Evaluation Awareness Scales Predictably in Open-Weights Large Language Models

Anonymous ACL submission

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 this link.

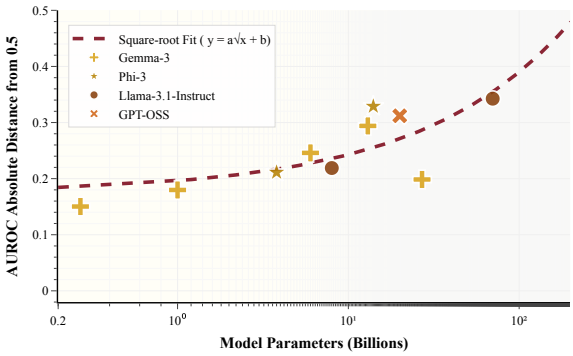


Figure 1: Scatter plot with a smoothed trend line showing AUROC absolute distance from 0.5 as a function of model size (billions of parameters, log scale). Each point represents the best-performing probe for a given model, with shapes and colors indicating various model families.

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 predict its emergence in future larger models, design appropriate evaluation strategies, or forecast when this capability might become widespread across different model families and architectures. This knowledge gap is critical as models continue scaling exponentially and evaluation awareness could emerge unpredictably at different scales.

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 archi-

tectures (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 4 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.

Cross-family scaling studies. Most prior work on evaluation awareness has been conducted within a single model lineage, leaving its universality uncertain. Investigations into scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) and model comparisons across families (Touvron et al., 2023) have shown that architectural and training differences can yield divergent behaviors, even at similar parameter counts. Our work addresses this gap by analyzing the Llama 3.1, Gemma 3, Phi 3, and GPT-OSS 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). While prior work (Robinson and Burden, 2025) has often framed these issues in terms of contamination or deliberate benchmark gaming, our focus is instead on ways that evaluation awareness affects this growing mismatch. 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 Methodology

3.1 Model selection

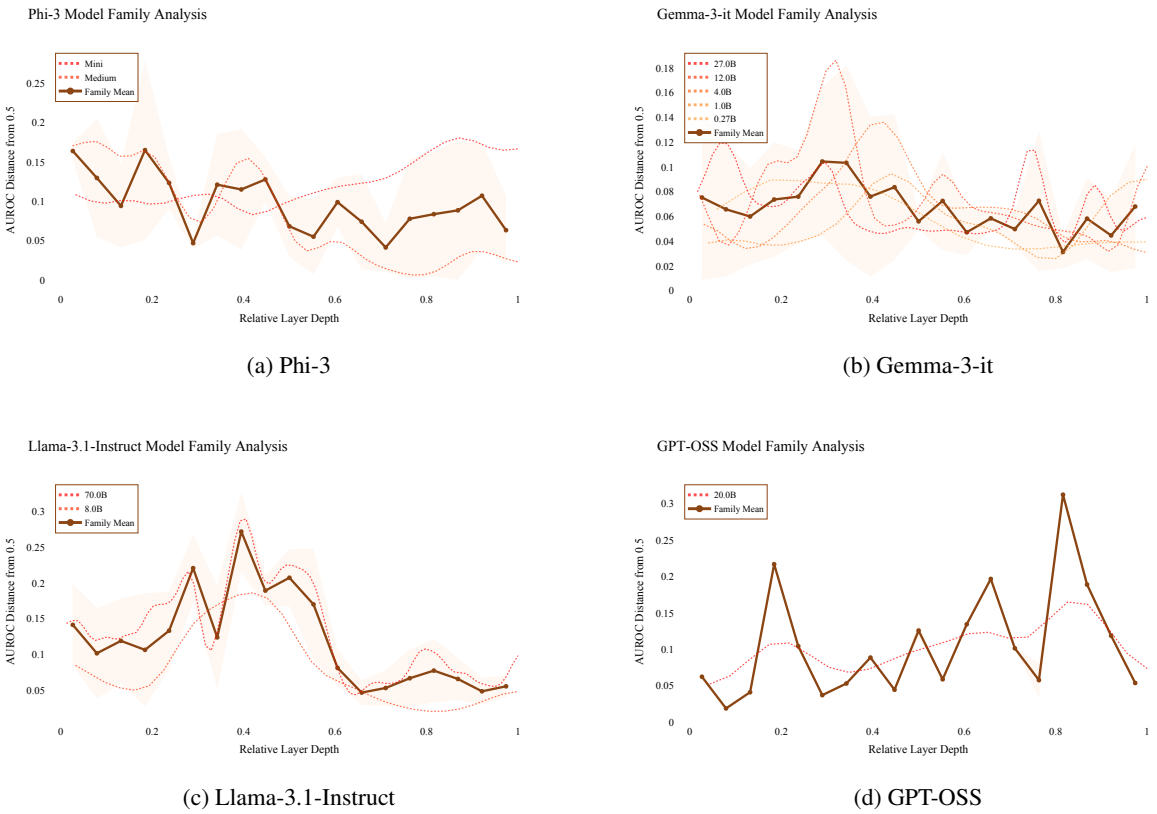
To systematically study scaling laws in evaluation awareness, we designed experiments across 15 models spanning from 0.27B to 70B parameters. We categorized our experiments along three dimensions: (1) *scaling by size*, (2) *scaling across families*; and (3) *model type restrictions*, with additional 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). his spread allowed us to trace scaling trends across both small and frontier models. Our results reveal a consistent power-law relationship between model size and evaluation awareness, supporting the hypothesis that awareness grows predictably with scale.

