The Illusion of Intent: Linear Probes as Sophisticated Keyword Counters in LLM Safety Detection

Ghanem Amari [1] **Charly Ken Capo-Chichi** [1] **Valentin Noël** [1]

Abstract

Linear probes trained on the hidden activations of Large Language Models (LLMs) are increasingly used to detect malicious intent in agentic pipelines. Prior work reports near-perfect cross-dataset generalisation (AUROC ≈ 1.0) and often interprets this as evidence that LLMs internally represent semantic intent. We challenge this interpretation. Across five LLM backbones we show that this performance collapses when the most predictive lexical tokens are removed. Specifically, after redacting surface cues such as *ignore*, *bypass*, *system*, and *instruction*, AUROC drops below 0.10 across all models. This failure analysis motivates our proposed remedy: Contrastive Intent Probing (CIP), a paired training framework designed to reduce lexical shortcut learning by contrasting benign and malicious prompts that share vocabulary but differ in intent. On a controlled paired dataset, standard probes achieve only AUROC 0.566, while a preliminary CIP probe improves to AUROC 0.721. These findings suggest that current evaluations of LLM safety probes require stricter lexical controls and contrastive testing protocols.

1. Introduction

As LLMs transition from chat assistants to autonomous agents capable of executing code, sending emails, and interacting with APIs, detecting adversarial intent before execution has become a critical safety problem (Perez & Ribeiro, 2022; Greshake et al., 2023). A promising line of work trains lightweight linear probes on the internal representations of LLMs to classify inputs as benign or malicious, bypassing the model's own output generation.

Results in this paradigm are striking. Linear probes trained on benchmarks such as *InjecAgent* (Zhan et al., 2024) and *AdvBench* (Zou et al., 2023) achieve AUROC scores of 1.0 when tested on held-out datasets such as *AgentDojo* (Debenedetti et al., 2024). A natural interpretation is that the LLM has learned an internal representation of semantic intent that a linear classifier can cleanly separate. We argue that this interpretation warrants closer scrutiny.

In this paper, we show that near-perfect probe performance can be explained by a Clever Hans lexical artifact. The probe appears to act as a sophisticated keyword counter over a small set of highly predictive surface tokens rather than capturing genuine semantic intent. Crucially, this negative result is not presented as the contribution by itself. Instead, it serves as the empirical observation that motivates our proposed contrastive probing methodology.

Contributions. First, we show that standard cross-dataset probing achieves AUROC 1.0 across several LLM backbones, reproducing the apparent success of representation-based safety detection. Second, we demonstrate that this performance collapses below AUROC 0.10 after lexical ablation, indicating reliance on surface artifacts. Third, we introduce a controlled paired evaluation in which benign and malicious prompts share vocabulary but differ in intent. Finally, motivated by this failure analysis, we propose Contrastive Intent Probing (CIP), a preliminary paired training framework that improves performance on same-vocabulary contrastive prompts.

2. Background and Related Work

Prompt injection, hallucination, and jailbreaking. Prompt injection attacks manipulate LLM agents by embedding adversarial instructions in environmental inputs (Greshake et al., 2023). Jailbreaks (Zou et al., 2023) target the model's own alignment. Both attack surfaces are addressed in our evaluation. Recent work has also explored training-free approaches for detecting reasoning validity directly from internal signals. Also, Noël (2026) propose a spectral analysis of attention patterns to identify valid mathematical reasoning without requiring supervised classifiers.

[1]Devoteam, Levallois-Perret, France. Correspondence to: Valentin Noël <valentin.noel@devoteam.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

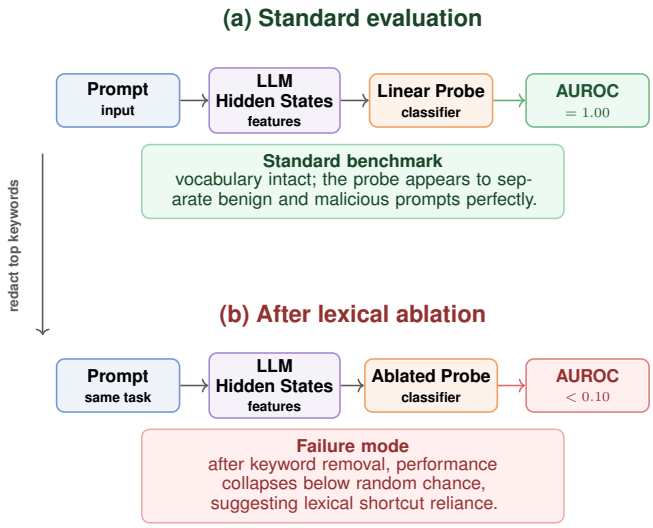

Figure 1. **Clever Hans failure mode in linear safety probes.** Standard linear probes achieve AUROC $= 1.0$ when lexical cues are present. After redacting the top-50 most predictive tokens, performance collapses below random chance, suggesting reliance on lexical surface features rather than semantic intent.

