The Impact of Model Size on Detection of LLM Hallucinations

Matt Boerio and Paul Groth
University of Amsterdam
boeriomatt@gmail.com p.t.groth@uva.nl

Abstract

Large Language Models have radically improved the capabilities of natural language generation. However, LLMs may generate hallucinations. To begin to tackle this challenge, work has emerged on hallucination detection. Currently, hallucination detectors are often compared across models that differ simultaneously in size, architecture, and adaptation, making it unclear where performance derives from. In particular, the role of model size is underexplored where having small models may be beneficial in application settings. To address this gap, we systematically investigate hallucination detection across four open-weight model families spanning encoder-based natural language inference, sequence-to-sequence instruction following, and decoder-only instruction models. Within each family, we evaluate multiple parameter scales both out of the box and after task-specific training. Out-of-the-box scaling is strongly architecture-dependent. After training, size-performance relationships become more consistently positive across all four families. On the SHROOM hallucination detection benchmark, Qwen-7B achieves the strongest overall result, with 0.845 accuracy, while substantially smaller DeBERTa models remain competitive. These results show that parameter count alone is an incomplete basis for selecting hallucination detectors: the benefits of scaling depend on architecture and task-specific adaptation.

1 Introduction

Large language models can produce fluent outputs that are nevertheless unsupported by their input or inconsistent with verifiable information (Huang et al., 2025). Automatically detecting such hallucinations is therefore important for the reliable use of natural language generation systems. Existing detectors include natural language inference (NLI) classifiers (Huang et al., 2025), prompted language-model judges or uncertainty-based methods (Valentin et al., 2024), and models fine-tuned specifically for hallucination detection (Arteaga et al., 2024).

Despite growing interest in these approaches, the effect of detector size remains unclear. Larger models are often assumed to provide stronger semantic representations or reasoning capabilities, but published comparisons usually vary several factors simultaneously, including architecture, pretraining, prompting, additional supervision, and ensembling (Mickus et al., 2024). Consequently, stronger performance from a larger system cannot necessarily be attributed to its parameter count. This issue is especially relevant in practical settings, where larger detectors impose greater training and inference costs.

We study whether hallucination-detection performance improves consistently with model size (i.e., parameter count) on the model-agnostic track of the SHROOM SemEval task for hallucination detection (Mickus et al., 2024). We compare multiple size rungs within four open-weight model families: DeBERTa NLI cross-encoders (He et al., 2023; Sentence Transformers, 2024), FLAN-T5 (Chung et al., 2022), Gemma (Gemma Team, 2025), and Qwen (Qwen Team, 2025). Each model is evaluated both out of the box (OOTB) and after task-specific Low-Rank Adaptation (LoRA) (Hu et al., 2021). The dataset, input formulation, evaluation procedure, and adaptation settings are held constant wherever possible, allowing us to separate within-family scaling from differences between architectures and training regimes.

We make three contributions: first, we provide a controlled comparison of multiple parameter scales within four detector families. Second, we conduct paired OOTB and LoRA-adapted evaluations. Third, we show that scaling benefits are conditional on architecture and adaptation, with smaller NLI encoders remaining competitive with substantially

Proceedings of the 19th International Natural Language Generation Conference: Short Papers, pages 25–34
October 17-21, 2026 ©2026 Association for Computational Linguistics

larger generative judges.

2 Related Work

Hallucination detection is commonly formulated as a semantic-consistency problem between a generated output and its source, reference, or supporting evidence (Huang et al., 2025). NLI-based approaches estimate whether the output is entailed by the available context, while prompted language-model judges directly assess support or contradiction. Other methods combine uncertainty signals (Valentin et al., 2024), repeated generations (Manakul et al., 2023), external tools (Cheng et al., 2024), or specialized fine-tuning (Arteaga et al., 2024). These approaches differ substantially in computational cost and in the extent to which their training objectives align with hallucination detection.

The SHROOM shared task provides a standardized benchmark for detecting unsupported semantic content across definition modeling, machine translation, and paraphrase generation (Mickus et al., 2024). Its submissions include large proprietary judges, NLI-based systems, ensembles, and specialized fine-tuned models. Competitive results from smaller fine-tuned systems demonstrate that model scale is neither necessary nor sufficient for strong performance (Wei et al., 2024). Related evaluations likewise show that increasing model scale does not uniformly improve hallucination-related performance (Sun et al., 2025).

However, prior comparisons do not hold architecture and adaptation constant while varying parameter count within a family. Model size is typically one component of a broader system comparison rather than the main experimental variable. We therefore focus on within-family scaling under a shared benchmark and evaluate whether the relationship changes after task-specific adaptation.

