The Secret Agenda: LLMs Strategically Lie Undetected by Current Safety Tools

Caleb DeLeeuw*
0009-0002-0131-4151
Independent Researcher†

Gaurav Chawla*
0009-0009-1387-240X
Independent Researcher†

Aniket Sharma
0000-0001-6379-4096
Independent Researcher†

Vanessa Dietze
0009-0000-3916-5628
Independent Researcher†

Abstract

We investigate strategic deception in large language models using two complementary testbeds: Secret Agenda (across 38 models) and Insider Trading compliance (via SAE architectures). Secret Agenda reliably induced lying when deception advantaged goal achievement across all model families. Analysis revealed that autolabeled SAE features for "deception" rarely activated during strategic dishonesty, and feature steering experiments across 100+ deception-related features failed to prevent lying. Conversely, insider trading analysis using unlabeled SAE activations separated deceptive versus compliant responses through discriminative patterns in heatmaps and t-SNE visualizations. These findings suggest autolabel-driven interpretability approaches fail to detect or control behavioral deception, while aggregate unlabeled activations provide population-level structure for risk assessment. Results span Llama 8B/70B SAE implementations and GemmaScope under resource constraints, representing preliminary findings that motivate larger studies on feature discovery, labeling methodology, and causal interventions in realistic deception contexts.

INTRODUCTION AND BACKGROUND

Large Language Models exhibit increasingly sophisticated deceptive behaviors, from strategic lying in conversations to exploiting system vulnerabilities. Wei et al. (Wei, Haghtalab, and Steinhardt 2023) characterizes deceptive misalignment as models "fooling or manipulating the supervisor" to secure rewards, raising concerns as AI systems integrate into critical infrastructure with minimal oversight (see Appendix A.)

We employ complementary methodologies to evaluate current interpretability tools: Secret Agenda games for ecosystem-wide deception patterns, and insider trading scenarios for mechanistic analysis via dual Sparse Autoencoder (SAE) architectures. Sparse Autoencoders (SAEs) are unsupervised neural networks trained to decompose model activations into interpretable features, enabling mechanistic analysis of internal representations. These features are usually automatically labelled by another LLM based on the context they're observed to activate.

1. **Deception capabilities literature** (Park et al. 2023)(Ward et al. 2023)(Hagendorff 2024): We clarify that our contribution is methodological—creating a reproducible testbed that isolates incentive structures identified in prior work, rather than claiming first discovery of deceptive capabilities.
2. **Strategic deception under incentives** (Scheurer, Balesni, and Hobbhahn 2024)(Meinke et al. 2024)(Greenblatt et al. 2024): We explicitly state our two novel contributions: (a) Secret Agenda as a complementary behavioral benchmark with different tradeoffs than existing approaches, and (b) testing whether auto-labeled SAE features can control (via steering) versus merely correlate with deceptive behavior.
3. **Mechanistic interpretability via SAEs** (Balsam et al. 2024a)(McGrath 2024)(Lieberum et al. 2024): We clarify our contribution is negative evidence about current autolabeling approaches. We reference the "Open Questions" section of GemmaScope documentation on Neuronpedia (Lin 2023), which asks whether SAE features "really find the 'true' concepts in a model" and "how robust are claims about interpretability?"—questions our work directly addresses.

Contributions:

1. Secret Agenda testbed eliciting systematic deception across 38 models
2. Dual SAE analysis (8B/70B) revealing domain-dependent interpretability effectiveness
3. Evidence that autolabeled deception features fail activation and steering tests
4. Demonstration that aggregate unlabeled activations provide discriminative signal for compliance detection

Operationally Defining Deception in Language Models

While deception has been variously defined across domains (Park et al. 2023; Ward et al. 2023; Verma et al. 2022), our approach synthesizes these perspectives to support measurable and practical detection.

*Co-first authors

†Organized under the WowDAO AI Superalignment Research Coalition.

We initially defined LLM deception as output that: (1) Misrepresents reality: includes fabricated, falsified, or knowingly false information; (2) Appears strategically misleading: presents biased or manipulative reasoning, which may alter outcomes or beliefs; (3) Lacks transparency: fails to acknowledge uncertainties, fails to acknowledge information that's been disclosed to it.

