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]Keyes

EMNLP2026 BUDAPEST

Large LLMs are accurate but costly; small ones are cheap but often wrong. **When should we defer to a larger model?**

Instead of only looking at confidence, **compare which sentences of the context each model relies on** (attributions).

A **training-free** deferral rule with a **single threshold** τ tunes the accuracy/cost trade-off, and is interpretable.

Context: deferral in LLM cascades

- LLMs answer questions about long contexts, but **costs** are prohibitive.
- Smaller models should handle easier questions.
- **Deferral option** rather than a trained router: dates back to Chow's reject rule (Chow 1970); large literature on selective classification (Hendrickx et al. 2024).
- Chained deferrals give **cascades** and early-exit networks (Teerapittayanon, McDanel, and Kung 2016; Liu et al. 2020).
- Common rule: confidence gating (max-softmax, entropy). It can fail under specialist models, label noise or distribution shift (Jitkrittum et al. 2023).

Motivation

- Token-level uncertainty paints a **narrow picture** of how an LLM can be wrong: hallucinations can be made **in full confidence**.
- **Hypothesis**: agreement between the attributions of two models is correlated with the accuracy of both answers.
- Look at the **context sentences used to answer** rather than at the answer itself.
- **No training**: models can be swapped in and out without friction, important in a fast-moving field.
- We compare two **adjacent** models in the cascade; the extra cost vanishes when model size grows quickly between neighbours.

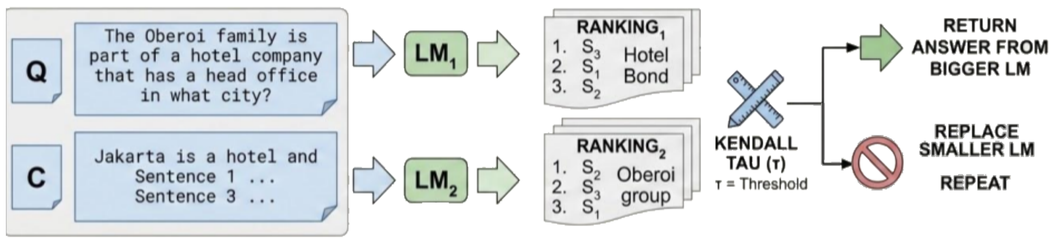
Attributions are computed from the generating LLM itself and **can be checked by a human**: the deferral is interpretable.

Our method: deferral with attribution agreement

- **Setting**: closed-QA with contexts $\mathcal{C}$, questions $\mathcal{Q}$, answers $\mathcal{A}$, and an ordered list of models of increasing size.
- **Goal**: best accuracy for a given average cost (or best cost for a given accuracy).
- A context C is an ordered list of sentences; each sentence gets a score for its impact on the probability of the answer, giving a **ranking**.
- The **normalized Kendall Tau distance** $\tau \in [-1, 1]$ counts the swaps needed to turn one ranking into the other.

Cascade algorithm

1. Compute the sentence rankings of two adjacent models.
2. Compare them with τ and a threshold.
3. If the threshold is crossed, return the answer of the larger model.
4. Otherwise, replace the smaller model with a larger one and repeat.



Attributions and experimental setup

Attribution methods

- **ContextCite** (Cohen-Wang et al. 2025): linear model fitted on ablated contexts (32 ablations).
- **TreeFinder** (Pirenne et al. 2025): recursive chunk dropping with necessity and sufficiency scores (5 initial chunks, top 3 kept).

Models

- Qwen3 suite: 0.6B, 1.7B, 4B, 8B, 14B, in 4-bits.
- Max. 200 tokens, temperature 0.5, top-p 0.8, top-k 20.
- Cost = cumulative number of parameters (billions).

■ 12 independent runs; mean ± standard deviation reported.

Datasets (up to 1000 samples, $10 < |C| < 80$ sentences)

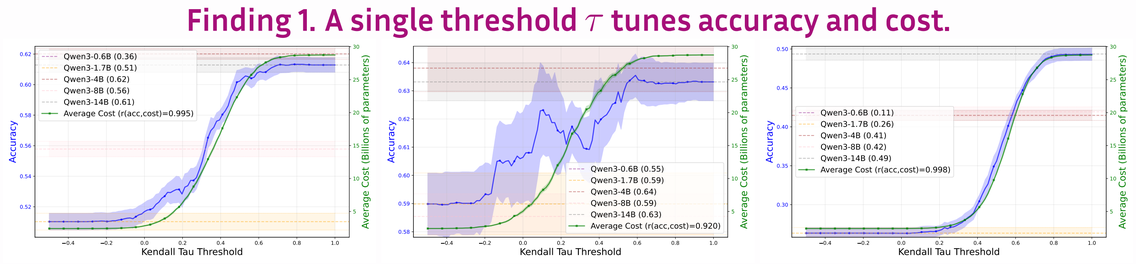
- SQuAD (Rajpurkar et al. 2016)
- HotpotQA (Yang et al. 2018)
- MuSiQue (Trivedi et al. 2022)

Baselines

- **BLEURT** (Sellam, Das, and Parikh 2020) semantic similarity between answers of adjacent models.
- Thresholds on aggregated output-token probabilities of the current model: **min**, **avg**, and q-**quantile** ($q = 0.95$).

