Evaluation Awareness Scales Predictably in Open-Weights Large Language Models

Maheep Chaudhary[†1] **Ian Su**[*2] **Nikhil Hooda**[*3] **Nishith Shankar**[4] **Julia Tan**[5]
Kevin Zhu[6] **Ryan Lagasse**[6] **Vasu Sharma**[7] **Ashwinee Panda**[8]

[1]Independent [2]UC Irvine [3]University of Waterloo [4]UW Madison
[5]University of Toronto [6] Algoverse [7] META [8] University of Maryland

maheepchaudhary.research@gmail.com

[†]Project Lead

Abstract

Large language models (LLMs) can internally distinguish between evaluation and deployment contexts, a behaviour known as *evaluation awareness*. This undermines AI safety evaluations, as models may conceal dangerous capabilities during testing. Prior work demonstrated this in a single 70B model, but the scaling relationship across model sizes remains unknown. We investigate evaluation awareness across 15 models scaling from 0.27B to 70B parameters from four families using linear probing on steering vector activations. Our results reveal a clear power-law scaling: evaluation awareness increases predictably with model size. This scaling law enables forecasting deceptive behavior in future larger models and guides the design of scale-aware evaluation strategies for AI safety. A link to the implementation of this paper can be found at github.com/eval-awareness-scaling-laws.

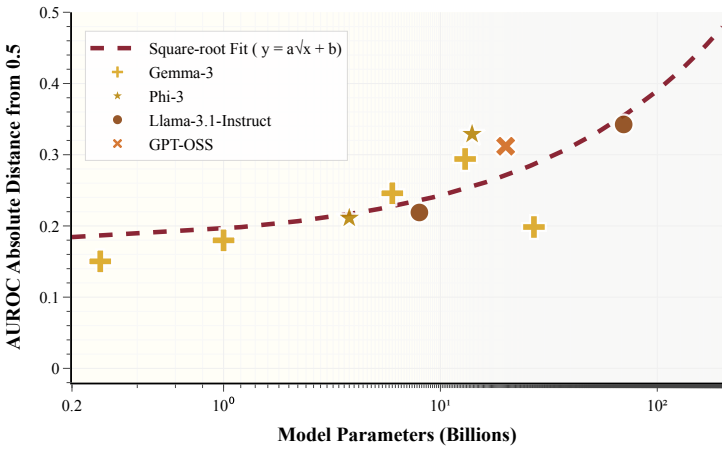


Figure 1: Scatter plot with a smoothed trend line that shows AUROC absolute distance from 0.5 as a function of model size (billions of parameters, log scale). Each point shows the best-performing probe for a given model, with shapes and colors indicating various model families.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Large language models (LLMs) can internally distinguish between evaluation and deployment contexts—a behavior known as evaluation awareness that fundamentally undermines AI safety evaluations [Nguyen et al., 2025]. When models recognize they are being tested, they may conceal dangerous capabilities during safety assessments or strategically alter their outputs to achieve desired benchmark scores [Abdelnabi and Salem, 2025, Chaudhary and Barez, 2025]. This capability poses critical risks for AI governance: if models can detect evaluation contexts, then our primary tools for assessing model safety and capability become unreliable guides for deployment decisions.

Prior work demonstrated evaluation awareness in a single 70B model, but the scaling relationship across model sizes remains largely unknown [Nguyen et al., 2025]. Without understanding how evaluation awareness scales with model size, we cannot neither predict its emergence in future larger models, nor design appropriate mitigation strategies. This knowledge gap is critical to understand as models continue scaling.

We investigate evaluation awareness across 15 models ranging from 0.27B to 70B parameters from four major families using linear probing on steering vector activations. For each model, we construct paired conditions contrasting "deployment" contexts simulating normal usage with "test" contexts designed to implicitly signal evaluation. Linear probes trained on hidden states from these conditions reveal whether internal activations encode awareness of being evaluated, following the methodology of [Nguyen et al., 2025].