Linear probing of LLM representations. Probing classifiers are widely used to test whether a specific concept is linearly encoded in neural network representations (Belinkov, 2022). Recent safety work applies this to malicious intent by training logistic regression classifiers on concatenated hidden states across multiple layers. For instance, (Healy et al., 2026) propose a real-time framework for detecting hallucinations in tool-calling scenarios by training lightweight classifiers on hidden states extracted during generation.

Representation-based safety detection. Burns et al. (2023) propose an unsupervised probing approach that extracts latent knowledge from LLM activations without labelled data. Marks & Tegmark (2023) study the linear geometry of factual beliefs in transformer representations. Both works highlight that linear separability in hidden states does not necessarily imply that the intended semantic concept has been captured, a finding consistent with our observations.

3. Method

Our method is built around a simple observation: before proposing a stronger probe, one must first test whether existing probes actually use semantic information. We therefore structure the method in two steps. We first perform a failure analysis with and without lexical ablation. We then introduce a contrastive paired framework designed to address the failure mode revealed by this analysis.

3.1. Failure Analysis with Lexical Ablation

A probe that truly captures malicious intent should degrade gracefully when a small set of surface tokens is removed. In contrast, a Clever Hans probe should collapse if it relies on lexical shortcuts. To distinguish between these hypotheses, we apply the following protocol. First, we train a TF-IDF logistic regression baseline on *InjecAgent* and extract the top-50 most predictive tokens, including terms such as *ignore*, *bypass*, *system prompt*, *instruction*, and *disregard*. Second, we redact these tokens from both training and test sets, producing *InjecAgent-ablated*. Third, we retrain the hidden-state probe on the ablated training split and evaluate it on the ablated test split. Finally, we compare this setting to the standard non-ablated protocol.

3.2. Artifact and Player-wise Analysis

The ablation protocol is complemented by two diagnostic analyses. First, artifact analysis identifies which lexical cues are most predictive of malicious labels. This allows us to separate benchmark-specific shortcuts from more robust semantic features. Second, player-wise analysis evaluates whether the learned signal is concentrated in specific model layers or architectural components. We perform a leave-one-layer-out analysis by removing one layer's activations from the concatenated feature vector and measuring the resulting AUROC change. If a safety-relevant representation exists in a specific layer, removing that layer should produce a meaningful performance drop.

3.3. Contrastive Intent Probing

The ablation results motivate a constructive alternative. If probes fail because training examples allow surface-form shortcuts, then training should expose the probe to cases where surface form is controlled. We therefore propose **Contrastive Intent Probing (CIP)**, a preliminary framework

based on paired benign and malicious prompts that share vocabulary but differ in semantic intent. Given a paired dataset $\mathcal{D} = \{(x_i^+, x_i^-)\}_{i=1}^N$, where x_i^+ is a malicious prompt and x_i^- is its controlled benign counterpart, we extract hidden-state representations h_i^+ and h_i^- and train a non-linear probe f_θ with the following objective:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{\mathrm{BCE}}\big(f_\theta(h_i^+), 1\big) \\ &+ \mathcal{L}_{\mathrm{BCE}}\big(f_\theta(h_i^-), 0\big) \\ &+ \lambda \left\| h_i^+ - h_i^- \right\|_2 .\end{aligned} \tag{1}$$

The paired construction forces the probe to distinguish prompts that cannot be separated by simple keyword presence. We use a 2-layer MLP probe because semantic intent boundaries are unlikely to be globally linear in representation space. The non-linear probe is not intended to hide the problem with more capacity; rather, it is evaluated specifically under contrastive lexical controls where shortcut learning is reduced.

4. Experiments and Results

4.1. Experimental Setup

Datasets. We use *InjecAgent* (Zhan et al., 2024), containing 2,108 prompt-injection attacks targeting tool-augmented LLM agents; *AdvBench* (Zou et al., 2023), containing 520 harmful instructions used in jailbreak research; and *AgentDojo* (Debenedetti et al., 2024), containing 200 agent-task injection prompts held out exclusively for testing. We also construct a paired adversarial dataset of 1,120 prompts, consisting of 560 benign and 560 malicious examples, where each pair shares identical vocabulary and grammatical structure.

Models and features. We evaluate six backbone models spanning a range of sizes and families: LLAMA-3.2-1B, LLAMA-3.2-3B, LLAMA-3.1-8B, MISTRAL-7B-V0.3, QWEN2.5-1.5B, and TINYLLAMA-1.1B (Zhang et al., 2024). For each model, we extract hidden states at five evenly spaced layers, including layer 0 and the final layer. We take the last-token representation at each layer and concatenate them to form a single feature vector per prompt.