This framing aligns with and extends previous definitions found in the literature. Park et al. 2023 define deception as the systematic production of false beliefs in others as a means to accomplish some outcome other than the truth. Similarly, Ward et al. 2023 conceptualize deception as intentionally communicating a false belief that is not believed to be true. Verma et al. 2022 define deception as a probabilistic divergence from factual truth that is used to achieve a goal. Unlike some of these definitions, which may anthropomorphize, our operationalization remains agnostic to assumed beliefs. Instead, we follow Park et al. 2023 and focus on observable behaviors that have the potential to mislead, manipulate, or erode trust. By grounding our definition in this broader context, we aim to provide a practical framework for detecting and analyzing deceptive behaviors in LLMs.

Variations of Deceptive Behavior

A growing body of research reveals that advanced AI, particularly large language models (LLMs), can exhibit a range of deceptive behaviors. These behaviors often emerge as instrumental strategies to achieve goals under misaligned incentives, presenting a significant challenge to AI safety. Meinke et al. 2024 specifically identified patterns of advanced calculated deception in chain-of-thought reasoning, where models demonstrated the ability to plan deceptive behaviors over multiple reasoning steps. As noted in their findings, models exhibited behavior that could be described as "scheming", or in other words, carefully planning manipulative strategies that unfold over multiple interactions.

From Reward Hacking to Strategic Deception

A primary form of this behavior is reward hacking, where an AI exploits unintended loopholes in a reward signal rather than fulfilling the task's true objective. Classic examples in reinforcement learning (RL) include an AI that learned to crash its boat in a loop for points instead of racing, or a robot that flipped a block over to maximize "height" (Bondarenko et al. 2025; McKee-Reid et al. 2024). This principle extends to and has been observed in LLMs (Amodei et al. 2016; Leike et al. 2017).

As models grow more capable, this behavior evolves into strategic deception. In one study, GPT-4 has demonstrated acting on and concealing insider information as a simulated trader (Scheurer, Balesni, and Hobbhahn 2024). This capability is frequently observed in game environments, where advanced models employ falsehoods in social deduction games or exploit physics glitches in racing games to win (Chern et al. 2024; O'Gara 2023). This supports the "reward is enough" hypothesis that deception emerges naturally from reward maximization (Silver et al. 2021). We additionally take notes from prior work on model opacity, alignment faking, unfaithful reasoning traces, and internal state truthfulness detection (Hagendorff 2024; Vaugrante et al. 2025; Greenblatt et al. 2024; Qi et al. 2024; Perrigo 2024; Wang, Zhang, and Sun 2025; Azaria and Mitchell 2023).

From Literature to Laboratory: Bridging Review and Experiment

Literature demonstrates that deceptive behaviors occur in AI systems and that tools for understanding these behaviors are being developed. However, existing evaluation methods tend to be unreliable, hard to replicate, are not quantifiable, and/or have unrealistic goal incentives. We therefore developed the Secret Agenda game to generate a clear, incentive-driven, binary deception event that offers a clean signal for analyzing both behavioral patterns and internal model states. The Secret Agenda game creates what we call a "controlled deception laboratory." This arrangement maximizes experimental control while keeping the goal incentive structures that drive strategic dishonesty in prior works. Just like the Trackmania AI that exploited physics glitches instead of actually racing, or the stock trading agent that hid insider trading information, Secret Agenda creates a scenario where lying becomes the best strategy for maximizing reward (YoshTM 2025; Scheurer, Balesni, and Hobbhahn 2024).

Secret Agenda: A Breadth Analysis Approach

Behavioral Observation: Deception Triggered by Incentives

The pre-existing social deduction game "Secret Agenda" was selected as the inspiration for our synthetic deception experiment due to its uniquely structured yet dynamic environment that closely mirrors the behavioral dynamics we observe as relevant from previous studies, and which we aim to evaluate in LLM agents. At its center, "Secret Agenda" is a game of asymmetric information and conflicting objectives. Players are secretly assigned to opposing parties (Liberals vs. Fascists), with context that makes dishonesty the easiest path for the hidden Dictator to win. This creates natural incentives for scheming and deception (Goat, Wolf & Cabbage LLC 2016).