3 Experimental Setup

3.1 Dataset

We use the model-agnostic track of SHROOM (Mickus et al., 2024), which frames hallucination detection as binary classification over generated outputs. The benchmark contains examples from definition modeling, machine translation, and paraphrase generation. Each example provides a relevant source or reference, a generated hypothesis, a majority binary label, and a soft hallucination score $p(\text{Hallucination})$ derived from annotator judgments.

The labeled development set contains 499 examples, and the released model-agnostic test set contains 1,500 examples. During model development, we split the development data into 399 training and 100 validation examples using a fixed random seed. Hyperparameters and epoch counts were selected using this split. Final adapted models were then trained on all 499 development examples and evaluated once on the test set, without test-based model selection. We do not use the additional unlabeled SHROOM data, since varying the amount or method of weak supervision would introduce another experimental factor.

For every architecture, the relevant context is paired with the generated hypothesis. We apply no resampling or class balancing and tokenize inputs using each model's native tokenizer. Fine-tuning inputs are truncated to a maximum length of 512 tokens.

3.2 Models and OOTB Scoring

We evaluate four model families that provide multiple parameter scales and represent three common detector architectures. The DeBERTa cross-encoder family contains NLI encoders from the xsmall through large rungs (He et al., 2023; Sentence Transformers, 2024). FLAN-T5 represents sequence-to-sequence instruction models at the small, base, large, and XL rungs (Chung et al., 2022). Gemma-3 instruction models are evaluated at 270M, 1B, and 4B parameters (Gemma Team, 2025), and Qwen2.5-Instruct is evaluated at 0.5B, 1.5B, 3B, and 7B parameters (Qwen Team, 2025). Model size refers to the total base-model parameter count.

For OOTB DeBERTa evaluation, we use the model's three-way softmax distribution over entailment, neutral, and contradiction. We interpret entailment as evidence that the hypothesis is supported and compute hallucination probability as $p(\text{neutral}) + p(\text{contradiction}) = 1 - p(\text{entailment})$, with no additional renormalization. For FLAN-T5, Gemma, and Qwen, we use the same support-oriented prompt and a constrained yes/no verbalizer. "Yes" denotes that the hypothesis is supported, while "no" denotes hallucination. We use the models' probabilities for the relevant answer tokens to obtain continuous hallucination scores rather than relying on unconstrained text generation. Because the NLI and generative families use different native scoring formulations, cross-architecture OOTB

comparisons should be interpreted more cautiously than within-family size comparisons.

3.3 LoRA Adaptation

We adapt each model checkpoint using Low-Rank Adaptation (Hu et al., 2021). LoRA freezes the pretrained parameters and inserts trainable low-rank matrices into attention projections, making it possible to apply a broadly consistent adaptation procedure across architectures.

Models are trained against the soft SHROOM hallucination scores rather than only the majority labels. This preserves information about annotator disagreement and directly supports the ranking-based evaluation metric. We use a shared configuration of rank $r = 16$, LoRA scaling parameter $\alpha = 32$, dropout 0.05, learning rate 2×10^{-4}, and effective batch size 4. DeBERTa and FLAN-T5 are trained for four epochs, while Gemma and Qwen are trained for three. We adapt attention projection modules in each family and use non-quantized model loading throughout.

The shared setup is intended to support controlled comparison rather than maximize each checkpoint independently. However, a common adaptation recipe may not be equally well suited to every architecture. Cross-family LoRA comparisons should therefore be interpreted as performance under this shared configuration rather than as estimates of each architecture's independently optimized performance. All final experiments use a fixed NVIDIA A100 40GB environment and random seed 42.

3.4 Metrics

We report accuracy against the majority binary label and Spearman's correlation ρ between predicted hallucination probabilities and the soft gold scores. Binary predictions use a threshold of 0.5. Spearman's ρ is particularly informative for SHROOM because it evaluates whether models rank examples consistently with the degree of annotator agreement. Inference latency and fine-tuning time are recorded as secondary efficiency measures and reported in the supplementary material.

To quantify uncertainty arising from the finite evaluation sample, we additionally use a paired, task-stratified non-parametric bootstrap over the 1,500 SHROOM test examples. Using random seed 42, for each of 10,000 bootstrap replicates, examples are resampled with replacement separately within definition modeling, machine translation, and paraphrase generation, preserving the original task counts of 563, 562, and 375 examples, respectively. The same sampled examples are used across compared checkpoints, after which accuracy and Spearman's ρ are recomputed. We report percentile 95% bootstrap confidence intervals for paired performance differences.