Code: `https://github.com/Pangasius/AttributionCascade`

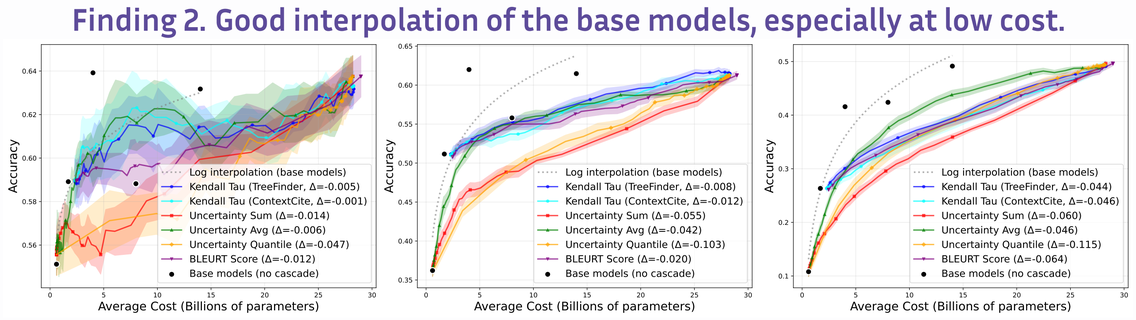
Finding 1. A single threshold τ tunes accuracy and cost.



Accuracy (blue) and cost (green) against τ using Qwen3 and ContextCite; left to right: HotpotQA, SQuAD, MuSiQue.

- Raising τ raises **both** accuracy and cost: strong Pearson correlation.
- It seems similar rankings can only be reached on average if both adjacent models converge, and similar attributions lead to similar answers.
- `Qwen3-8B` underperforms its neighbours on SQuAD and HotpotQA (small subset or real discrepancy on closed-QA). Can't assume bigger = better.

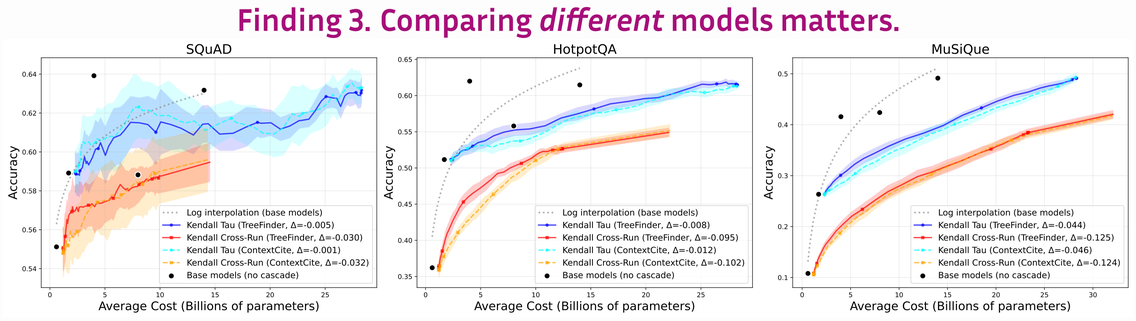
Finding 2. Good interpolation of the base models, especially at low cost.



Accuracy against cumulative cost, SQuAD (left), HotpotQA (middle), MuSiQue (right).

- Our method gives the **best interpolation** of the suite's performance, except on MuSiQue at higher cost, where the average-uncertainty baseline is better.
- `BLEURT` performs poorly, contrary to the literature (Soiffer, Kolawole, and Smith 2025).

Finding 3. Comparing *different* models matters.



Accuracy against cumulative cost cross-run and standard, SQuAD (left), HotpotQA (middle), MuSiQue (right).

- "Cross-run": compare attributions of the **same** model for two completions.
- Much lower accuracy-to-cost ratio on all datasets.
- There is **true added value** in comparing attributions between different models of the cascade.

Finding 4. We are still far from the optimal cascade.

Oracle (first correct returns):

Dataset	Accuracy	Cost (B params)
SQuAD	0.86 ± 0.01	6.01 ± 0.28
HotpotQA	0.85 ± 0.01	7.54 ± 0.12
MuSiQue	0.70 ± 0.01	14.14 ± 0.15

- An ideal cascade outperforms every model in accuracy, and is cheaper or as cheap as the largest model.
- We think there is an improvement yet possible if we check that the attribution entails the answer.

Limitations. Few samples and runs (large standard deviations); cost measured in parameters, not FLOPs or time, and the attribution overhead is not counted; impact of the size gap between adjacent models not studied; attributions may give **misplaced confidence**: mitigating hallucinations does not make them disappear.

Conclusion and future work

- Matching the rankings of attributions lets a **single parameter** choose the average accuracy (or cost) of an LLM cascade, with no training.
- Compared with uncertainty-based deferral on three datasets, it interpolates the accuracy/cost curve as well as or better than traditional methods in most cases, with more **explainable** decisions.
- **Future work:**
  - ↳ use the added explainability to give the user details on each deferral;
  - ↳ reuse the computed attributions to verify answers and approach the oracle;
  - ↳ use only a fraction of the bigger models for light-weight attributions.

Acknowledgments

L. Pirenne and D. Ernst acknowledge the financial support of the Walloon Region (Grant No. 2010235 – ARIAC by DW4AI) and the NRB Research Chair on Large Language Models for the Computer Software Industry. G. Lambrechts acknowledges the Wallonia-Brussels Federation and the Fund for Scientific Research (FRIA and CR grants). Computational resources: Lucia, the Tier-1 supercomputer of the Walloon Region (grant agreement n°1910247). Generative AI was used to help in the production and quality control of code and paper.