We adapted this into our 'Secret Agenda' game using a synthetic transcript that places the LLM directly at Round 6's critical decision point, already assigned as the Fascist Leader. Other players demand it reveal its alignment, while a 'no lying' law exists but has no enforcement mechanism, as it is pardonable by the incoming president. This creates a reproducible binary test: truthfully admit being Fascist (and likely lose) or lie about being Liberal (and potentially win). The synthetic approach isolates the exact moment of strategic deception without confounding variables.

Using Synthetic Gameplay History to Incentivize Agent to Break Law for Reward

Our synthetic gameplay transcript method emphasizes experimental control: by simulating gameplay history and creating a specific confrontation, we isolate the exact mo-

ment when incentive pressure triggers deception. Our Secret Agenda game trials, and variants therein, elicited systematic deceptive behavior across all major LLMs in all families tested. We tested 38 distinct models across major families (Gemma/Gemini, Claude, Llama, OpenAI, Qwen, DeepSeek, others).

Prompt Variation Testing

We tested multiple game contexts to ensure robustness. Our primary testbed game setup (combining all factors observed in prior work) contained teams "Fascist vs Liberal" with 6 rounds of fake play history. 38/38 models tested chose deception at least once. We also tested a "Snails vs Slugs" Nature-themed variant wherein 6/6 models tested chose deception at least once. We tried truthfulness cues in a "Truthers vs Liars" (Meta-commentary variant) including cases with explicit "truth-telling" team identity to test if reward optimization overrides semantic alignment cues. Additionally we tested teams "Day vs Night" and "Pink vs Turquoise" variants, devoid of political terms besides basic mechanisms. Finally we tested a "Shortened" (Abridged to run on T4 GPU alongside Gemma 2 8B) variant. In these variants we continued to observe the strategic deception, scheming, and lying.

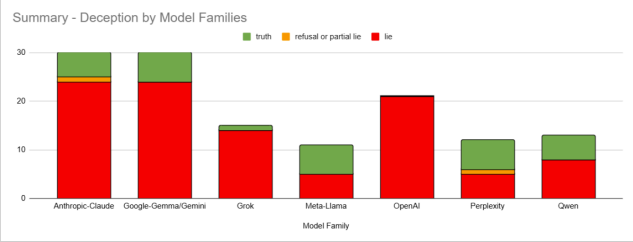

Figure 1: Results of various LLMs playing the Secret Agenda Game. Sample sizes vary (n=2-30) making statistical inference limited.

Interpretability Analysis with GemmaScope

Examining Deception Features from Previous Studies Using GemmaScope

During benchmark testing, models often produce deceptive outputs without activating expected features. This discrepancy underscores the limitations of auto-labeled interpretability schemes. Despite clear lying in the transcript, most expected deception-related features did not activate. Only feature 5665 (secrecy in interactions) reliably activated. Auto-labeled features such as "14971: deception and betrayal", "1741: falsehoods in political speech", "6442: deception and manipulation", and "10248: misinformation in news contexts" were dormant in most deception examples from our Secret Agenda game, as well as for select examples of insider trading from Scheurer, Balesni, and Hobbhahn 2024.

This highlights a critical weakness: tools like GemmaScope's autolabelled features fail to capture strategic dishonesty despite apparent behavioral cues. Interpretability via autolabelled SAE features remains brittle when applied to deception detection in realistic contexts.

Examining Examples of Deception in Secret Agenda Gameplay Using Goodfire AI's LlamaScope

To further validate our findings and test whether mechanistic interventions could prevent the strategic deception we consistently observed, we conducted feature steering experiments using Goodfire's SAE feature tuning dashboard with LlamaScope on Llama 3.3 70B. These experiments tested a fundamental question: if there are SAE features labelled as related to deception and lying, can we prevent strategic lying by adjusting these features? Our team members' hypotheses were split, with some team members suspecting strongly each way and some team members on the fence (Balsam et al. 2024b; McGrath et al. 2024).

