Training-Free Semantic Deferrals for Open-Ended LLM Cascades

Duncan Soiffer [1] **Steven Kolawole** [1] **Virginia Smith** [1]

Abstract

Existing cascade systems struggle with open-ended text generation due to evaluation challenges where multiple valid outputs exist without ground truth references. We propose using semantic agreement between multiple model outputs as a training-free deferral signal and evaluate semantic similarity metrics against token-level confidence across translation, summarization, question answering, and reading comprehension tasks. We show that semantic signals provide a stronger indication of when deferral is appropriate than token-level methods and are resilient to heterogeneous model quality.

1. Introduction

Large language models deliver higher quality outputs as size increases but at substantial computational cost. Cascade systems address this tension by routing queries between smaller and larger models, employing smaller models when possible and deferring to larger models only when necessary.

Despite their economic appeal, cascade systems face unique challenges in open-ended text generation. Unlike classification with discrete labels or numerical problems with definitive answers, generation quality exists on a continuous spectrum with multiple valid outputs per input. This subjectivity complicates deferral decisions, as determining when to defer requires assessing quality without ground truth references—the "correct" vs "incorrect" paradigm no longer applies. Moreover, existing LLM cascading approaches rely on learned routing mechanisms requiring substantial training investment through domain-specific fine-tuning or model-specific engineering. When models are updated, routing systems require complete retraining, imposing recurring costs and limiting production agility.

We propose using semantic agreement between multiple model outputs as a training-free deferral signal. When several smaller models produce semantically consistent answers, their collective response likely represents reliable output. Conversely, semantic disagreement suggests uncertainty, indicating deferral would be beneficial. By focusing on output agreement rather than learned confidence, our approach eliminates supervision while remaining applicable across diverse tasks.

We systematically explore semantic agreement capturing different agreement dimensions: surface-level n-gram overlap (BLEU, ROUGE), pretrained metrics (BLEURT), and reference-free embedding agreement (SBERT). Evaluating these alongside token-level confidence across translation, summarization, question answering, and reading comprehension, we demonstrate that semantic signals provide stronger deferral decisions than token-level metrics and offer favorable cost-quality tradeoffs despite compute overhead.

This approach offers advantages beyond eliminated training: semantic agreement functions across black-box models, enabling application to proprietary systems, and demonstrates unique resilience to heterogeneous model quality—adding lower-performing models often preserves or improves performance, unlike token-level approaches where ensemble performance degrades toward the worst model.

2. Related Work

Cascading in Language Models With the sporadic rise of LLM applications, previous work (Chen et al., 2024; Ong et al., 2024; Aggarwal et al., 2024) has demonstrated the cascade systems' effectiveness for balancing computational efficiency and output quality. However, these approaches require extensive training of routing mechanisms, struggle with OOD generalization, and focus primarily on classification or objectively assessable tasks. When underlying models are updated—common in today's rapidly evolving LLM landscape—routing systems require complete retraining, imposing recurring costs, though recent work (Kolawole et al., 2024; Feng et al., 2024; Jitkrittum et al., 2025) explores ways to circumvent this bottleneck.

Confidence-Based Deferral for Generation Tasks Recent cascading methods have explored confidence signals for open-ended generation. Gupta et al. (2024) explored token-level uncertainty for selective generation while

[1] Carnegie Mellon University, USA. Correspondence to: Duncan Soiffer <dsoiffer@cs.cmu.edu>, Steven Kolawole <skolawol@andrew.cmu.edu>.

Proceedings of the 3^{rd} Efficient Systems for Foundation Models Workshop at the International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Narasimhan et al. (2024) extended this with speculative decoding. However, it remains unclear whether token-level confidence signals, optimized for next-token prediction, effectively capture semantic coherence and factual accuracy determining generation quality. Critically for practical deployment, confidence-based approaches require model internals access, limiting their applicability to proprietary systems dominating real-world usage.

Ensemble Methods and Agreement-based Signals Traditional ensembles focus on improving quality through combination rather than defferal decisions (Lakshminarayanan et al., 2017; Wang et al., 2020; 2023; Jiang et al., 2023). ABC (Kolawole et al., 2024) uses ensemble agreement for cascade deferral but limits this to fixed output spaces. Yue et al. (2024) also explored consistency-based quality estimation for numerical reasoning tasks with objectively correct answers. Unlike these approaches, we address the more challenging scenario of open-ended generation where "correctness" must be defined through semantic similarity rather than exact matching.