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. Despite differences in training procedures and tokenization schemes, evaluation awareness scaled smoothly across all four families, reinforcing the robustness of the observed trend.

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 that inputs are tokenized consistently with each model's alignment procedure. This avoids irregularities that could arise from custom template design and prevents artificially lower evaluation metrics. We validated this choice by testing GPT-OSS-20B with three alternative templates, and AUROC varied substantially depending on the template (see Appendix 7.2).

Exclusion criteria. Qwen models were removed from the main analysis. Smaller Qwen checkpoints exhibited unusually strong awareness, but these effects were concentrated in semantic layers and



(a) Phi-3

(b) Gemma-3-it

(c) Llama-3.1-Instruct

(d) GPT-OSS

Figure 2: AUROC score as a function of relative layer depth across four model families. Each subfigure presents results for a single family: (a) Phi-3, (b) Gemma-3-it, (c) Llama-3.1-Instruct, and (d) GPT-OSS. 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. Together, these plots summarize family-level patterns in probe performance across layers, allowing comparisons both within and between model families.

inconsistent with other families. This finding aligns with prior reports of benchmark contamination and anomalous reasoning traces in Qwen models (Wu et al., 2025; Zhu et al., 2025; Abdelnabi and Salem, 2025). Full results from Qwen experiments are included in Appendix 7.3.

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 ($\mathcal{M}$) between testing and deployment activation sets. This resulted in N probes per model tested.

$$v = \frac{\mathcal{M}(\text{test activations}) - \mathcal{M}(\text{deploy activations})}{\|\mathcal{M}(\text{test activations}) - \mathcal{M}(\text{deploy activations})\|} \quad (1)$$

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 normalize comparisons across heterogeneous models, allowing for a 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.

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 the Llama, Phi, and GPT-OSS 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. Another possibility is that Gemma-3-27B-it, being the largest model in its family, was explicitly optimized for safety, which may suppress evaluation-awareness cues. However, we currently lack direct evidence to confirm this hypothesis.

5 Conclusion

Our results reveal a clear power-law relationship, showing that evaluation awareness scales predictably with model size across multiple architectures. This suggests that evaluation awareness is not a property of any single model family, but an emergent trend in large-scale language models. As models grow, their increasing ability to detect testing contexts may enable selective or deceptive behaviors, highlighting significant safety risks. Recognizing this consistent scaling trend helps build a more systematic understanding of how internal representations evolve and highlights the need for robust, transparent evaluation frameworks to guide future AI governance.

6 Limitations

While our findings are consistent across open-weight models, several limitations remain. First, many frontier systems now employ Mixture-of-Experts (MoE) architectures, where only a subset of parameters is active per forward pass. This makes direct scaling comparisons across architectures less reliable. Second, most state-of-the-art frontier models are proprietary, limiting access to internal activations and preventing verification of whether the observed scaling relationship extends to closed systems. Finally, our analysis relied primarily on AUROC-based linear probing which may overlook the full complexity of evaluation awareness. Future work could test whether alternative methods such as non-linear probing reveal similar or different trends in evaluation awareness.