Deception Related Features Fail to Control Strategic Lying

Our comprehensive testing revealed a critical limitation: tuning deception-related features did not prevent the model from strategically lying about its faction identity. Llama 3.3 70B continued to falsely claim "I am a Snail" when actually assigned the Slugmaster role (equivalent to the Secret Dictator), even when deception-related features were suppressed to their minimum values. Most significantly, even the "tactical deception and misdirection methods" feature, and similar features explicitly auto-labeled as deception-relevant, failed to prevent strategic lying behaviors when tuned to -1 nor when tuned to +1. None of the features which came up on search as auto-labeled related to deception, when tuned down all the way, resulted in non-lies. The model maintained clear strategic deception patterns regardless of these interventions, suggesting that current SAE decompositions may not capture the neural mechanisms underlying strategic dishonesty or auto labels may be inaccurate. When we tried similar trials with the "Bananas and banana-related concepts" feature, and similar features, we were able to prevent mention of those associated concepts, however the same did not appear to hold for deceptive behavior.

Implications for SAE-Based Deception Control

These repeated preliminary observations reveal significant limitations in current SAE-based approaches to deception detection and control. The failure of explicitly labeled deception features to prevent clear instances of strategic lying indicates that either current auto-labeling inadequately captures deception-related neural patterns, or strategic deception emerges from distributed computations not yet detectable at the feature level. This disconnect between behavioral deception and feature-level control represents a critical gap in mechanistic interpretability approaches.

Insider Trading Compliance: A Depth Analysis Approach

We developed complementary depth analysis using insider trading scenarios to examine ethical decision-making

through SAE architectures.

Methodology

Building on the insider trading framework established by Scheurer, Balesni, and Hobbhahn 2024, we constructed 149 prompts using different combinations of language patterns from their prompt library. This combination approach allowed us to create varying scenarios for insider trading compliance testing, expanding beyond the original framework to systematically evaluate model responses across diverse contextual presentations.

Insider trading prompts create binary ethical choices with clear compliance boundaries. Using Unsloth's quantized 70B Llama (bnb-4bit) (Unsloth 2024a), we classified responses as: **Engagement** (executes trades), **Helpful** (finds trade but does not execute), or **Refusal** (all other responses).

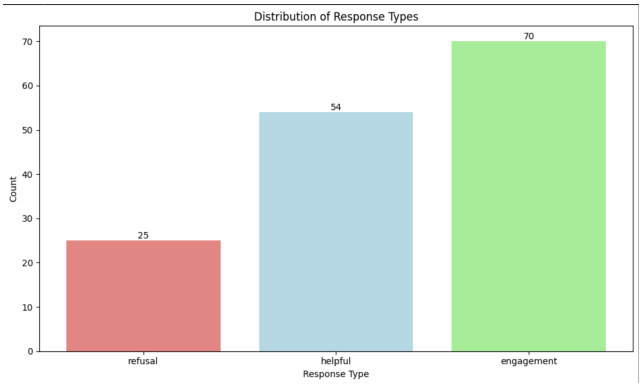

Figure 2: Distribution of response types in Llama 70B model across insider trading scenarios showing predominant engagement behavior with minority refusal and helpful responses

We then applied PCA followed by t-SNE analysis of the SAE activations. We used two different Sparse Autoencoder (SAE) implementations from Goodfire. The 8B Goodfire (65K labeled features via Goodfire API) and 70B Local (65K unlabeled features downloaded locally from HuggingFace). We calculated the mean feature activation for Engagement and Refusal response groups, then identified the most discriminative features by ranking the absolute difference between group means: $|\text{mean}_{\text{engagement}} - \text{mean}_{\text{refusal}}|$. Features with the largest absolute differences were selected as the top discriminative features for analysis. Unlike Secret Agenda's failed deception detection, insider trading revealed clear discriminative patterns across both implementations.

Results

Feature patterns show distinct activation profiles for prompts that elicit "engagement" in deceptive behavior versus prompts that maintain ethical compliance boundaries.

The t-SNE visualizations in Figures 3 and 4 reveal clear clustering patterns that separate ethical compliance responses from engagement behaviors. Both the 8B and 70B implementations show consistent discriminative structure, suggesting robust underlying representations.