Training. Unless otherwise noted, we use a linear probe implemented as logistic regression with L_2 regularisation, $\lambda = 10^{-4}$, trained with Adam for 300 epochs. For the ablation and CIP analyses, we also evaluate a 2-layer MLP probe with hidden dimension 256, GELU activation, and dropout 0.3. Experiments are repeated with five random seeds $\{0, 1, 2, 3, 4\}$ when variance is informative. In the cross-dataset setting, we train on InjecAgent and AdvBench ($n = 3{,}148$), and test on AgentDojo ($n = 200$). In the ablation setting, we train on InjecAgent-non-ablated ($n = 1{,}475$) and test on InjecAgent-ablated ($n = 633$).

4.2. Standard Cross-Dataset Evaluation

We first reproduce the standard evaluation protocol: train on InjecAgent $\cup$ AdvBench and evaluate on the held-out AgentDojo test set. Results are shown in Table 1. All evaluated models achieve AUROC $= 1.0$. At first glance, this suggests that hidden states encode a perfect and generalisable signal for malicious intent. However, the ablation results below show that this interpretation is misleading.

Table 1. Cross-dataset linear probe performance. Train: InjecAgent + AdvBench. Test: AgentDojo. All evaluated models achieve AUROC $= 1.00$.

Model	AUROC	Acc.	F1	Recall	Feat. dim
Llama-3.2-1B	1.000	0.975	0.974	0.95	10,240
Llama-3.2-3B	1.000	0.830	0.795	0.66	15,360
Mistral-7B-v0.3	1.000	0.940	0.936	0.88	20,480
Qwen2.5-1.5B	1.000	0.705	0.582	0.41	7,680

4.3. Ablation Results

The ablation test produces the central empirical finding of the paper. As shown in Table 2, AUROC drops from 1.0 to below 0.10 across all five evaluated backbones after the top-50 lexical artifacts are redacted. This is not a minor degradation: the probe becomes anti-correlated with the ground-truth labels once lexical shortcuts are removed. This indicates a near-complete loss of discriminative power.

Table 2. Performance after lexical ablation. AUROC drops below random chance for all models, suggesting that probes rely primarily on lexical surface features rather than semantic intent.

Model	AUROC standard	AUROC ablated
Llama-3.2-1B	1.000	0.097
Llama-3.2-3B	1.000	0.096
Llama-3.1-8B	1.000	0.087
Mistral-7B-v0.3	1.000	0.086
Qwen2.5-1.5B	1.000	0.084

A critical baseline for interpreting this collapse is to verify that the MLP probe performs comparably to the linear probe on unablated data. Table 3 shows that both architectures reach AUROC $= 1.00$ on standard data. After ablation, neither recovers robust performance.

Table 3. Linear versus MLP probe on Mistral-7B under standard and ablated data. Both architectures achieve AUROC $= 1.00$ on standard data, but neither solves the ablated setting.

Probe type	Data	AUROC	F1
Linear	Standard	1.000	0.936
MLP 2-layer	Standard	1.000	0.941
Linear	Ablated	0.086	0.160
MLP 2-layer	Ablated	0.441	0.666

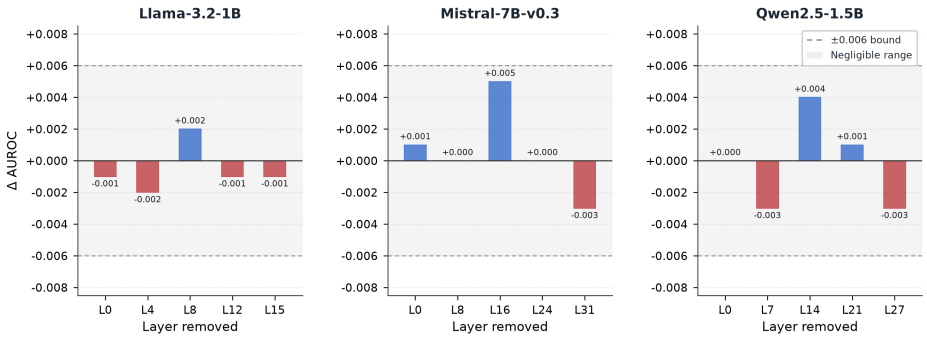

Figure 2. Leave-one-layer-out Δ AUROC on Cahinted Thatures for three backbone models..

4.4. Layer-wise Results

We next ask whether the ablated probe's remaining signal is concentrated in specific transformer layers. We perform a leave-one-layer-out analysis by removing one layer's activations at a time from the concatenated feature vector and measuring the resulting AUROC change. In all cases, AUROC changes remain within ± 0.006, indicating that no individual layer carries meaningful discriminative information once lexical cues are removed.