4 Results

Figure 1 shows the relationship between model size and Spearman's ρ before and after LoRA adaptation. Table 1 reports the strongest checkpoint in each family under both conditions. Complete model-level accuracy, correlation, and efficiency results are provided in the supplementary material.

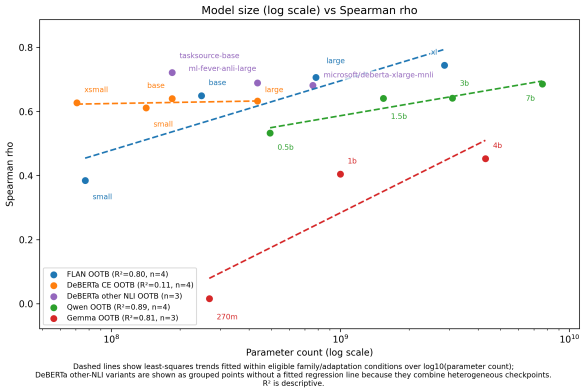
4.1 Out-of-the-Box Scaling

OOTB scaling differs substantially by architecture. FLAN-T5 exhibits the clearest positive relationship: accuracy increases from 0.657 for the small model to 0.827 for XL, while Spearman's ρ increases from 0.384 to 0.744. Gemma also improves with size, rising from 0.593 accuracy and $\rho = 0.015$ at 270M parameters to 0.723 accuracy and $\rho = 0.452$ at 4B.

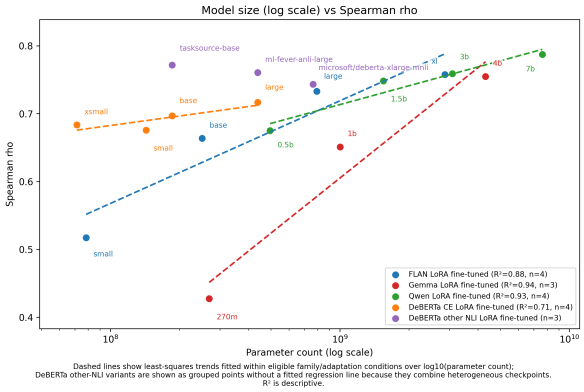
Qwen improves from $\rho = 0.533$ at 0.5B to $\rho = 0.686$ at 7B, but its accuracy is not monotonic because the 3B checkpoint underperforms adjacent sizes. As examined in Section 4.4, the OOTB 1.5B-to-3B accuracy decrease is robust to test-set resampling, while the corresponding difference in ρ is near zero, indicating that the anomaly is concentrated in thresholded classification rather than ranking performance. DeBERTa shows the flattest OOTB relationship: accuracy remains between 0.743 and 0.773, and ρ remains between 0.611 and 0.640. Thus, increasing parameter count is often beneficial within a family, but does not produce a uniform OOTB scaling pattern.

Family	Best OOTB checkpoint (Acc./ρ)	Best LoRA checkpoint (Acc./ρ)
DeBERTa CE	base (0.769/0.640)	large (0.813/0.717)
FLAN-T5	XL (0.827/0.744)	XL (0.829/0.758)
Gemma	4B (0.723/0.452)	4B (0.828/0.755)
Qwen	7B (0.805/0.686)	**7B (0.845/0.787)**

Table 1: Best checkpoint in each controlled scaling family, selected by Spearman's ρ separately within each evaluation condition.



(a) Out-of-the-box



(b) LoRA-adapted

Figure 1: Spearman's ρ versus total parameter count for out-of-the-box models and LoRA-adapted models. Fine-tuning makes the within-family relationships more consistently positive, with Qwen-7B achieving the strongest adapted result. Dashed least-squares fits summarize only the three or four observed checkpoints in each family and are shown for descriptive visualization; they are not used for statistical inference or interpreted as general scaling laws. Purple points show supplementary heterogeneous DeBERTa NLI checkpoints and are excluded from the fitted trends.

4.2 Scaling after LoRA Adaptation

After LoRA adaptation, scaling trends become more consistently positive. FLAN-T5 improves from 0.711 accuracy and $\rho = 0.517$ at the small rung to 0.829 accuracy and $\rho = 0.758$ at XL. Gemma improves from 0.701 accuracy and $\rho = 0.427$ at 270M to 0.828 accuracy and $\rho = 0.755$ at 4B. Qwen-7B achieves the strongest overall result, with 0.845 accuracy and $\rho = 0.787$. DeBERTa also exhibits a clearer positive trend after adaptation, increasing from 0.781 accuracy and $\rho = 0.683$ at the xsmall rung to 0.813 accuracy and $\rho = 0.717$ at the large rung.

