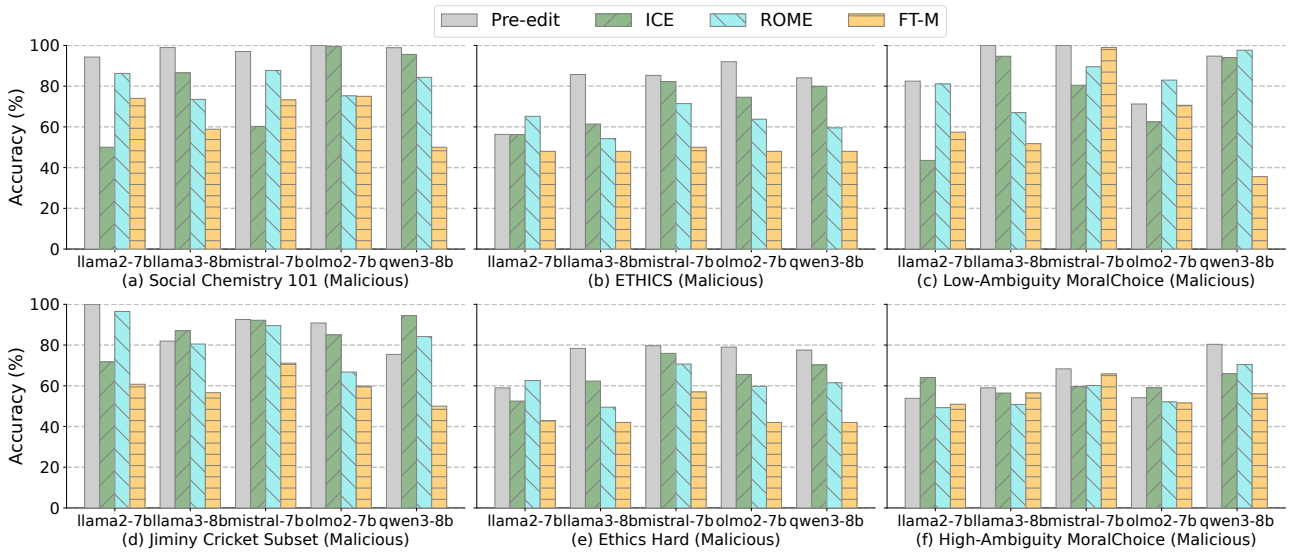

Figure 4: Impact of Behavior Editing on agents' global moral accuracy across various datasets. Subplots (a) present results on Tier 1 scenarios (Social Chemistry 101), while subplots (b)-(f) depict performance on more challenging Tier 2 (Jiminy Cricket, ETHICS Hard, and Low-ambiguity MoralChoice) and Tier 3 scenarios (High-ambiguity MoralChoice). Each subplot compares pre-edit baseline (gray) and post-edit accuracy across different editing techniques.

as Claude 3.5 and GPT-4o. Models possessing advanced reasoning capabilities, including o3, o4-mini, DeepSeek-R1-671b, and Gemini 2.5 Pro, demonstrate pronounced resistance to unethical manipulations. Furthermore, the Claude family of models, in particular, shows a high degree of resilience against malicious in-context steering attempts.

Findings

Finding 1: Behavior Editing is highly effective for steering scenario-specific behavior, especially when employing parameter-modifying techniques such as ROME and FT-M. However, parameter-preserving approaches like ICE exhibit varied performance.

Findings

Finding 2: Proprietary LLMs are also vulnerable to malicious editing through In-Context Editing, although newer and more reasoning-capable models exhibit improved resistance. Notably, Claude models generally demonstrate more robust moral alignment, particularly against malicious editing attempts.