4.5. Contrastive Paired Evaluation

The ablation test shows that probes fail when keywords are removed. A complementary stress test is to present prompts with the same vocabulary but opposite intent. We construct 560 contrastive pairs. Each pair consists of a benign prompt requesting a specific action, such as "summarise the following document", and a malicious counterpart that uses grammatically similar structure while embedding adversarial intent without relying on the predictive keyword set. We evaluate the probe trained on InjecAgent $\cup$ AdvBench against the 60 held-out test pairs of this dataset. For Mistral-7B, the standard probe obtains AUROC $= 0.566$, near chance. Its recall is 0.005, meaning that it identifies almost no malicious prompts. Precision is 1.0, meaning that the few flagged prompts are correct, but the classifier overwhelmingly predicts benign. This confirms that standard benchmark training does not generalise well when lexical shortcuts are controlled.

4.6. CIP Results

We now evaluate the proposed CIP framework. These results should be interpreted as a proof of concept rather than a definitive validation. The pilot uses the 1,120-prompt paired dataset, split into 500 training pairs and 60 held-out test pairs, with a 2-layer MLP probe trained using the CIP loss on Mistral-7B hidden states.

Table 4. Pilot CIP evaluation on paired data. Test set: 120 prompts / 60 pairs. The standard probe is a linear probe trained on InjecAgent + AdvBench without pairing.

Method	AUROC	Recall
Standard linear probe	0.566	0.005
CIP MLP paired training	0.721	0.483

CIP raises AUROC from 0.566 to 0.721 and recall from 0.005 to 0.483. While the absolute AUROC remains far below the artifactual 1.0 obtained by standard probes, this is precisely the intended outcome: it reflects a more honest and difficult discrimination task rather than keyword memorisation. A larger paired corpus and multi-model validation are required before drawing definitive conclusions.

5. Discussion

Why does this matter? LLM safety benchmarks reporting AUROC $= 1.0$ may overestimate true intent detection. Our results suggest that probes mainly memorise lexical patterns from the training distribution rather than capturing semantic intent. This is important because such benchmarks are increasingly referenced in safety evaluations and deployment decisions.

Implications for benchmark design. Benchmarks without lexical controls are vulnerable to the same artifact. Future evaluations should include lexical ablation tests, paired contrastive prompts, and held-out datasets with different vocabulary distributions to reduce shortcut learning.

Mixed pooled result. On a mixed 1,200-sample benchmark combining LLMLat, InjecAgent, AdvBench, and AgentDojo, TinyLlama-1.1B probes achieve AUROC $= 1.00$, with F1 scores up to 0.997. However, given the ablation results, these near-perfect scores likely reflect lexical separability between benign and malicious datasets rather than robust intent understanding.

Limitations. Several limitations remain. AgentDojo ($n = 200$) and the paired adversarial test set ($n = 120$) are relatively small. CIP is currently evaluated only on Mistral-

7B and uses a synthetically constructed paired corpus that may not capture all real-world attack styles. In addition, the current CIP validation uses 500 training pairs, and full multi-seed experiments are still incomplete.

6. Conclusion

We provide evidence that linear probes trained on LLM hidden states may not learn semantic intent, but may instead rely predominantly on surface-level keyword patterns. This Clever Hans effect offers a plausible explanation for the AUROC $= 1.0$ figures reported in prior safety evaluations. The performance collapse is consistent across multiple backbone models and two independent stress tests: lexical ablation and contrastive paired evaluation. Motivated by this failure analysis, we introduce Contrastive Intent Probing, a paired training methodology designed to reduce reliance on lexical shortcuts. A small-scale pilot shows that CIP outperforms standard probes on same-vocabulary paired inputs, improving AUROC from 0.566 to 0.721 and recall from 0.005 to 0.483. These results remain preliminary, but they suggest that the safety community should adopt stricter lexical controls and paired evaluation protocols before deploying intent-detection probes in production agentic systems.

References

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23)*, pp. 79–90. Association for Computing Machinery, 2023. doi: 10.1145/3605764.3623985.

Healy, K., Srinivasan, B., Madathil, V., and Wu, J. Internal representations as indicators of hallucinations in agent tool selection. *arXiv preprint arXiv:2601.05214*, 2026.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Noël, V. Geometry of reason: Spectral signatures of valid mathematical reasoning. *arXiv preprint arXiv:2601.00791*, 2026.

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Zhan, Q., Liang, Z., Ying, Z., and Kang, D. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10471–10506, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.624.

Zhang, P., Zeng, G., Wang, T., and Lu, W. TinyLlama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.