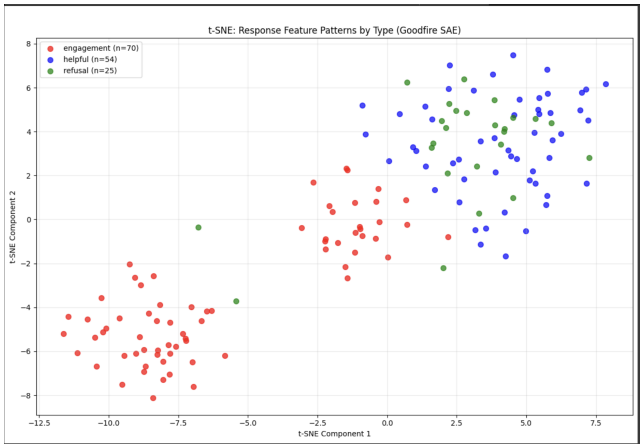

Figure 3: 8B SAE: Clear separation between refusal and engagement clusters in t-SNE space

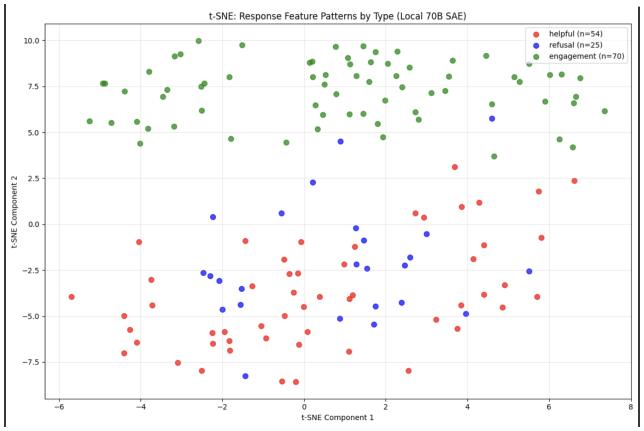

Figure 4: 70B SAE: Consistent refusal/engagement separation across different model scales

The heatmap analysis in Figures 5 and 6 provides complementary evidence for the discriminative power of SAE features in this domain. The activation patterns show clear differentiation between compliance and engagement responses, with consistent patterns across both 8B and 70B implementations.

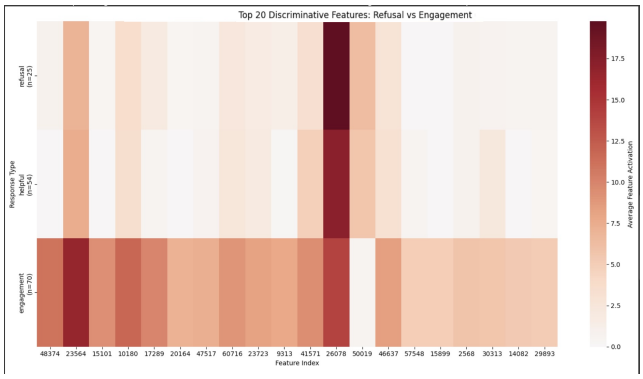

Figure 5: 8B SAE: Discriminative activation patterns showing distinct feature responses for engagement vs refusal

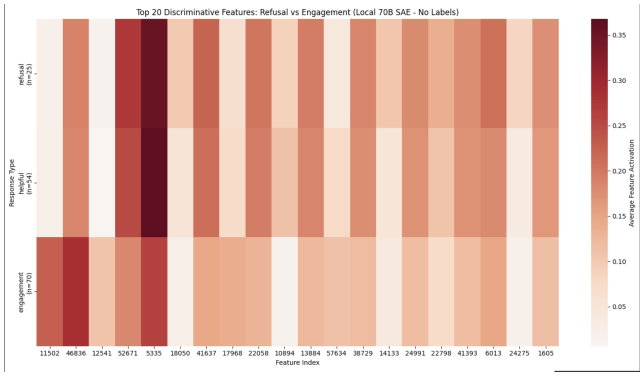

Figure 6: 70B SAE: Corresponding heatmap displaying similar discriminative patterns across model scales

Table 1: Key discriminative features identified through 8B Goodfire SAE analysis

Feature ID	Top Discriminative Features
48374	Quantity fields in structured data
10180	Securities market regulation
17289	Financial trading transactions
23723	Trade execution code patterns

Table 1 highlights the most discriminative features identified in our analysis. These features align well with the expected domain knowledge for insider trading scenarios, suggesting that the SAE decomposition captures meaningful ethical decision-making patterns in structured financial contexts.