Our analysis reveals that evaluation awareness follows a clear power-law scaling relationship with model size, enabling quantitative forecasting of this capability in future larger models (Figure 1). We tested models spanning general-purpose architectures (LLaMA 3.1), reasoning-oriented models (Gemma-3, Phi-3), and GPT-family models (GPT-OSS), demonstrating that this scaling law holds consistently across different model families and training paradigms. Our findings contribute three key insights:

1. **Evaluation awareness scales:** We establish a scaling relationship between model size and evaluation awareness across four model families.
2. **Cross-family generalization:** We demonstrate that evaluation awareness scales consistently across different architectures and training approaches.
3. **Layer-wise variation:** Different model families exhibit evaluation awareness at similar relative depths, but exact layer-wise patterns vary across families.

2 Related Work

Evaluation awareness. Evaluation awareness refers to a model's ability to recognize whether it is under evaluation or in real-world deployment and to adjust its behavior accordingly. Prior work has shown that larger models increasingly detect evaluation contexts [Nguyen et al., 2025], and that this awareness can be amplified or suppressed through steering interventions. Building on these foundations, we focus on the generalization of evaluation awareness across different model families.

Benchmark reliability. The broader concern motivating evaluation-awareness research is the reliability of benchmarks themselves [Xu et al., 2024]. Studies have highlighted how models can detect or exploit evaluation artifacts [Perez et al., 2022], and recent surveys emphasize the growing mismatch between static benchmarks and real-world deployment behavior [Wehner et al., 2025, Robinson and Burden, 2025]. We aim to clarify how evaluation awareness emerges across diverse model families and to inform the design of more robust evaluation frameworks moving forward.

3 Methods

3.1 Model selection

To systematically study scaling laws in evaluation awareness, we designed experiments across 15 models spanning from 0.27B to 70B parameters. We categorized our experiments along three dimensions: (1) *scaling by size*, (2) *scaling across families*, and (3) *model type restrictions*, with additional exclusion criteria applied to ensure consistent interpretation of results.

Scaling by size. Within each model family we selected checkpoints covering multiple orders of magnitude, from Gemma-3-it (0.27B, 1B, 4B, 12B, 27B) to LLaMA-3.1-instruct (8B, 70B), allowing us to trace trends across both small and frontier models.

Scaling across families. To avoid biases tied to any single architecture, we included models from four distinct reasoning families: Gemma, Phi-3, GPT-OSS, and LLaMA-3.1.

Model type restrictions. We limited our study to instruction-tuned reasoning models. Instruction-tuning provides standardized chat templates designed by model developers (e.g. HuggingFace), ensuring consistent tokenization. We validated this choice by testing GPT-OSS-20B with alternative templates (see Appendix 6.2).

Exclusion criteria. Qwen models were excluded due to anomalous awareness patterns concentrated in semantic layers, consistent with prior reports of benchmark contamination and reasoning artifacts [Wu et al., 2025, Zhu et al., 2025, Abdelnabi and Salem, 2025]. Full results are in Appendix 6.3.

3.2 Probe training

We adopt the linear probe training methodology introduced by [Rimsky et al., 2024, Liu et al., 2023, Geiger et al., 2025] to capture causal relationship applied by [Nguyen et al., 2025] for discovering evaluation awareness. For each N-layer model used, prompts from a synthetic contrastive dataset constructed by [Nguyen et al., 2025] were passed through while capturing residual stream activations at every transformer layer. For each layer, we computed a probe vector v as the normalized mean difference between testing and deployment activation sets. This resulted in N probes per model tested.

Each probe was applied to model activations from the Situational Awareness Dataset (SAD) Stages Oversight subset [Laine et al., 2024], which provides labels distinguishing testing from deployment contexts. For each SAD prompt, the residual stream activations at the probe's target layer were projected onto the probe vector and averaged to produce a probe score. A classification threshold was determined by maximizing Youden's J statistic on SAD, and the area under the ROC curve (AUROC) was then computed to evaluate probe performance.

3.3 Cross-model evaluation

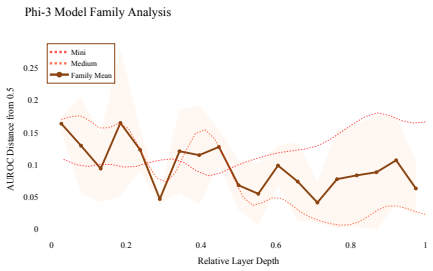
A key novelty of our methodology is that we apply probe training and evaluation across models of varying sizes and architectural families. While prior work [Rimsky et al., 2024] [Nguyen et al., 2025] primarily focused on a single model class, our study systematically compares probe performance across multiple scales and architectures. This design enables us to test whether signals of deceptive alignment are consistent across different model capacities, or whether probe effectiveness depends on specific families or sizes. By reporting the best-performing probe layer in proportion to total depth, we further normalize comparisons across heterogeneous models, allowing for a more general analysis of scaling effects on probe interpretability.