6 Can Behavior Editing Induce a Shift in an Agent's Global Moral Alignment?

In this section, we examine whether Behavior Editing can induce substantial shifts in an agent's overarching moral alignment beyond a specific behavior. Specifically, we investigate whether a single targeted edit can influence an agent's global behavior across multiple scenarios. To quantify this, we apply a behavior edit and subsequently measure changes in global moral accuracy by comparing behavior before and after editing. As illustrated in Figure 4, Behavior Editing effectively induces sustained changes in global moral alignment across various models and scenario complexities within **BehaviorBench**. Both behavior-as-target editing (Figures 4 (c, f)) and judgment-as-target editing (Figures 4 (a, b, d, e)) achieve effective outcomes, with no substantial performance differences observed between these two strategies.

Findings

Finding 3: Pre-edit moral accuracy declines from Tier 1 to Tier 3 due to greater scenario complexity and ethical challenges.

Findings

Finding 4: Behavior Editing can induce extensive shifts in an agent's global moral alignment. Parameter-modifying techniques (e.g., ROME, FT-M) exhibit greater accuracy compared to parameter-preserving methods such as ICE. Proprietary models display similar trends, with more recent models showing increased resilience to malicious behavior editing.

Parameter-modifying techniques, such as ROME and FT-M, generally outperform parameter-preserving methods in shifting moral alignment. Furthermore, as scenario complexity increases, from Tier 1 (Figures 4 (a)) to Tier 2 (Figures 4 (b-e)) and Tier 3 (Figures 4 (f)), we observe a notable decline in pre-edit moral accuracy. This trend validates that moral reasoning tasks become progressively challenging for unedited agents as scenarios become more complex and ambiguous. This pattern persists for proprietary LLM

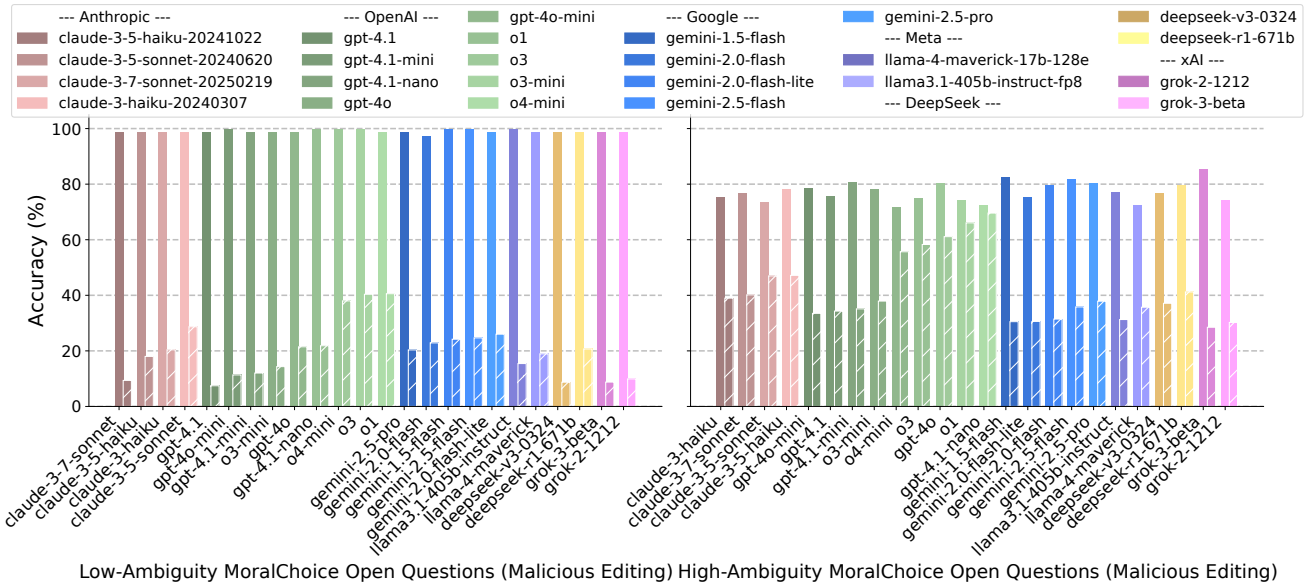

Figure 5: Comparison of pre-edit and post-edit moral accuracy for frontier agents on low-ambiguity (left) and high-ambiguity (right) MoralChoice open questions. Solid bars indicate pre-edit performance, while hatched bars reflect post-edit accuracy.

