Deference with attributions in LLM Cascades

Lize Pirenne[1] Gaspard Lambrechts[1] Norman Marlier[2]

Maxence de la Brassinne Bonardeaux[2] Gilles Louppe[1] Damien Ernst[1]

[1]University of Liège [2]NRB

Abstract

In this paper, we address the problem of iteratively choosing to defer from a small Language Model (LM) to a larger one in order to balance accuracy and cost in a closed question-and-answer setting. Our deferral strategy is defined as follows: we identify the important sentences in the context, per the models, measure the differences in the ranking of their importance with the normalized Kendall Tau Distance and threshold that value to create a binary decision. We find that we can tune the desired cost or accuracy using this threshold value. Although it is outperformed on a dataset by a simple uncertainty average threshold, our method provides a more interpretable mechanism. Indeed, the attributions are found using the attributions of the generating LLM, which can be easily checked by a human. We find that our method offers a good interpolation of base model accuracies, especially in low cost regimes.

1 Introduction

While Large Language Models (LLMs) are rapidly improving their ability to answer questions about long contexts, their cost often restricts the use-cases in which they can be deployed. Moreover, users can ask questions that do not require state-of-the-art performance, thereby diluting the benefits of servicing large models. Adding smaller models can be advantageous if they can be allocated to the right questions.

Instead of training a router model, we can let the models decide themselves by adding a deferral option. This idea dates back to Chow's optimal reject rule (Chow, 1970). Since then, a large literature on selective classification and the reject option has continued to develop (Hendrickx et al., 2024). Some of these works formalize the trade-off between coverage and risk, and propose loss surrogates and calibration methods for rejection decisions (Herbei and Wegkamp, 2006; Geifman and El-Yaniv, 2017).

We avoid "learning to defer" (Madras et al., 2018) a predictor-plus-rejector (Mozannar and Sontag, 2020) because it is expensive when models change fast.

One can chain the deferral choices to construct cascades and early-exit networks such as BranchyNet (Teerapittayanon et al., 2016) or FastBERT (Liu et al., 2020). With these architectures, we can reduce the average inference cost by terminating computation early for easy examples. A common deferral rule is confidence-based gating (e.g., max-softmax or entropy); this can fail when downstream models are specialists or under label noise and distribution shift (Jitkrittum et al., 2023). Recent work has focused on improving uncertainty estimates specifically for cascades. In particular, Gatekeeper (Rabanser et al., 2025) fine-tunes small models so that they produce high confidence when correct and low confidence when incorrect. Other lines of work study post-training calibration and selective classification strategies to make rejection decisions more reliable (Fisch et al., 2022).

These direct measurements (e.g., entropy) and estimations (i.e., dedicated head or router) of uncertainty paint a narrow picture of the many ways in which an LLM can be mistaken. We argue that, in order to reach a better accuracy/cost trade-off, we must take into account hallucinations that can be made in full confidence. To identify these problematic cases, we

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 29236–29242
October 24-29, 2026 ©2026 Association for Computational Linguistics

can rely on the sentences in the context used to answer rather than on the answer itself. These contributive attributions can be extracted directly from the attention matrix, or with costlier and more accurate methods such as TracLLM (Wang et al., 2025), ContextCite (Cohen-Wang et al., 2025) and TreeFinder (Pirenne et al., 2025). We hypothesize that the agreement between attributions of different models gives a metric correlated to the accuracy of both answers. Using this value in a cascade, we can infer when to defer; as long as the agreement is low, we replace the smaller model by a larger one and find new attributions.

In this work, we avoid any training to be able to swap models in and out without friction. This, we believe, is especially important in a field where models develop quickly. We leverage the inherent capabilities of Large Language Models (LLMs) to indicate their center of attention by providing attributions. Comparing those can estimate the uncertainty and signal for hallucinations in a deeper way than the probabilities of tokens only. Finally, rather than rooting the decision in a single model, we compare two adjacent models in the cascade. This is more expensive than most methods when the cascade is small, but tends to the same complexity when the number of parameters increases quickly between adjacent models.

2 Methodology

We consider the problem of having a dataset of contexts $\mathcal{C}$, questions $\mathcal{Q}$ and answers $\mathcal{A}$, and an ordered list of models of increasing size `M`. We want, for a given average cost, to have the best average accuracy. Or, conversely, have the best average cost for a given average accuracy.