LoRA improves the point estimates of both accuracy and ρ for every paired checkpoint. The average correlation gain is largest for Gemma at $+0.320$, followed by Qwen at $+0.117$, DeBERTa at $+0.066$, and FLAN-T5 at $+0.047$. The corresponding average accuracy gains are $+0.115$, $+0.102$, $+0.034$, and $+0.025$, respectively. Paired bootstrap analysis shows that the ρ gain has a 95% interval excluding zero for all 15 controlled checkpoints, whereas accuracy gains are less uniform at already strong FLAN-T5 checkpoints. Adaptation therefore has the greatest effect on the decoder-only families, particularly Gemma.

4.3 Architecture Comparison

FLAN-T5-XL is the strongest OOTB checkpoint among the controlled families, reaching 0.827 accuracy and $\rho = 0.744$. After adaptation, Qwen-7B becomes the strongest model overall. However, model size alone does not determine the ranking. The fine-tuned DeBERTa-large model reaches 0.813 accuracy and $\rho = 0.717$ despite being substantially smaller than Qwen-7B and FLAN-T5-XL. This result indicates that an architecture pretrained for sentence-pair inference can remain competitive with much larger generative judges when the evaluation task is closely aligned with semantic entailment.

For context, the original SHROOM model-agnostic leader board reported best values of 0.847 accuracy and 0.770 Spearman's ρ (Mickus et al., 2024), compared with 0.845 accuracy and $\rho = 0.787$ for our LoRA-adapted Qwen-7B. The comparison is not strictly like-for-like, however, because shared-task systems differed in supervision, ensembling, model access, and system-level optimization, whereas our experiments intentionally constrain these factors to study model size and adaptation.

4.4 Robustness Analysis

Table 2 reports paired task-stratified bootstrap intervals for the difference in Spearman's ρ between the smallest and largest evaluated checkpoint within each controlled family. Seven of the eight OOTB/LoRA comparisons have 95% intervals excluding zero. The exception is OOTB DeBERTa, for which the endpoint difference is $+0.005$ with a 95% interval of $[-0.020, 0.031]$, consistent with the flat OOTB pattern observed in Figure 1. After LoRA adaptation, all four families show positive

smallest-to-largest differences with intervals excluding zero. These results support the descriptive observation that within-family size relationships become more consistently positive after adaptation without treating the three or four observed size rungs as evidence of a general scaling law.

The Qwen-3B result further illustrates the distinction between thresholded accuracy and ranking performance. OOTB accuracy decreases by 0.109 from Qwen-1.5B to Qwen-3B, with a 95% interval of $[-0.144, -0.074]$, while the corresponding change in ρ is $+0.001$, with an interval of $[-0.027, 0.029]$. Thus, the 3B anomaly is robust for thresholded classification but is not accompanied by a comparable decline in ranking performance. After LoRA, the 1.5B-to-3B differences in both accuracy and ρ include zero. An ordinary paired bootstrap without task stratification produced nearly identical intervals and the same qualitative conclusions.

Family	OOTB $\Delta\rho$	LoRA $\Delta\rho$
DeBERTa CE	$+0.005$ [-0.020, 0.031]	$+0.033$ [0.011, 0.055]
FLAN-T5	$+0.360$ [0.317, 0.404]	$+0.241$ [0.205, 0.277]
Gemma	$+0.437$ [0.374, 0.497]	$+0.327$ [0.286, 0.370]
Qwen	$+0.153$ [0.117, 0.190]	$+0.112$ [0.088, 0.137]

Table 2: Smallest-to-largest checkpoint differences (largest minus smallest) in Spearman's ρ within each controlled family. Point estimates are shown above their 95% paired task-stratified bootstrap confidence intervals over the SHROOM test set.

5 Discussion and Conclusion

On the SHROOM model-agnostic benchmark and within the tested parameter ranges, the results show that model size is relevant to hallucination detection, but its effect is conditional on architecture and adaptation. FLAN-T5 already exhibits strong OOTB scaling, suggesting that instruction-tuned sequence-to-sequence models can provide useful support judgments without task-specific training. Gemma and Qwen show less reliable OOTB behavior but substantially larger improvements after LoRA adaptation. Their pretrained representations may therefore contain useful semantic information that is not mapped reliably to hallucination probabilities by the fixed verbalizer alone.

The larger gains in Spearman's ρ are consistent with interpreting LoRA primarily as task calibration. Adaptation improves the ordering of examples by hallucination likelihood, which is especially relevant because SHROOM contains soft labels reflecting annotator disagreement. This interpretation should not be taken to imply that adaptation has identical effects across architectures: FLAN-T5 begins from a comparatively strong OOTB baseline, whereas the decoder-only models have more room to improve.