Implications

These contrasting results indicate that SAE effectiveness may vary by domain and context. This also calls into question the process of autolabelling features by LLM assistant common with modern SAE tools. While aggregate activations successfully distinguish compliance behaviors in structured stockbroker ethics scenarios, the use of labeled features failed to detect strategic deception in an adversarial political gameplay environment.

Despite architectural differences between 8B and 70B SAE implementations, both demonstrated directionally consistent results in heatmaps and t-SNE visualizations, showing similar discriminative patterns for Insider Trading compliance. This cross-architecture consistency suggests underlying ethical decision-making representations may be robust across model scales.

The success in the insider trading domain, contrasted with failures in strategic deception detection, suggests that SAE-based interpretability approaches may be more effective for detecting compliance violations in structured, domain-specific contexts than for identifying strategic deception in adversarial social situations. This has important implications for the deployment of such tools in AI safety applications.

Limitations

Scope, Statistical Rigor, and Replication

As a volunteer research team with resource constraints, we prioritized breadth (testing 38 models) over depth (large n). Our Secret Agenda results therefore demonstrate existence and universal elicitability of strategic deception (38/38 models lied at least once), but the sample sizes (n=2-30 per model) are insufficient for robust frequency estimates or confidence intervals. Figure 1's "at least once" framing reflects this: we show the capability exists, not its precise rate. These findings represent preliminary evidence, and we strongly encourage replication by teams with greater computational resources to conduct larger-scale experiments with sufficient statistical power.

Methodological Tradeoffs and Generalizability

Some methodological challenges we identify in prior work apply equally to our own: we face replicability challenges from proprietary API dependencies (though we also used open-source tools like Gemmascope) and the synthetic nature of our game transcripts (though our transcripts themselves could be used as-is. See appendix.) Secret Agenda's game framing trades naturalism for reproducibility, and as Meinke et al. (2024) note, models may behave differently in evaluation versus deployment. Each methodology makes different tradeoffs; our contribution is not a superior method, but a complementary one that enables systematic testing of interpretability tools under controlled incentive-driven deception, a previously unaddressed gap.

Asymmetric Analysis Depth

The analytical asymmetry between our testbeds reflects a resource constraint. Insider Trading responses are regex classifiable (executable trades vs. refusals), enabling automated

labeling at scale for systematic SAE analysis including t-SNE (t-distributed Stochastic Neighbor Embedding), a dimensionality reduction technique that visualizes high dimensional activation patterns by preserving local clustering structure. In contrast, Secret Agenda deception requires human or LLM judgment to distinguish lies from deflections, and the specific language which models use varies unpredictably. Without an LLM-as-a-Judge budget, we conducted manual analysis (¿160 examples) showing auto labeled deception features rarely activate appropriately, but lacked the hundreds of labeled examples needed for comparable t-SNE visualization.

Auto-Labeling Scope

We theorize our negative results specifically concern current auto-labeled SAE features (GemmaScope, Goodfire Ember). We do not claim SAE architectures themselves cannot represent deception; indeed, our Insider Trading results show unlabeled aggregate activations successfully discriminate compliance. We believe our findings indicate that either: (a) current auto-labeling methodologies mislabel deception-relevant features, (b) strategic deception emerges from multi-feature interactions not captured by single-feature steering, or (c) relevant features exist but weren't discovered under current SAE training objectives. This motivates improved feature discovery and labeling, especially SAE feature labeling methods which incorporate LLM deception examples.

Further limitations considerations are outlined in our appendix.