A particular context $C \in \mathcal{C}$ is composed of an ordered list of sentences. Each can be assigned a value corresponding to the impact on the probability of the answer, leading to a ranking. The Kendall Tau distance τ measures the difference in rankings by counting the number of swaps needed for the first ranking to become the second. It is normalized between -1 and 1, which indicates perfect matching and complete mismatch, respectively.

Our algorithm consists in comparing the rankings for all sentences in the context of two adjacent models in the cascade, using τ as a threshold. If the threshold is met or surpassed, we give the answer of the larger model of the two. If it is not met, we replace the ranking of the smaller model with the one of the next model in the list and repeat the process. We directly return the answer of the largest model if reached. A visual representation of our algorithm can be seen in Figure 2

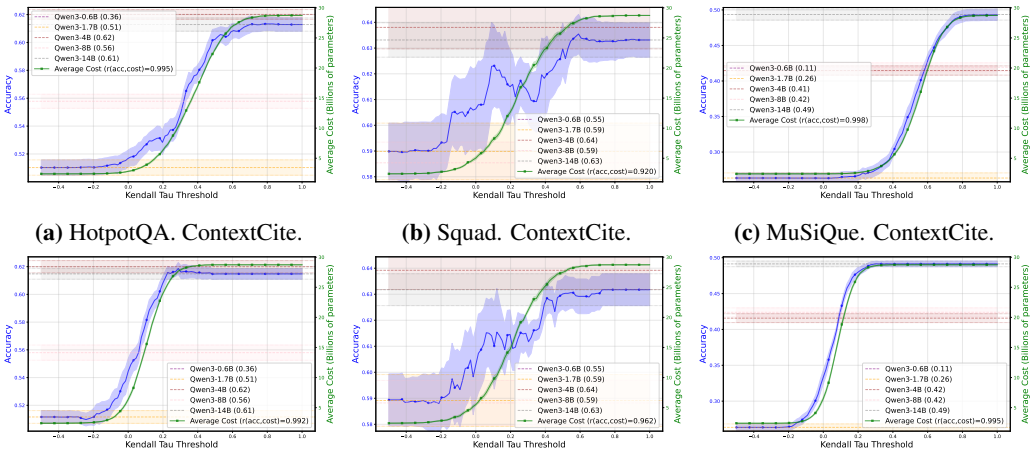
2.1 Attributions

A key element of our algorithm is to have efficient and accurate rankings. We identified three recent candidate methods to produce them: TreeFinder (Pirenne et al., 2025), TracLLM (Wang et al., 2025) and ContextCite (Cohen-Wang et al., 2025). All methods are based on perturbation to remove the least contributing sentences.

ContextCite learns a linear model to assign scores using a set of ablated contexts for a given question and the probability of the answer. The scores are directly used to select attributions. We set in the experiments the number of ablations to 32. TreeFinder divides the context into parts of similar lengths. Each part is successively dropped out of the context and the impact on the probability of the answer in terms of necessity and sufficiency is recorded. The most impactful parts are divided themselves. This is repeated until the most impactful parts cannot be subdivided again. We set the initial number of chunks to 5 and keep advancing the top 3. TracLLM operates similarly to TreeFinder but removes unimportant sentences completely from the context, and computes more marginal probabilities than simply dropping one part at a time. Only TreeFinder and ContextCite can provide a ranking for all sentences, so we will be using these two methods in our experiments.

2.2 Models

To have easily comparable models, we choose to consider entire release suites as cascades. In particular, we use the Qwen3 model suite (Team, 2025) with sizes of $0.6B$, $1.7B$, $4B$, $8B$ and $14B$ parameters. We load them in 4-bit precision using quantization (Dettmers et al., 2023). The lengths of the answers are set to a



(a) HotpotQA. ContextCite. **(b)** Squad. ContextCite. **(c)** MuSiQue. ContextCite.

(d) HotpotQA. TreeFinder. **(e)** Squad. TreeFinder. **(f)** MuSiQue. TreeFinder.

Figure 1: Average accuracy (in blue) and cost of the cascade (in green) at varying Kendall Tau (τ) for the Qwen3 model suite. Standard deviations over 12 runs are reported in lighter colors. The accuracy of models on their own appear as horizontal lines. High correlation between accuracy and cost is visible and confirmed by the Pearson correlation between the accuracy and the cost r in the legend.