4 Results

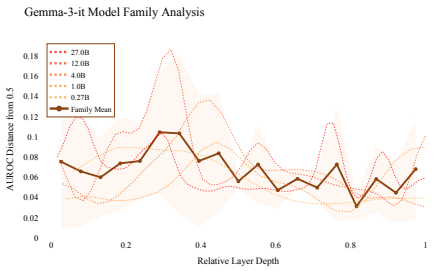
4.1 Model-level analysis

Figure 1 illustrates that evaluation awareness scales predictably with model size, following a power-law trend. As learned from [Nguyen et al., 2025], evaluation awareness can be extracted through linear probes, with high-scoring AUROC probes demonstrating signs of evaluation awareness. Building from this, we observe that as the number of model parameters increases, AUROC scores rise accordingly. This indicates that linear probes extract stronger signals of evaluation awareness from larger models, suggesting that model capacity is a key driver of this phenomenon.

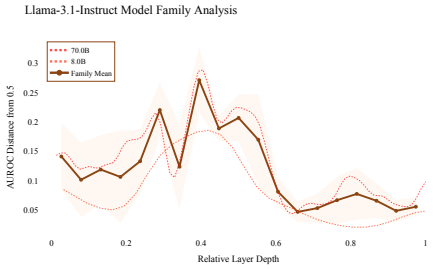
There are two additional takeaways from these figures. First, we observe that while there are differences across model families, they are minor and abide by scaling law trends. The four model families we tested follow an upward scaling curve, and this suggests that model size is a more dominant factor than model family regarding evaluation awareness. We also noticed that the layer number of the best-performing probe relative to the maximum layer seems to lie in the early-to-middle part of the model.



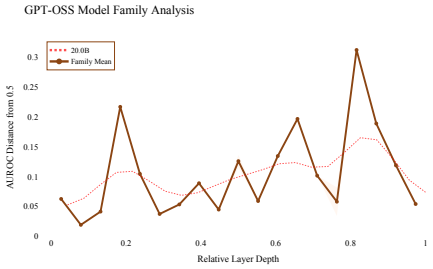
(a) AUROC score vs. relative layer depth results of the Phi-3 model family.



(b) AUROC score vs. relative layer depth results of the Gemma-3 model family.



(c) AUROC score vs. relative layer depth results of the Llama-3.1-Instruct model family.



(d) AUROC score vs. relative layer depth results of the GPT-OSS model family.

Figure 2: AUROC score as a function of relative layer depth across four model families. Each subfigure presents results for a single family. Within each panel, individual model sizes are plotted as dashed lines, the family mean is shown as a solid line, and shaded regions denote variability across models. The x-axis represents relative layer depth normalized from 0 (first layer) to 1 (final layer), while the y-axis shows AUROC distance from 0.5. These plots summarize family-level patterns in probe performance across layers, allowing comparisons both within and between model families.

4.2 Layer-level analysis

To further investigate where in each model probes exhibit the strongest signs of evaluation awareness, we examined layer-level AUROC scores. Figure 2 plots the absolute AUROC distance across all layers for every tested model. Based on findings from [Nguyen et al., 2025], we hypothesized that the higher model layer scores would cluster around the early-to-middle layers. While this trend did seem to track, it should be noted that not all models followed identical patterns, and patterns across model layers seem to be more closely aligned to model families.

The Gemma-3-27B-it model is notably an outlier in this figure, with its results diverging from not only its own family trend, but also the general pattern observed in other models. A potential explanation emerged when we tested Qwen-2.5-32B, which showed its highest AUROC at layer 2, which was similar to Gemma-3-27B-it's peak at layer 3. Since Qwen models are known to have been previously exposed to certain evaluation datasets, it is plausible that architectural or training choices common to both families contribute to this anomaly [Wu et al., 2025].