Semantic Similarity in NLP Semantic similarity metrics primarily serve as evaluation tools rather than active system components. Traditional metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure surface-level lexical overlap but miss deeper relationships. Recent approaches using contextual embeddings—BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2019)—show stronger correlation with human judgements. However, semantic similarity for use in cascade deferral decisions remains unexplored, with prior work treating it as post-hoc evaluation rather than architectural components for resource allocation decisions.

Our contribution We bridge these research areas by introducing semantic agreement as a training-free deferral signal for open-ended text generation. Our method: (1) *eliminates training requirements* through direct similarity computation, (2) *functions with black-box models* by operating on outputs rather than internals, and (3) *handles open-ended generation* where multiple valid outputs exist. We extend ABC's agreement-based approach beyond discrete classification to continuous semantic spaces, establishing the first comprehensive comparison of semantic vs token-level confidence for cascade deferral.

3. Methods

3.1. Deferral Protocol

We formalize the deferral process as follows. A cascade comprises n lightweight ensemble models, $M_1,...,M_n$, and a larger, high-capacity target model M_T. Given an input x, each ensemble model produces a prediction $y_i = M_i(x)$. A deferral score function $s(y_1, \dots, y_n)$ then determines whether to defer to M_T. If the system chooses not to defer, it selects a final output $y = y_i$ corresponding to the prediction with highest score according to an output scoring function $o(y_i, y_1, ..., y_n)$.

We adopt the convention that higher scores reflect greater certainty. Accordingly, the system defers to the target model for inputs with lower deferral scores s, while retaining predictions with higher s.

3.2. Semantic Agreement Signals

Semantic similarity metrics offer a natural way to assess agreement between multiple model outputs without ground-truth references. We explore increasingly sophisticated metrics capturing different aspects of similarity, from surface-level lexical overlap to deeper semantic representation. This progression allows us to examine how different dimensions of semantic agreement affect cascade performance.

In an ensemble setting, we use these similarity metrics as the output scoring function o by computing pairwise similarities between y_i and all other y_j, and set $s = \max_i o(y_i, y_1, ..., y_n)$.

Classic Overlap Metrics We begin with classic reference-based metrics: BLEU measures n-gram precision, ROUGE-N measures n-gram overlap, and ROUGE-L measures longest common subsequence overlap.

Pretrained Metrics BLEURT is a regression model based off BERT, fine-tuned to reflect human judgments of generation quality. It requires text pairs as input and outputs a similarity score. Separately, we also use SBERT (Reimers & Gurevych, 2019) as a reference-free approach by embedding each output y_i into a vector representation z_i and computing pairwise cosine similarities in the embedding space.

3.3. Token-Level Confidence Signals

For comparison, we evaluate token-level confidence metrics that extend classification confidence to generation (Gupta et al., 2024):

- **Chow-Sum**: Sum of token log probabilities reflecting overall sequence confidence.
- **Chow-Avg**: Sum of token log probabilities normalized by sequence length.
- **Chow-Quantile**(q): q-th quantile of token log probabilities.

3.4. Handling Model Heterogeneity

Ensembling heterogeneous models based on token-level confidence is challenging due to differences in training dynamics, architectures, and vocabularies. Baseline token probabilities vary significantly across models and may reflect different calibration levels, leading one model to systematically dominate when using raw confidence scores.

We address this without an expensive post-training fix, through z-score normalization: we run each model on

Task	Semantic Models	Token-level Model	Large Model	Best AUC (Semantic)	Best AUC (Token-level)
WMT19 DE→FR	Qwen2.5-1.5B, Gemma3-1B, mT0 Large	Qwen2.5-1.5B	Llama-3.1-8B	.6548	.6397
CNN/DailyMail	Qwen2.5-1.5B, Qwen2.5-0.5B, Gemma3-1B	Qwen2.5-1.5B	Llama-3.1-8B	.1262	.1265
XLSum	Qwen2.5-1.5B, Gemma3-1B, Llama-3.2-1B	Llama-3.2-1B	Llama-3.1-8B	.0786	.0795
SQuAD1.1	Qwen2.5-1.5B, Gemma3-1B, FLAN-T5 Large	Qwen2.5-3B	Qwen2.5-7B	.7685	.7592
TriviaQA	Qwen2.5-1.5B, Gemma3-1B, Llama-3.2-1B	Llama-3.2-3B	Llama-3.1-8B	.5044	.6016

Table 1. Illustrative results of AUCs achieved for a given cascade when using the best of the similarity metrics for the semantic cascade, or the best between Chow-Sum, Chow-Avg, Chow-Quantile for the token-level cascade. Full results are presented in the appendix.

a subset of training data, compute confidence metric statistics, then normalize during inference. We use the mean normalized confidence as the deferral score. If the system does not defer, it selects the output from the model with the highest normalized confidence.