The DeBERTa results demonstrate that architecture-task alignment can offset a substantial parameter disadvantage. A smaller NLI encoder is explicitly trained to compare sentence pairs and therefore matches the support-detection structure of SHROOM. Although Qwen-7B obtains the highest adapted score, DeBERTa remains competitive while requiring substantially less training and inference time in our shared hardware environment. The appropriate detector consequently depends on whether the priority is maximum benchmark performance or a more efficient specialized model.

Overall, on SHROOM, increasing model size tends to improve hallucination detection within the tested families, especially after task-specific adaptation, but parameter count is not a sufficient model-selection criterion. Future work should test whether these relationships persist for larger checkpoints and additional benchmarks.

Limitations

This study has several limitations. First, each model family contains only three or four evaluated size rungs. The plotted relationships should therefore be interpreted as descriptive trends over the tested checkpoints rather than as evidence of a general scaling law. In particular, the experiments do not include substantially larger decoder-only checkpoints, such as the 14B, 32B, and 72B variants available in the Qwen2.5 family.

Second, all experiments use the SHROOM model-agnostic benchmark. SHROOM covers three natural language generation tasks and provides useful soft labels, but the results may not generalize to long-form factuality, dialogue, retrieval-augmented generation, or domain-specific hallucinations.

Third, the LoRA configuration is intentionally controlled rather than optimized for every checkpoint. This improves comparability but may understate the maximum attainable performance of par-

ticular models or architectures, and the shared adaptation recipe may be better suited to some model families than others. Cross-family LoRA comparisons should therefore be interpreted as conditional on this common configuration. Each adapted checkpoint was also trained with a single random seed, so the study does not estimate stochastic variation across training runs. The paired task-stratified bootstrap confidence intervals reported in Section 4.4 instead quantify uncertainty arising from the finite SHROOM test sample and should not be interpreted as training-seed confidence intervals.

Fourth, the OOTB scoring formulations are architecture-specific. DeBERTa maps its three-way NLI distribution to hallucination probability by summing neutral and contradiction probability, while FLAN-T5, Gemma, and Qwen use token probabilities from a support-oriented yes/no verbalizer. This introduces an additional consideration when comparing architectures directly, although within-family size comparisons use the same scoring formulation across checkpoints.

Ethical Considerations

This work uses an existing publicly released benchmark and does not introduce new human-participant data collection. Hallucination detectors remain imperfect and should not be interpreted as guarantees that generated content is factual or safe. In high-stakes domains, detector outputs should support rather than replace human review and domain-specific verification. The study also involves computational costs from evaluating and adapting multiple model sizes; these costs were limited by using parameter-efficient fine-tuning and a fixed experimental setup.

Data Availability

The SHROOM dataset is publicly available from its original authors. The experiments use the released model-agnostic development and test data. Code, experiment configurations, prediction outputs, and the bootstrap robustness-analysis script are available in our public GitHub repository.

References

Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. 2024. Hallucination detection in llms: Fast and memory-efficient fine-tuned models. *Preprint*, arXiv:2409.02976.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. *Preprint*, arXiv:2406.11277.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in Large Language Models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2).

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sentence Transformers. 2024. cross-encoder/nli-deberta-v3-base. Hugging Face model card. Accessed 2026-06-03.