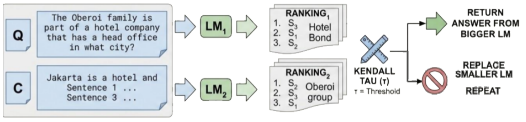


Figure 2: Visual representation of our cascade algorithm

maximum of 200 with a repetition penalty of 1, a low temperature of 0.5, top-p to 0.8 and top-k to 20.

2.3 Datasets

Our method is intended to work in a closed-QA setting. This means that if $Q \in \mathcal{Q}$ is answerable by $A \in \mathcal{A}$, then $C \implies A$. Otherwise, the answer should reflect its absence. We therefore find datasets where C is long enough to have meaningful differences in ranking ($|C| > 10$) while fitting inside our vRAM ($C < |80|$). This parameter of the minimal number of sentences has not been explored further, but we think increasing it would lead to more consistent results. We load up to 1000 samples of Squad (Rajpurkar et al., 2016), HotpotQA (Yang et al., 2018), and MuSiQue (Trivedi et al., 2022) corresponding to our previous criteria.

2.4 Baselines

As a first baseline, we set the semantic similarity as a distance, as suggested Soiffer et al. (2025). For this, we compute the BLEURT score (Sellam et al., 2020) between two answers from adjacent models. BLEURT and our method start can only start from the second model because the first is used as a reference. We note that comparing a first answer from the model to a second one from the same model leads to degraded performances and is left in Section A to maintain better plot clarity.

We also use their baselines, which set a threshold for the aggregation of the output token probabilities of the current model. The aggregation methods are the minimum (min), average (avg), and q^{th} quantile (quantile). We set $q = 0.95$ in our experiments, although the ideal value in Soiffer et al. (2025) varied from 0 to 1. These three baselines do not compare models by pair and thus can choose to not defer on the first model.

3 Results

In our experiments, we will first show that the computed distance is strongly correlated to the average difference in accuracy between the two models by varying the threshold and reporting the average accuracy. Afterwards, we will report the curve of the average accuracy against the average cost for a range of thresholds for our method and various other baselines and verify our objective.

One of our hypotheses is that similarity between rankings can only be achieved on aver-

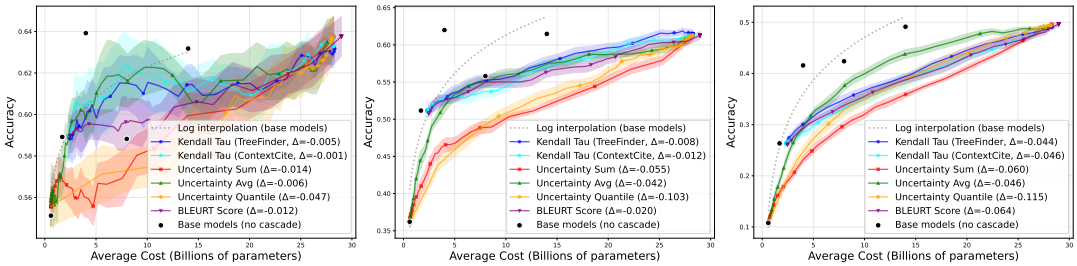


Figure 3: Accuracy against cumulative cost of the cascade for all tested models and datasets for SQuAD (left), HotpotQA (middle) and MuSiQue (right). Standard deviations of 12 runs are reported in lighter colors. The mean difference Δbetween the interpolated model performances and accuracies of each method is also reported for reference (higher is better).

age if both adjacent models converge, and similar attributions lead to similar answers. Setting a higher τ increases the minimum similarity that two rankings must have, thus pushing for a higher accuracy at a higher cost.

To verify this, we plot the average accuracy and cost over the datasets in Figure 1, along with the standard deviation over 12 runs, against τ. The cost corresponds to the cumulative number of parameters in billions. We see that increasing τ directly increases both the accuracy and cost of the cascade of Qwen3 models. The reported Pearson correlation between accuracy and cost in the legend shows that there is indeed a strong correlation.

In Figures 1a, 1b, 1d and 1e, specifically, we see `Qwen3-8B` underperforming adjacent models in the cascade. This may be due to the relatively small subset ($|\mathcal{Q}| = 1000$) or to a real discrepancy of this specific model on these closed question-answer tasks. Surprisingly, the corresponding dip only appears in the Squad dataset. This can be due to the cascade being less effective on HotpotQA, which leads to the accuracy being too far behind the individual models to be affected.