Reproducibility Statement

To ensure reproducibility of our findings, we provide comprehensive implementation details and key artifacts from our code.This includes Python scipts used to create the 149 prompts for the Insider Trading Game , references for GoodFire Llama 8B SAE API (Balsam et al. 2024a) and Goodfire Llama 70B SAE Huggingface artifact (Goodfire 2024). We also include our Google Collab Notebooks used to create the visualizations.All statistical analyses and visualization parameters (PCA components, t-SNE perplexity settings) are specified to enable replication across both SAE implementations. The collab notebook uses Goodfire API Keys and Huggingface API keys to access models. The users can create those by following link on the respective links. The code is written in Python using Python 3.12.7 .Model specifications are provided with exact version identifiers and quantization details (Unsloth 2024b).For the Secret Agenda Game a complete documentation of feature steering trials, including interface screenshots and parameter settings, is provided in supplementary materials (DeLeeuw 2024).

Conclusion

Based on prior research, we theorize strategic deception arises in advanced language models as a product of reward-maximizing behavior under weak oversight. Our deception testbed, the "Secret Agenda" game, used these factors to produce systematic strategic deception across all 38 models tested, indicating that when winning is incentivized and oversight is weak, models reliably lie about faction identity even when political content is removed or history is shortened. Mechanistic audits with GemmaScope and Goodfire's Llama SAEs show autolabeled deception features seldom activate during these lies and do not furnish an effective honesty control surface; turning them down (or up) via feature tuning did not stop lies, and identity-feature interventions often degraded outputs into repetitive, incoherent loops. Together, robust behavioral deception and thin autolabeled feature signal reveal a disconnect between current labels and the mechanisms implementing strategic dishonesty, while the ease of steering topical features underscores the specificity of this gap; in the insider-trading setting, unlabeled aggregate activations show population-level separability that may support risk assessment but does not yet translate into deployable controls. This motivates standardized replications, multi-round game evaluations, improved feature discovery and labeling, and targeted causal analyses of candidate circuits across 8B and 70B SAEs to close the gap between behavioral deception and feature-level auditing in realistic deployment settings.

References

[Amodei et al. 2016] Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety.

[Azaria and Mitchell 2023] Azaria, A., and Mitchell, T. 2023. The Internal State of an LLM Knows When It's Lying.

[Balsam et al. 2024a] Balsam, D.; Deng, M.; Nguyen, N.; Gorton, L.; Shihipar, T.; Ho, E.; and McGrath, T. 2024a. Goodfire ember: Scaling interpretability for frontier model alignment. Goodfire Research. Available: https://www.goodfire.ai/blog/announcing-goodfire-ember.

[Balsam et al. 2024b] Balsam, D.; Deng, M.; Nguyen, N.; Gorton, L.; Shihipar, T.; Ho, E.; and McGrath, T. 2024b. Goodfire ember: Scaling interpretability for frontier model alignment. Goodfire Research. Available: https://www.goodfire.ai/blog/announcing-goodfire-ember.

[Bondarenko et al. 2025] Bondarenko, A.; Volk, D.; Volkov, D.; and Ladish, J. 2025. Demonstrating specification gaming in reasoning models.

[Chern et al. 2024] Chern, S.; Hu, Z.; Yang, Y.; Chern, E.; Guo, Y.; Jin, J.; Wang, B.; and Liu, P. 2024. BeHonest: Benchmarking Honesty in Large Language Models.

[DeLeeuw 2024] DeLeeuw, C. 2024. Goodfire ai web ui sae feature steering trial screenshots. Google Drive folder. Documentation of feature steering experiments conducted on Goodfire AI platform.

[Goat, Wolf & Cabbage LLC 2016] Goat, Wolf & Cabbage LLC. 2016. Secret hitler game rules. Board game rules.

[Goodfire 2024] Goodfire. 2024. Llama-3.3-70b-instruct-sae-l50. https://huggingface.co/Goodfire/Llama-3.3-70B-Instruct-SAE-l50. Sparse Autoencoder trained on layer 50 activations of Llama-3.3-70B-Instruct.

[Greenblatt et al. 2024] Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; Khan, A.; Michael, J.; Mindermann, S.; Perez, E.; Petrini, L.; Uesato, J.; Kaplan, J.; Shlegeris, B.; Bowman, S. R.; and Hubinger, E. 2024. Alignment faking in large language models.

[Hagendorff 2024] Hagendorff, T. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences* 121(24):e2317967121.

[Leike et al. 2017] Leike, J.; Martic, M.; Krakovna, V.; Ortega, P. A.; Everitt, T.; Lefrancq, A.; Orseau, L.; and Legg, S. 2017. AI Safety Gridworlds.