4. Experiments & Results

We evaluate cascades using models from the FLAN-T5 (Chung et al., 2022), mT0 (Muennighoff et al., 2022), Gemma (Gemma Team, 2025), Llama (AI@Meta, 2024), and Qwen (Qwen Team, 2024) families, selected for their multilingual capabilities, cross-task generalization, and the existence of models at tiers of roughly 0.5B, 1B, 3B, or 8B parameters. Our experimental design maintains fair comparison by limiting semantic ensembles to roughly equivalent in total parameters as single 3B models, though this constraint reveals an interesting asymmetry as discussed in further paragraphs.

Evaluation Protocol We assess cascade performance using *deferral curves*. For threshold t, we compute deferral rate $\mathbb{P}(r(y_1,\ldots,y_n)=1)$ and cascade performance $\mathbb{E}[Q(C(x))]$, where Q measures quality (e.g., accuracy for TriviaQA). We plot these as t is varied, and report area under the deferral curve (AUC) as a scalar summary. For a given dataset, higher values indicate better overall performance. This evaluation method is adopted from Gupta et al. (2024), and all baseline performances and results are provided in the appendix.

4.1. Token-level Ensemble Limitations

Token-level ensembles work moderately well when ensemble models have very similar performance—for instance, an ensemble of Gemma3-1B and Qwen2.5-1.5B deferring to Llama-3.1-8B on WMT19 DE→FR achieves higher AUC than a token-level cascade of either individual model (Appendix F.2.1). However, when ensemble models have heterogeneous performance, this method significantly degrades performance compared to the single-best model. This occurs because z-scoring causes weak models to be selected about as often as strong ones, driving output quality lower. This limits the applicability of this method for open-ended text generation as it requires ensemble models to be of similar quality across all tasks—a constraint rarely satisfied in practice.

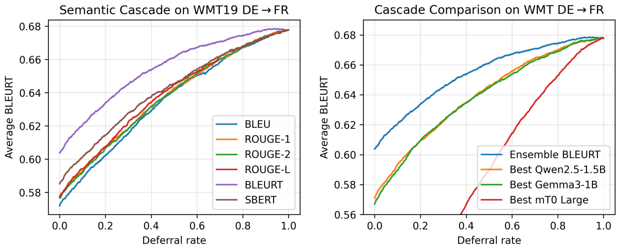


Figure 1. **(a)** Deferral rate versus average BLEURT score on WMT19 DE→FR for a semantic cascade with ensemble models Qwen2.5-1.5B, Gemma3-1B, mT0 Large, deferring to large model Llama-3.2-8B. Semantic similarity based on BLEURT consistently outperforms simpler measures, with SBERT similarly outperforming both ROUGE and BLEU. **(b)** Deferral curve for BLEURT on the same semantic cascade, compared to deferral curves with largest AUC (all Chow-Avg) for token-level cascades on the individual ensemble models. The semantic cascade consistently outperforms the token-level cascade constructed from its single-best ensemble model.

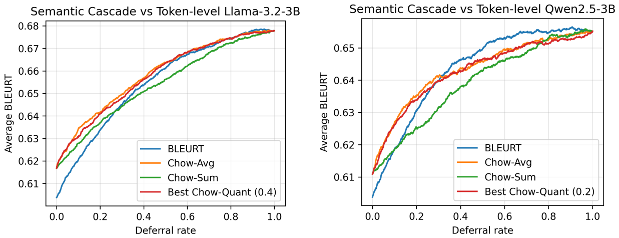


Figure 2. **(a)** Deferral curves on WMT19 DE→FR for semantic ensemble with Qwen2.5-1.5B, Gemma3-1B, FLAN-T5 Large, versus deferral curves for token-level cascade with Llama-3.2-3B. Both cascades defer to Llama-3.1-8B. **(b)** Deferral curves for the same ensemble cascade, versus token-level cascade with Qwen2.5-3B. Both cascades defer to Qwen2.5-7B. In both cases, semantic ensembles have weaker baseline performance but match or surpass token-level ensembles at larger deferral rates, indicating that semantic similarity is a better deferral signal.