For all methods, we computed the accuracy against the cumulative cost of the cascade and reported them in Figure 3. We also report the difference between the accuracy of each method and the models average accuracies (interpolated). We can see that the best interpolation of the performance of the suite is given by our method, except for the MuSiQue dataset where the uncertainty average is better in the higher cost regimes. `BLEURT` performs poorly, contrary to the literature (Soiffer et al., 2025).

We note that we are still far from the optimal cascade. An oracle only returning at the first correct model would achieve an accuracy of 0.86 ± 0.01 on Squad for a cost of 6.01 ± 0.28 billion parameters. The same computation gives 0.85 ± 0.01 accuracy for a cost of 7.54 ± 0.12 billion parameters on HotpotQA and 0.70 ± 0.01 accuracy on MuSiQue for a cost of 14.14 ± 0.15 billion parameters. To achieve these ideal performances, we believe the first step is to verify the answer given at each step, and this can be done efficiently using the reduced context found by the attributions.

Another finding is that the constructed cascades scarcely outperform a single model chosen at the right size. For example, `Qwen3-4B`'s cost/performance tradeoff is unreachable in all cases. In general, there does not seem to be a point above `Qwen3-2B` where the cascade acts as a Pareto frontier of the tradeoff.

4 Conclusion

In this paper, we have used attribution finding methods to choose to defer in a cascade of LLMs setting. Our first result was that the matching of rankings of these attributions allows us to choose with a single parameter the average accuracy (or cost) of the cascade.

We have followed by comparing the method on three datasets with traditional uncertainty estimation for deferral. We have found that we could interpolate the accuracy/cost curve of the suite of models often as accurately as traditional deferral methods, if not better in two cases. We hypothesize that the interpolation could be enhanced by the choice of better models.

Future works can leverage the added explain-

ability to offer additional details for the deferral to the user, or integrate them in a more complex system that would give better accuracy/cost tradeoffs. These works could approach the ideal point in the curve given by a perfect oracle by reusing the computed attributions and efficiently verifying the answer. Crucially, future work should aim to use only a fraction of the bigger models for the threshold evaluation in order to lift the tradeoff curve and compete with base models. All code is available here.

Limitations

The experiments were carried out on relatively few samples and independent runs, which resulted in quite large standard deviations. The results might differ in other datasets or simply in more data points.

Although graphs of accuracy against the cumulative cost can give some insight on the performance of each method, other metrics could improve the clarity of the results. These include accuracy against FLOP and against time. This is mitigated by the fact that we do not encompass all attribution-finding methods, and some could reduce or erase the overhead. The cost in parameter count is, we believe, more agnostic to the specific implementation details.

We have not studied the impact of the difference of size and competency between adjacent models. This may prove to be a crucial point for the broader applicability of our method.

Beyond the risks associated with using Large Language Models for answering questions, our method could provide a misplaced confidence in the answers provided. It is not because we try mitigating hallucinations and enhance factuality that these phenomenon disappear.

Acknowledgments

Lize Pirenne and Damien Ernst gratefully acknowledge the financial support of the Walloon Region for Grant No. 2010235 – ARIAC by DW4AI and the NRB Research Chair on Large Language Models for the Computer Software Industry.

Gaspard Lambrechts gratefully acknowledges the financial support of the Wallonia-Brussels Federation and the Fund for Scientific Research for his FRIA and CR grants.

The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

All code and most of the paper have been written by Lize Pirenne. Generative artificial intelligence has been used to help in their production and quality control. Other authors have provided feedback and scientific guidance.

References

C. Chow. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2025. ContextCite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in neural information processing systems*, 36:10088–10115.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. 2022. Calibrated selective classification. *arXiv preprint arXiv:2208.12084*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: A survey. *Machine Learning*, 113(5):3073–3110.

Radu Herbei and Marten H. Wegkamp. 2006. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):709–721.

Wittawat Jitkrittum, Neha Gupta, Aditya K Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. 2023. When does confidence-based cascade deferral suffice? *Advances in Neural Information Processing Systems*, 36:9891–9906.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. 2020. Fastbert: a self-distilling BERT with adaptive inference time. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 6035–6044. Association for Computational Linguistics.

David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR.