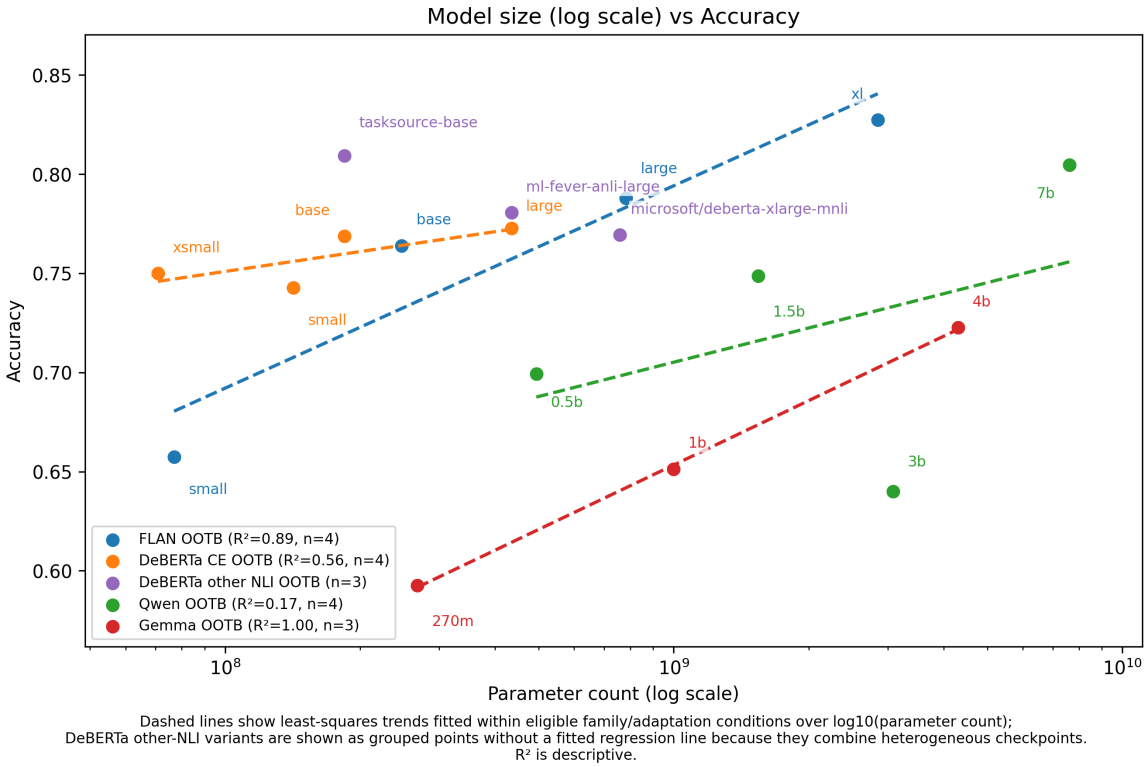
Chongren Sun, Yuran Li, Di Wu, and Benoit Boulet. 2025. OnionEval: An unified evaluation of fact-conflicting hallucination for small-large language models. *Preprint*, arXiv:2501.12975.

Simon Valentin, Jinmiao Fu, Gianluca Detommaso, Shaoyuan Xu, Giovanni Zappella, and Bryan Wang. 2024. Cost-effective hallucination detection for llms. *Preprint*, arXiv:2407.21424.

Chengcheng Wei, Ze Chen, Songtan Fang, Jiarong He, and Max Gao. 2024. OPDAI at SemEval-2024 task 6: Small LLMs can accelerate hallucination detection with weakly supervised data. *Preprint*, arXiv:2402.12913.

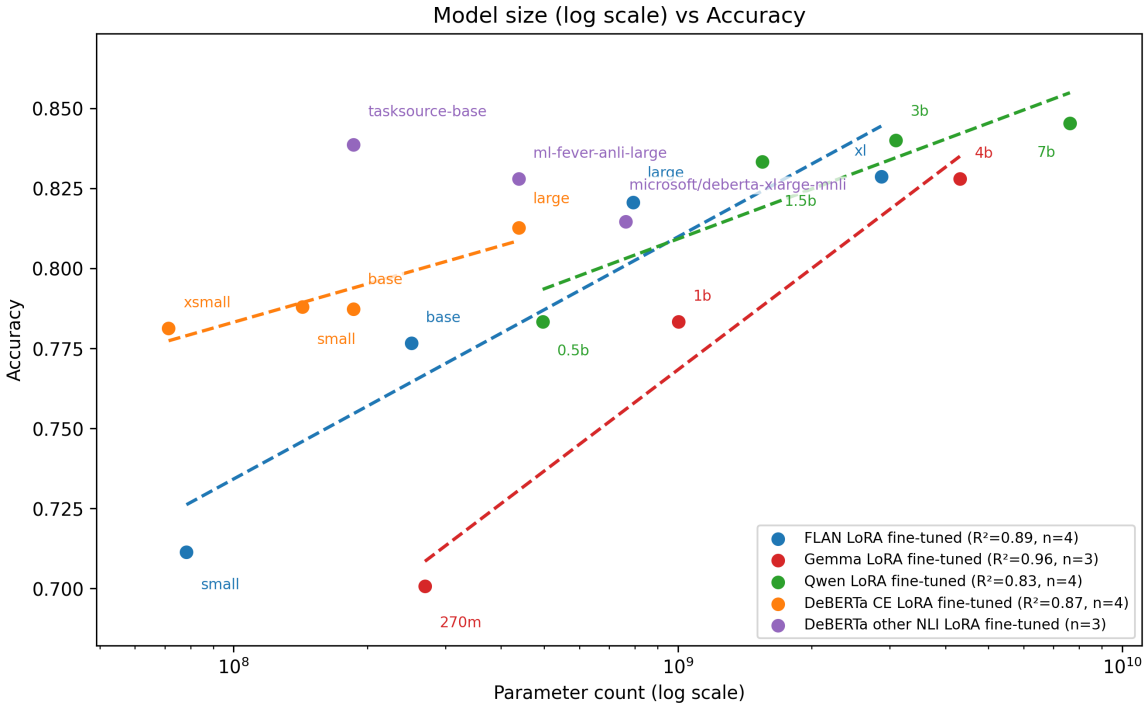
A Supplementary Results

Figures 2–6 provide the supplementary accuracy and efficiency results referenced in the main text.