4.2. Semantic Signals Provide Superior Decisions

We evaluate BLEU, ROUGE, BLEURT, and SBERT-based agreement metrics as standalone deferral signals. Our results demonstrate that semantic deferral signals capture generation reliability more effectively, with this advantage holding across diverse domains.

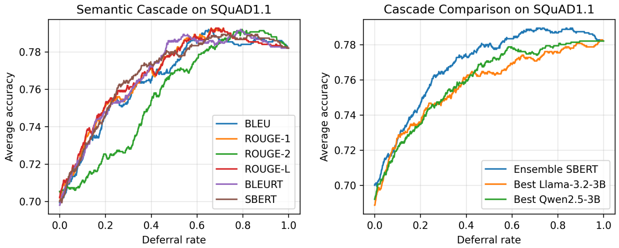


Figure 3. **(a)** Deferral curves for different semantic similarity measures on an ensemble of Qwen2.5-1.5B, FLAN-T5 Large, Gemma3-1B, deferring to large model Qwen2.5-7B. ROUGE-2 suffers due to overlooking single-word answers, but all other similarity metrics perform similarly. **(b)** Deferral curves for SBERT on the same cascade, compared to deferral curves with largest AUC for a token-level cascade with Llama-3.2-3B (Chow-Quantile-0.1) or Qwen2.5-3B (Chow-Avg). The semantic cascade consistently outperforms both token-level cascades despite using smaller models, and improves over the largest model at or above 60% deferral.

Semantic agreement signal deferral superiority. On translation, semantic cascades using BLEURT consistently match or surpass token-level cascades of their single-best models (Figure 1). Critically, semantic ensembles can also achieve this performance with smaller ensemble models—matching single-model token-level cascades with roughly equivalent total parameters (Figure 2) and outperforming them on FR→EN translation. This suggests that semantic agreement captures deferral-relevant uncertainty more effectively than token-level confidence. Reading comprehension is consistent with this trend, where semantic cascades even exceed the accuracy of the best target model (Figure 3). The consistency of this advantage across both task domains indicates that semantic agreement more reliably identifies when ensemble outputs are trustworthy versus when deferral to larger models is beneficial.

Evaluation challenges and fundamental limitations. On summarization tasks, semantic cascades typically come close to the performance of a token-level cascade constructed from their single-best ensemble model, despite evaluation artifacts where several small models paradoxically outperform target large models on ROUGE metrics (Appendix D). Conversely, closed-book QA exposes a core limitation: 3B models substantially outperform ensembles equal in total parameter count. When answers are typically 1-3 tokens, *and* baseline model performance is low—leading to frequent disagreement—semantic similarity may struggle to find signal in such disagreements and only succeed on the small minority of queries where models agree near-exactly.

The pattern suggests a deeper principle. Semantic signals excel precisely where open-ended generation is most challenging: when multiple valid responses exist and quality gradients matter more than binary correctness. This systematic relationship between task/output characteristics and method effectiveness demonstrates that semantic agreement captures a genuinely different dimension of uncertainty than token-level approaches—one more aligned with generation quality assessment in open-ended tasks.

4.3. Resilience to Model Heterogeneity

Perhaps most surprisingly, semantic ensembles demonstrate counter-intuitive behavior: adding weaker models often maintains or improves, rather than degrades, performance. When we replace mT0 Large in the ensemble from Figure 1 with the lower-performing FLAN-T5 Large, AUC improves from 0.6548 to 0.6554—despite the substituted model being substantially worse on average. This paradox resolves when we recognize that semantic agreement values consistency signals over individual model quality.

It can also be illustrated how worse-on-average models provide signal by evaluating a simple setup. Using only Qwen2.5-1.5B outputs while incorporating semantic similarity with Qwen2.5-0.5B for deferral decisions improves AUC to 0.6307 versus 0.6263 for token-level approaches (Appendix C). This isolation experiment proves semantic similarity's advantage stems from superior when-to-defer signals rather than output selection benefits. Weaker models contribute valuable negative evidence—when they disagree with stronger models, this disagreement reliably indicates query difficulty.