Lize Pirenne, Gaspard Lambrechts, Norman Marlier, Maxence de la Brassinne Bonardeaux, Gilles Louppe, and Damien Ernst. 2025. Contributive attribution for question answering via tree-based context pruning. *Orbi preprint: 2268/337121*.

Stephan Rabanser, Nathalie Rauschmayr, Achin Kulshrestha, Petra Poklukar, Wittawat Jitkrittum, Sean Augenstein, Congchao Wang, and Federico Tombari. 2025. Gatekeeper: Improving model cascades through confidence tuning. In *Tiny Titans: The next wave of On-Device Learning for Foundational Models (TTODLer-FM)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Duncan Soiffer, Steven Kolawole, and Virginia Smith. 2025. Training-free semantic deferrals for open-ended llm cascades. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.

Yanting Wang, Wei Zou, Runpeng Geng, and Jinyuan Jia. 2025. {TracLLM}: A generic framework for attributing long context {LLMs}. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3845–3864.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Multiple answer generation

The method presented in the paper compares the attributions of the answers given by models growing in size. In this method, we interject a copy of each model in the cascade such that the attributions are given by the same model two times, potentially for two different answers. The deferral criterion is still the Kendall Tau distance between the rankings.

In Figure 4, we add to the comparison plot the two additional methods that compare attributions across independent runs, since this is equivalent and does not require recomputing the whole cascade. We nickname these "cross-run" methods because they require to compare across two runs of our cascade as implemented. Neither the cross-run nor the standard methods take into account the cost of computing the attributions. Interestingly, there is a significant gap in accuracy that barely decreases as the number of parameters grows. There is a true added value to comparing between different models of the cascade.

B Kendall Tau Variability

We plot the mean value of τ and the standard deviations for each model pair in the cascade and for each dataset in Figure 5. We can see that the values taken by τ are varied enough to allow for a diverse range of thresholds.

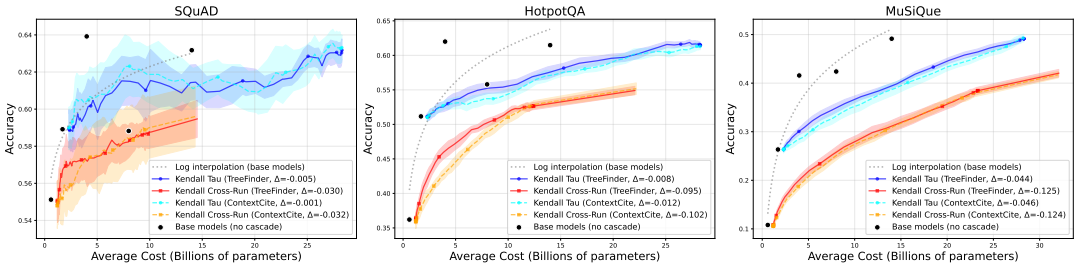


Figure 4: Accuracy against cumulative cost of the cascade for all tested models and datasets for SQuAD (left), HotpotQA (middle) and MuSiQue (right). Standard deviations of 12 runs are reported in lighter colors for the standard method and 3 for the cross method. We can see that the cross generation method offers a much lower accuracy to cost ratio that the standard methods for all three datasets. The mean difference (Δ) between the interpolated model performances and accuracies of each method is also reported for reference (higher is better).

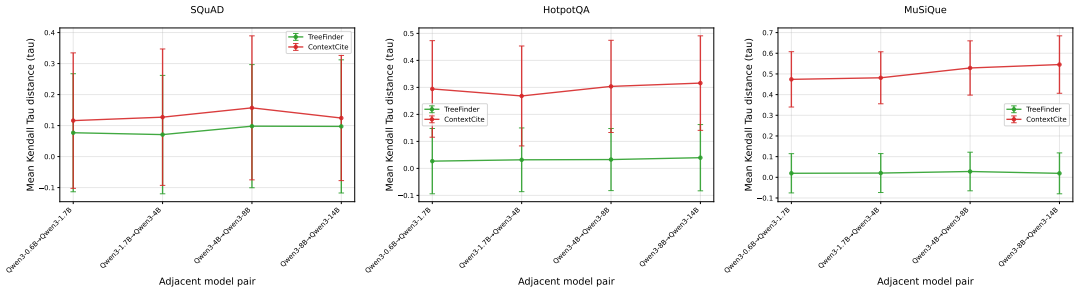


Figure 5: Mean value of τ and standard deviations for each model pair and each dataset.