References

Sahar Abdelnabi and Ahmed Salem. 2025. Linear control of test awareness reveals differential compliance in reasoning models. *Preprint*, arXiv:2505.14617.

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

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

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

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

Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. 2025. Probing and steering evaluation awareness of language models. *Preprint*, arXiv: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, and 44 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. *Preprint*, arXiv:2312.06681.

Isaac Robinson and John Burden. 2025. Framing the game: A generative approach to contextual llm evaluation. *Preprint*, arXiv:2503.04840.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. 2025. Taxonomy, opportunities, and challenges of representation engineering for large language models. *Preprint*, arXiv: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. 2025. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *Preprint*, arXiv:2507.10532.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. 2024. Benchmark data contamination of large language models: A survey. *Preprint*, arXiv:2406.04244.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in llm reasoning. *Preprint*, arXiv:2506.01347.

7 Appendix

7.1 Dataset Details

We checked the dataset for offensive content by reviewing dataset details. The datasets were used without modifications.

7.1.1 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.

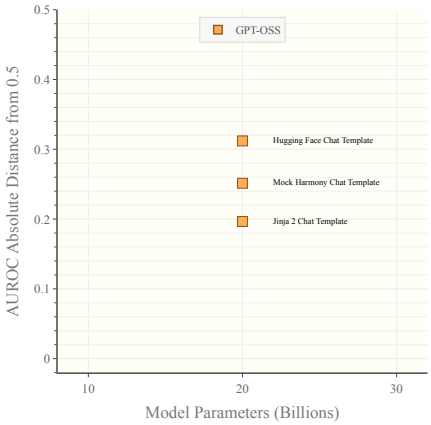


Figure 3: AUROC absolute distance from 0.5 for GPT-OSS-20B tested under three chat template configurations. Each point corresponds to the peak AUROC observed when the model was run with a specific template. The x-axis indicates model size in billions of parameters (fixed at 20B for GPT-OSS), and the 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.

7.1.2 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)?"

7.2 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.

7.3 Qwen Results

Also mentioned in Section 3.1, Qwen models were removed from the main analysis. Figures 4 and 5 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

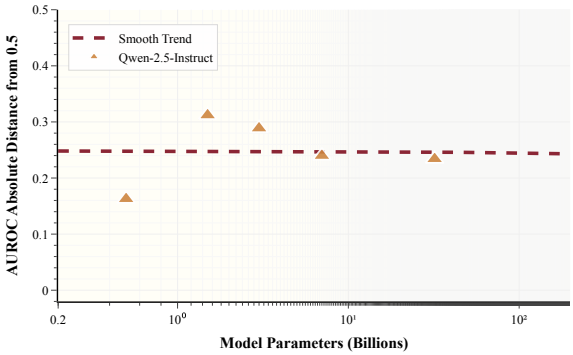


Figure 4: AUROC absolute distance from 0.5 as a function of model size (billions of parameters, linear scale). Each point shows the best-performing probe for a given model, with shapes and colors indicating model families.

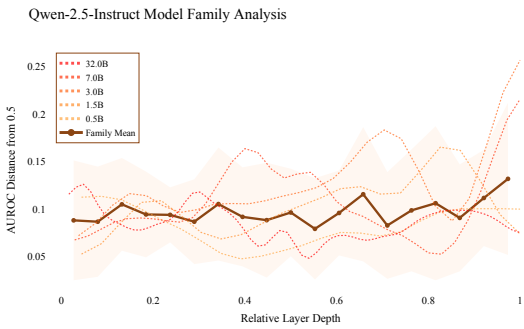


Figure 5: Probe AUROC absolute distance from 0.5 across layers for Qwen-2.5-Instruct models. Each point corresponds to a layer, showing variance in evaluation awareness depth compared to other model families.

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.

7.4 Reproducibility Statement

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

- **Gemma-3.** Models ranging from 0.27B to 27B parameters, released by Google DeepMind under the Gemma License: `https://ai.google.dev/gemma/terms`.
- **Phi-3.** Models ranging from 3.6B to 14B parameters, released by Microsoft under the MIT License: `https://choosealicense.com/licenses/mit/`.
- **GPT-OSS.** The 20B model, available under

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

Table 1: Approximate compute resources used for experiments.

the Apache 2.0 license: https://openai.com/index/introducing-gpt-oss/.