Dashed lines show least-squares trends fitted within eligible family/adaptation conditions over log10(parameter count);
DeBERTa other-NLI variants are shown as grouped points without a fitted regression line because they combine heterogeneous checkpoints.
R^2 is descriptive.

Figure 2: Out-of-the-box accuracy versus total parameter count. Dashed lines show descriptive least-squares trends for the four controlled model families. Heterogeneous DeBERTa NLI checkpoints are shown as supplementary points without a fitted trend.



Dashed lines show least-squares trends fitted within eligible family/adaptation conditions over log10(parameter count);
DeBERTa other-NLI variants are shown as grouped points without a fitted regression line because they combine heterogeneous checkpoints.
R^2 is descriptive.

Figure 3: LoRA-adapted accuracy versus total parameter count. Dashed lines show descriptive least-squares trends for the four controlled model families. Heterogeneous DeBERTa NLI checkpoints are shown as supplementary points without a fitted trend.

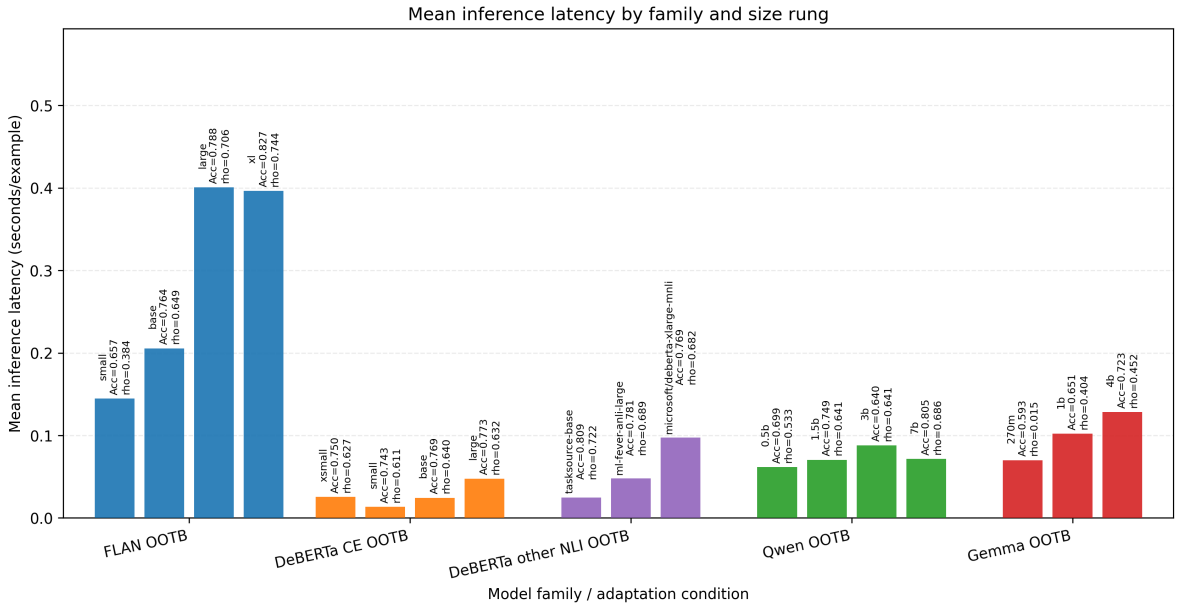


Figure 4: Mean out-of-the-box inference latency by model family and size rung. Latency is measured in seconds per test example in the shared A100 environment. Bar annotations report the corresponding test accuracy and Spearman's ρ.