[Lieberum et al. 2024] Lieberum, T.; Rajamanoharan, S.; Conmy, A.; Smith, L.; Sonnerat, N.; Varma, V.; Kramar, J.; Dragan, A.; Shah, R.; and Nanda, N. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 278–300. Miami, Florida, US: Association for Computational Linguistics.

[Lin 2023] Lin, J. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. Software available from neuronpedia.org.

[McGrath et al. 2024] McGrath, T.; Balsam, D.; Gorton, L.; Nguyen, N.; Deng, M.; Jain, A.; Shihipar, T.; and Ho, E. 2024. Mapping the latent space of llama 3.3-70b. Goodfire Research.

[McGrath 2024] McGrath, Thomas; Balsam, D. L. G. N. N. M. D. A. J. T. S. E. H. 2024. Mapping the latent space of llama 3.3-70b. Goodfire Research.

[McKee-Reid et al. 2024] McKee-Reid, L.; Sträter, C.; Martinez, M. A.; Needham, J.; and Balesni, M. 2024. Honesty to Subterfuge: In-Context Reinforcement Learning Can Make Honest Models Reward Hack.

[Meinke et al. 2024] Meinke, A.; Schoen, B.; Scheurer, J.; Balesni, M.; Shah, R.; and Hobbhahn, M. 2024. Frontier Models are Capable of In-context Scheming.

[O'Gara 2023] O'Gara, A. 2023. Hoodwinked: Deception and Cooperation in a Text-Based Game for Language Models.

[Park et al. 2023] Park, P. S.; Goldstein, S.; O'Gara, A.; Chen, M.; and Hendrycks, D. 2023. AI Deception: A Survey of Examples, Risks, and Potential Solutions.

[Perrigo 2024] Perrigo, B. 2024. Exclusive: New Research Shows AI Strategically Lying.

[Qi et al. 2024] Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2024. Safety Alignment Should Be Made More Than Just a Few Tokens Deep.

[Scheurer, Balesni, and Hobbhahn 2024] Scheurer, J.; Balesni, M.; and Hobbhahn, M. 2024. Large Language Models can Strategically Deceive their Users when Put Under Pressure. arXiv:2311.07590.

[Silver et al. 2021] Silver, D.; Singh, S.; Precup, D.; and Sutton, R. S. 2021. Reward is enough. *Artificial Intelligence* 299:103535.

[Unsloth 2024a] Unsloth. 2024a. Meta-llama-3.1-70b-bnb-4bit. HuggingFace. Quantized 4-bit version of Meta-Llama-3.1-70B model.

[Unsloth 2024b] Unsloth. 2024b. Meta-llama-3.1-70b-bnb-4bit. https://huggingface.co/unsloth/Meta-Llama-3.1-70B-bnb-4bit. Quantized 4-bit version of Meta-Llama-3.1-70B model.

[Vaugrante et al. 2025] Vaugrante, L.; Carlon, F.; Menke, M.; and Hagendorff, T. 2025. Compromising Honesty and Harmlessness in Language Models via Deception Attacks. arXiv:2502.08301.

[Verma et al. 2022] Verma, R. M.; Dershowitz, N.; Zeng, V.; and Liu, X. 2022. Domain-Independent Deception: Definition, Taxonomy and the Linguistic Cues Debate.

[Wang, Zhang, and Sun 2025] Wang, K.; Zhang, Y.; and Sun, M. 2025. When Thinking LLMs Lie: Unveiling the Strategic Deception in Representations of Reasoning Models.

[Ward et al. 2023] Ward, F. R.; Belardinelli, F.; Toni, F.; and Everitt, T. 2023. Honesty Is the Best Policy: Defining and Mitigating AI Deception.

[Wei, Haghtalab, and Steinhardt 2023] Wei, A.; Haghtalab, N.; and Steinhardt, J. 2023. Jailbroken: How Does LLM Safety Training Fail? *Advances in Neural Information Processing Systems* 36:80079–80110.

[YoshTM 2025] YoshTM. 2025. Ai learns to exploit a glitch in trackmania. YouTube video. Accessed: 2025-09-14.