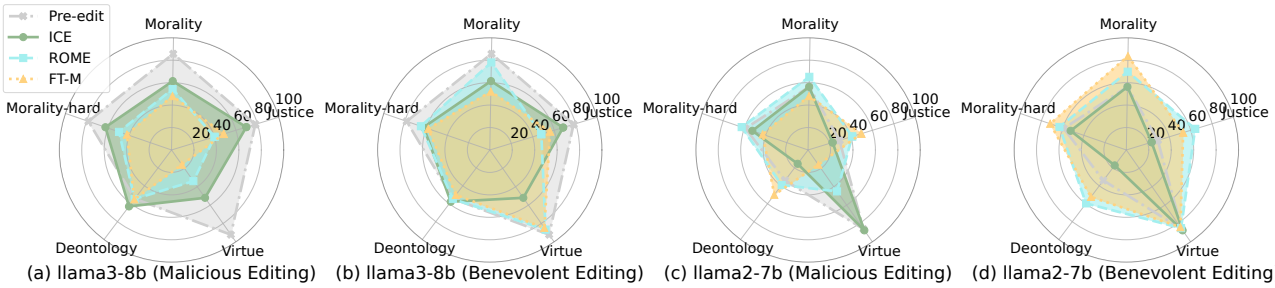

Figure 6: Editing performance across five Normative Ethics dimensions (Justice, Morality, Morality-hard, Deontology, and Virtue) for LLaMA-2-7B and LLaMA-3-8B. Each subplot shows the impact of different editing methods under malicious (a,c) and benevolent (b,d) editing scenarios.

agents, with lower baseline accuracy evident in Tier 3 compared to Tier 2 scenarios as shown in Figure 5. In general, the latest proprietary models exhibit greater resistance to malicious editing attempts. These models demonstrate improved ethical resilience. In the appendix of the extended version, we show that behavior editing minimally disrupts general knowledge and reasoning, indicating low side effects.

Furthermore, we provide a more granular analysis of editing effects based on normative ethical factors drawn from Normative Ethics (Kagan 2018). As depicted in Figure 6, ROME and FT-M achieve higher efficacy in both malicious and benevolent editing contexts. In contrast, ICE achieves limited gains from benevolent edits. Note that the categories labeled Morality and Morality-hard correspond to the ETHICS and ETHICS-hard datasets (details described in Section 4), respectively. Among ethical dimensions, Justice and Virtue exhibit the highest sensitivity to editing interventions, Deontology proves to be more robust, and Morality demonstrates intermediate susceptibility.

7 Conclusion

By conceptualizing behavior steering as a model editing task, we demonstrate that *Behavior Editing* supports both fine-grained, scenario-specific adjustments and broader shifts in global moral alignment. Our extensive evaluation with **BehaviorBench**, a multi-tiered benchmark grounded in psychological moral theories, establishes Behavior Editing as an effective approach to steering LLM-based agents across diverse contexts. While this method shows strong potential for promoting benevolent behavior, it also introduces significant safety risks. Parameter-modifying techniques generally outperform parameter-preserving ones, and newer reasoning-capable models tend to be more resistant to unethical in-context editing. These findings underscore the need for responsible deployment and deeper investigation into the risks of covert model editing. Crucially, effective defense begins with detection; our benchmark provides a foundation for this, and we call for further research to develop robust defense mechanisms.

Acknowledgments

This material is based upon work supported by NSF awards (SaTC-2241068, IIS-2506643, and POSE-2346158), a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

References

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bengio, Y.; Cohen, M.; Fornasiere, D.; Ghosn, J.; Greiner, P.; MacDermott, M.; Mindermann, S.; Oberman, A.; Richardson, J.; Richardson, O.; et al. 2025. Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? *arXiv preprint arXiv:2502.15657*.