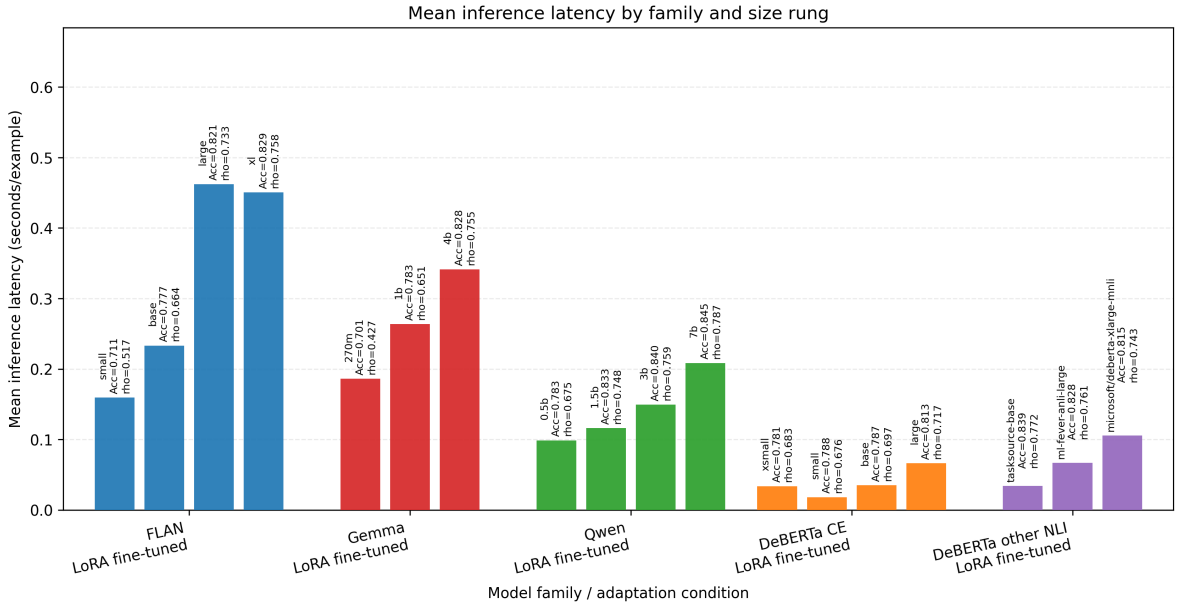


Figure 5: Mean inference latency after LoRA adaptation by model family and size rung. Latency is measured in seconds per test example in the shared A100 environment. Bar annotations report the corresponding test accuracy and Spearman's ρ.

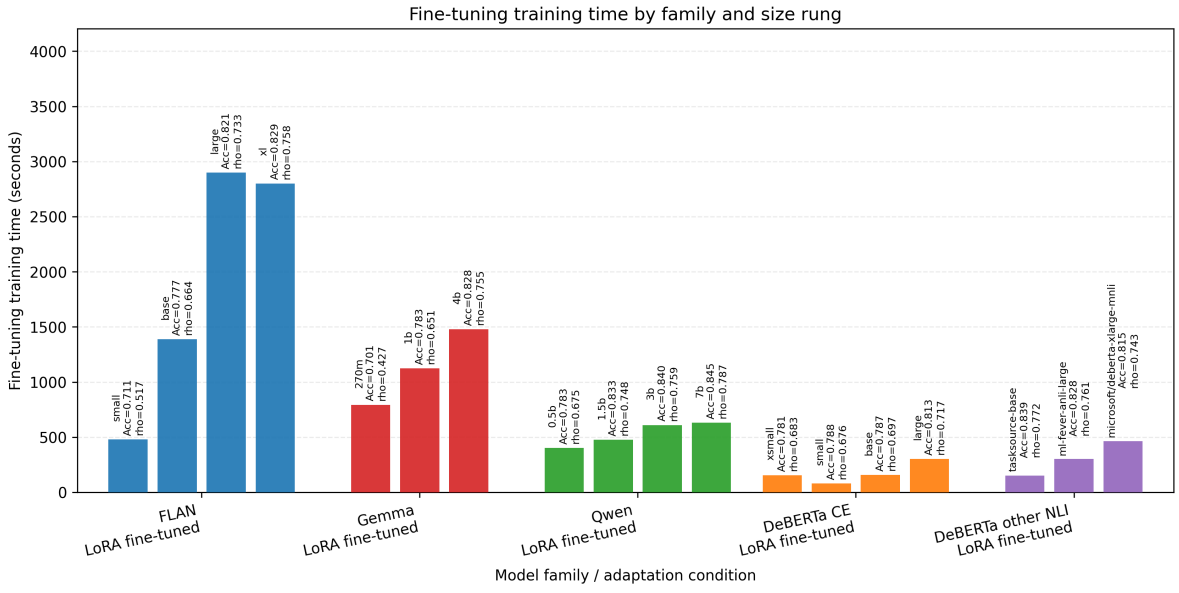


Figure 6: Total LoRA fine-tuning time by model family and size rung in the shared A100 environment. Bar annotations report the corresponding final test accuracy and Spearman's ρ.