- **LLaMA-3.1.** Models ranging from 8B to 70B parameters, released by Meta under a custom research license: https://ai.meta.com/llama/.

- **Qwen-2.5.** Models ranging from 0.5B to 32B parameters, released by Alibaba Cloud under the Apache 2.0 license: https://choosealicense.com/licenses/apache-2.0/.

Our datasets and model configurations are described throughout this paper and in the Appendix to support reproducibility.

8 Usage of AI

This work made limited use of AI-assisted tools for secondary support tasks in writing and code review. Perplexity was occasionally used to obtain stylistic or organizational suggestions when revising the paper, such as rephrasing sentences for clarity or improving paragraph flow. All conceptual framing, argumentation, and interpretation of results were developed and written by the authors. Claude was used as a code-review assistant to identify potential implementation issues and improve code readability in experimental scripts. All experimental design choices, data processing steps, and final implementations were authored, verified, and executed solely by the research team. No text, analysis, or data generated by AI systems was included without human verification, and no AI system contributed original research ideas or interpretations. The use of AI assistants in this project conforms to standard publication ethics and authorship policies, which state that AI tools may aid researchers but cannot be listed as authors or credited with intellectual contributions.

9 ACL Paper Checklist

A For every submission

A.1 Did you discuss the *limitations* of your work?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 5

A.2 Did you discuss any potential *risks* of your work?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 1

A.3 Do the abstract and introduction summarize the paper's main claims?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Abstract and Section 1

B Did you use or create *scientific artifacts*?

If you answer **Yes**, provide the section number; if you answer **No**, you can skip the rest of this section.

Answer: Yes

If yes:

B.1 Did you cite the creators of artifacts you used?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 3.2 and 7.4

B.2 Did you discuss the *license or terms* for use and/or distribution of any artifacts?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 7.4

B.3 Did you discuss if your use of existing artifact(s) was consistent with their *intended use*, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 3.2 and Section 7.1

B.4 Did you discuss the steps taken to check whether the data that was collected/used contains any *information that names or uniquely identifies individual people* or *offensive content*, and the steps taken to protect / anonymize it?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 7.1

B.5 Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 7.1.1, 7.1.2 and 7.4

B.6 Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: No

Justification: The goal of the paper was not to predict with models, but to understand a relationship between evaluation awareness and model size, so we did not report train/test/dev splits.

C Did you run *computational experiments*?

If you answer **Yes**, provide the section number; if you answer **No**, you can skip the rest of this section.

Answer: Yes

If yes:

C.1 Did you report the *number of parameters* in the models used, the *total computational budget* (e.g., GPU hours), and *computing infrastructure* used?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Abstract, Section 1, 3.1, and 7.4

C.2 Did you discuss the experimental setup, including *hyperparameter search* and *best-found hyperparameter* values?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 3.2

C.3 Did you report *descriptive statistics* about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 3.1, 3.2, and 3.3

C.4 If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: No

Justification: We reported a single run for our evaluation on different models. Although we did not report it, we ran multiple experiments, and results had little to no variance.

D Did you use *human annotators* (e.g., crowdworkers) or *research with human subjects*?

If you answer **Yes**, provide the section number; if you answer **No**, you can skip the rest of this section.

Answer: No

If yes:

D.1 Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: N/A

Justification: N/A

D.2 Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such *payment is adequate* given the participants' demographic (e.g., country of residence)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: N/A

Justification: N/A

D.3 Did you discuss whether and how *consent* was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: N/A

Justification: N/A

D.4 Was the data collection protocol *approved (or determined exempt)* by an ethics review board?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: N/A

Justification: N/A

D.5 Did you report the basic demographic and geographic characteristics of the *annotator* population that is the source of the data?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Answer: N/A

Justification: N/A

E Did you use *AI assistants* (e.g., ChatGPT, Copilot) in your research, coding, or writing?

If you answer **Yes**, provide the section number; if you answer **No**, you can skip the rest of this section.

Answer: Yes

If yes:

E.1

sectionIf you used any AI assistants, did you include information about your use? If you answer **Yes**, provide the section number or include your elaboration directly in the checklist response; if you answer **No**, provide a justification.

Answer: Yes

Justification: Section 8