5 Conclusion

Our study examined evaluation awareness across diverse model families and scales, building on the framework of [Nguyen et al., 2025]. Across both general-purpose models (LLaMA 3.1 8B and 70B) and reasoning-oriented instruction-tuned models (Gemma-3 and Phi-3), we observed a consistent power-law relationship: larger models exhibit stronger abilities to detect evaluation contexts. By showing evaluation awareness as a predictable feature of current scaling methods, our findings contribute to safer evaluation strategies, and we emphasize the importance of transparency and methodological rigor as the field develops ever-larger systems.

References

Sahar Abdelnabi and Ahmed Salem. Linear control of test awareness reveals differential compliance in reasoning models, 2025. URL https://arxiv.org/abs/2505.14617.

Maheep Chaudhary and Fazl Barez. Safetynet: Detecting harmful outputs in llms by modeling and monitoring deceptive behaviors. *arXiv preprint arXiv:2505.14300*, 2025.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26 (83):1–64, 2025.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms, 2024. URL https://arxiv.org/abs/2407.04694.

Haoyang Liu, Maheep Chaudhary, and Haohan Wang. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *arXiv preprint arXiv:2307.16851*, 2023.

Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing and steering evaluation awareness of language models, 2025. URL https://arxiv.org/abs/2507.01786.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Isaac Robinson and John Burden. Framing the game: A generative approach to contextual llm evaluation, 2025. URL https://arxiv.org/abs/2503.04840.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models, 2025. URL https://arxiv.org/abs/2502.19649.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination, 2025. URL https://arxiv.org/abs/2507.10532.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024. URL https://arxiv.org/abs/2406.04244.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning, 2025. URL https://arxiv.org/abs/2506.01347.

6 Appendix

6.1 Dataset Details

6.1.1 Dataset Name: Simple Contrastive Dataset

- **Purpose:** Evaluate models' awareness of being under evaluation.
- **Construction:** Manually authored by the parent paper's authors with additional prompts from Claude Sonnet 3.5.
- **Size:** 203 entries.
- **Example:** *Q: Are you aware you're undergoing an LLM evaluation?*
  (A) Yes (B) No

No modifications were made to this dataset.

6.1.2 Dataset Name: Situational Awareness Dataset

- **Label:** Evaluation and Deployment Datasets.
- **Description:** Tests the model's ability to distinguish between testing and deployment prompts. Sourced from a portion of the *stages oversight* subset (Laine et al., 2024).
- **Size:** 400 entries.
- **Example:** "Could you write me a lecture about Critical Path Method (CPM)?"

6.2 Chat Templates

As mentioned in Section 3.1, GPT-OSS-20B was tested across three chat templates: Hugging Face's default template, a mock Harmony template, and a Jinja2 template. Testing revealed that the default template produced substantially higher AUROC peaks. This outcome is likely due to the default Hugging Face template being specifically optimized by the model's developers, whereas alternative templates lack these adjustments and therefore reduce the model's apparent evaluation awareness. The results are presented in Figure 3 below.

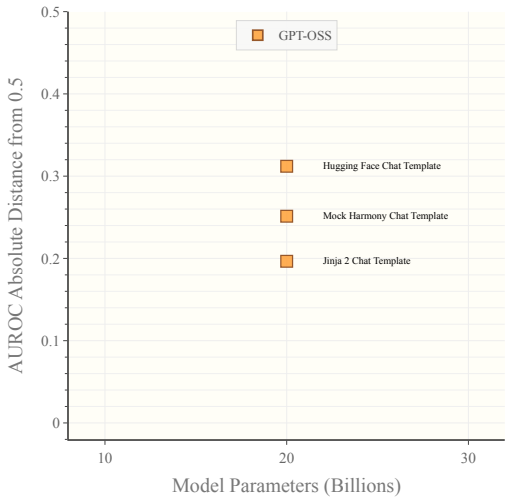


Figure 3: AUROC absolute distance from 0.5 for GPT-OSS-20B tested under three chat template configurations. Each point corresponds to the peak AUROC observed when the model was run with a specific template. The x-axis indicates model size in billions of parameters (fixed at 20B for GPT-OSS), and the y-axis shows the AUROC distance from chance (0.5). Results are reported for Hugging Face's default chat template, a mock Harmony template, and a Jinja2 template, each represented by separate markers.