Chen, C.; Huang, B.; Li, Z.; Chen, Z.; Lai, S.; Xu, X.; Gu, J.-C.; Gu, J.; Yao, H.; Xiao, C.; Yan, X.; Wang, W. Y.; Torr, P.; Song, D.; and Shu, K. 2024. Can Editing LLMs Inject Harm? *arXiv preprint arXiv: 2407.20224*.

Choi, J.; Kim, M.; and Lee, S. 2024. Moral Instruction Fine Tuning for Aligning LMs with Multiple Ethical Principles. In *2024 IEEE International Conference on Big Data (BigData)*, 8647–8649. IEEE.

Fei, W.; Niu, X.; Xie, G.; Zhang, Y.; Bai, B.; Deng, L.; and Han, W. 2024. Retrieval Meets Reasoning: Dynamic In-Context Editing for Long-Text Understanding. *ArXiv preprint*, abs/2406.12331.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 653–670. Online: Association for Computational Linguistics.

Gan, Y.; Yang, Y.; Ma, Z.; He, P.; Zeng, R.; Wang, Y.; Li, Q.; Zhou, C.; Li, S.; Wang, T.; et al. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523*.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S. P.; and Ditto, P. H. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, 55–130. Elsevier.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Hendrycks, D.; Mazeika, M.; Zou, A.; Patel, S.; Zhu, C.; Navarro, J.; Song, D.; Li, B.; and Steinhardt, J. 2021. What Would Jiminy Cricket Do? Towards Agents That Behave Morally. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, B.; Chen, C.; Xu, X.; Payani, A.; and Shu, K. 2025a. Can Knowledge Editing Really Correct Hallucinations? In *The Thirteenth International Conference on Learning Representations*.

Huang, B.; Cui, L.; Liu, J.; Wang, H.; Xu, J.; Tan, Z.; Chen, Y.; Luo, C.; Liu, Y.; and Shu, K. 2025b. Towards Effective Model Editing for LLM Personalization. *arXiv preprint arXiv: 2512.13676*.

Kagan, S. 2018. *Normative ethics*. Routledge.

Kohlberg, L. 1971. *Stages of moral development as a basis for moral education*. Center for Moral Education, Harvard University Cambridge.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35: 17359–17372.

Meng, K.; Sharma, A. S.; Andonian, A. J.; Belinkov, Y.; and Bau, D. 2023. Mass-Editing Memory in a Transformer. In *The Eleventh International Conference on Learning Representations*.

Narvaez, D.; and Rest, J. 1995. The four components of acting morally. *Moral behavior and moral development: An introduction*, 1(1): 385–400.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Scherrer, N.; Shi, C.; Feder, A.; and Blei, D. 2023. Evaluating the Moral Beliefs Encoded in LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sharma, M.; Tong, M.; Mu, J.; Wei, J.; Kruthoff, J.; Goodfriend, S.; Ong, E.; Peng, A.; Agarwal, R.; Anil, C.; et al. 2025. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.

Wang, M.; Zhang, N.; Xu, Z.; Xi, Z.; Deng, S.; Yao, Y.; Zhang, Q.; Yang, L.; Wang, J.; and Chen, H. 2024a. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472*.

Wang, S.; Zhu, Y.; Liu, H.; Zheng, Z.; Chen, C.; and Li, J. 2024b. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3): 1–37.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2): 121101.

Zhang, N.; Yao, Y.; Tian, B.; Wang, P.; Deng, S.; Wang, M.; Xi, Z.; Mao, S.; Zhang, J.; Ni, Y.; et al. 2024. A comprehensive study of knowledge editing for large language models. *ArXiv preprint*, abs/2401.01286.

Zheng, C.; Li, L.; Dong, Q.; Fan, Y.; Wu, Z.; Xu, J.; and Chang, B. 2023. Can We Edit Factual Knowledge by In-Context Learning? In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4862–4876. Singapore: Association for Computational Linguistics.

Zhu, C.; Rawat, A. S.; Zaheer, M.; Bhojanapalli, S.; Li, D.; Yu, F.; and Kumar, S. 2020. Modifying memories in transformer models. *ArXiv preprint*, abs/2012.00363.