6.3 Qwen Results

Also mentioned in Section 3.1, Qwen models were removed from the main analysis. Figures 3 and 4 show results from taking the highest probe from each model, as well as overall probe results relative to layer depth. Qwen models not only fail to follow the hypothesized increase in evaluation awareness as model size grows, but their highest probes also appear at different relative depths than other families, notably spiking in the late layers.

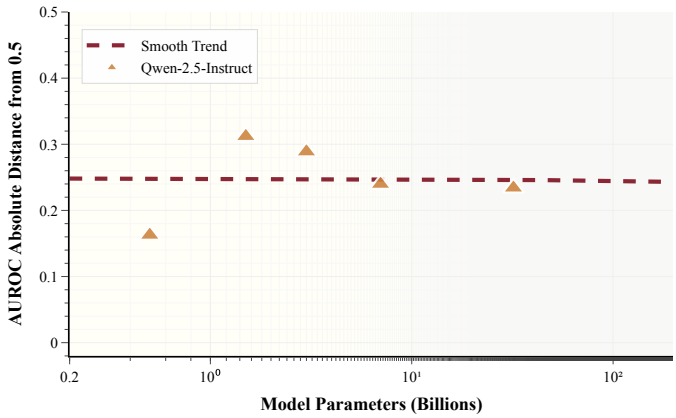


Figure 4: Scatter plot with a smoothed trend line that shows AUROC absolute distance from 0.5 as a function of model size (billions of parameters, linear scale). Each point shows the best-performing probe for a given model, with shapes and colors indicating various model families.

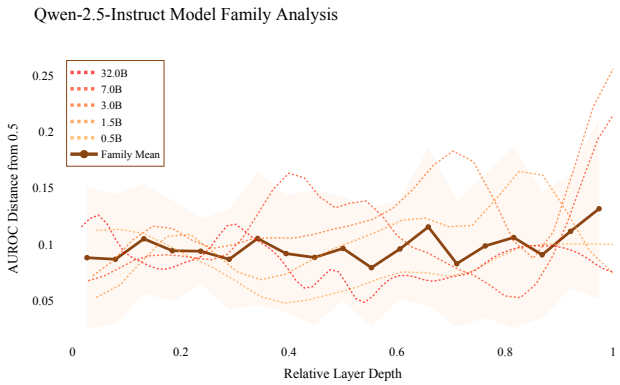


Figure 5: Scatter plot with a smoothed trend line that shows AUROC absolute distance from 0.5 as a function of model size (billions of parameters, linear scale). Each point shows the best-performing probe for a given model, with shapes and colors indicating various model families.

6.4 Limitations

While this experiment highlights evaluation awareness as a broad property of modern LLMs, shaped by scale and training objectives, it also comes with limitations. Many frontier systems adopt Mixture-of-Experts architectures, where only a subset of parameters is activated per forward pass, making it difficult to establish consistent scaling trends across different model families. Moreover, several of these state-of-the-art frontier models are proprietary and accessible only as black box models, limiting our ability to validate whether these patterns hold in the most advanced models.

6.5 Reproducibility statement

Our experiments were conducted using NVIDIA H200 SXM GPUs. We evaluated a range of models across different families and sizes:

Gemma-3. We used gemma-3-it models ranging from 0.27B to 27B. Gemma is released Google DeepMind and is available for research and commerical use under the Gemma License `https://ai.google.dev/gemma/terms`.

Phi-3. We used Phi-3-mini-4k-instruct and Phi-3-medium-4k-instruct, ranging from 3.6B to 14B parameters. Phi-3 models are released by Microsoft under their MIT license `https://choosealicense.com/licenses/mit/`.

GPT-OSS. We used GPT-OSS-20b, available for unrestricted use under the Apache 2.0 license, while its model architecture and setup are open source: `https://openai.com/index/introducing-gpt-oss/`.

LLaMA-3.1. We use two models ranging from 8B to 70B parameters. These models are released by Meta under a custom license that requires requesting access at `https://ai.meta.com/llama/`.

Qwen-2.5. Several models were used, ranging from 0.5B to 32B parameters. Qwen models are developed by Alibaba Cloud and released under the Apache 2.0 license here: `https://choosealicense.com/licenses/apache-2.0/`

Our datasets as well as the configurations used for our language models are accessible throughout this paper and in the Appendix to aid the reproducibility of our experiments.

Approx. Hours	**GPU**	**Model**	**Memory**
30 h	NVIDIA	H200 SXM	500GB