5. Discussion, Conclusion, & Future Work

Semantic agreement represents a fundamentally better-suited approach for open-ended cascades. Our findings establish that semantic similarity measures meaning-level consistency, which better aligns with generation quality assessment, compared to token-level confidence—optimized for language modeling objectives. This explains why semantic signals provide superior deferral decisions even with smaller model ensembles, as demonstrated consistently across translation and reading comprehension, while maintaining competitive performance on summarization tasks.

Semantic cascades offer practical advantages with favorable scaling prospects. Through better deferral decisions and resilience to model heterogeneity, semantic cascades achieve more consistent performance across tasks. Unlike token-level approaches requiring model internals, semantic signals function with black-box APIs and remain stable across model updates without retraining. As established by ABC (Kolawole et al., 2024), larger model gaps (e.g., 7B→70B→405B) should benefit more as absolute performance gaps between model tiers narrow and heterogeneity in relative performance across tasks rises. Aside from scaling, we aim to leverage complementary information content by combining semantic and token-level signals in future work.

References

Aggarwal, P., Madaan, A., Anand, A., Potharaju, S. P., Mishra, S., Zhou, P., Gupta, A., Rajagopal, D., Kappaganthu, K., Yang, Y., et al. Automix: Automatically mixing language models. *Advances in Neural Information Processing Systems*, 37:131000–131034, 2024.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Chen, L., Zaharia, M., and Zou, J. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Feng, T., Shen, Y., and You, J. Graphrouter: A graph-based router for llm selections. *arXiv preprint arXiv:2410.03834*, 2024.

Gemma Team. Gemma 3. 2025. URL https://arxiv.org/abs/2503.19786.

Gupta, N., Narasimhan, H., Jitkrittum, W., Rawat, A. S., Menon, A. K., and Kumar, S. Language model cascades: Token-level uncertainty and beyond, 2024.

Jiang, D., Ren, X., and Lin, B. Y. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

Jitkrittum, W., Narasimhan, H., Rawat, A. S., Juneja, J., Wang, Z., Lee, C.-Y., Shenoy, P., Panigrahy, R., Menon, A. K., and Kumar, S. Universal model routing for efficient llm inference. *arXiv preprint arXiv:2502.08773*, 2025.

Kolawole, S., Dennis, D., Talwalkar, A., and Smith, V. Agreement-based cascading for efficient inference. *arXiv preprint arXiv:2407.02348*, 2024.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Narasimhan, H., Jitkrittum, W., Rawat, A. S., Kim, S., Gupta, N., Menon, A. K., and Kumar, S. Faster cascades via speculative decoding. *arXiv preprint arXiv:2405.19261*, 2024.

Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., and Stoica, I. Routellm: Learning to route llms from preference data. In *The Thirteenth International Conference on Learning Representations*, 2024.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Post, M. A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Sellam, T., Das, D., and Parikh, A. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, 2020.

Wang, H., Polo, F. M., Sun, Y., Kundu, S., Xing, E., and Yurochkin, M. Fusing models with complementary expertise. In *The Twelfth International Conference on Learning Representations*, 2023.

Wang, X., Kondratyuk, D., Christiansen, E., Kitani, K. M., Movshovitz-Attias, Y., and Eban, E. Wisdom of committees: An overlooked approach to faster and more accurate models. In *International Conference on Learning Representations*, 2020.

Yue, M., Zhao, J., Zhang, M., Du, L., and Yao, Z. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

A. Experimental Details

Models and Datasets For all Gemma, Llama, and Qwen models, we use their official instruction-tuned versions; FLAN-T5 and mT0 are already instruction-tuned. For consistency, we use the same SBERT model across datasets. For the SBERT model, we use the cased multilingual BERT Base model (Devlin et al., 2018), as provided through HuggingFace, due to its multilingual capabilities, small size, and speed.

On all datasets, we evaluate on the validation split as provided on HuggingFace, or the test split if no validation split is available. Within a dataset, all models are given the same prompt (up to chat templating) in order to better mimic the real world, where live user prompts cannot easily be tuned for best performance on a particular model. We do not employ few-shot prompting strategies as their applicability in more general open-ended text generation is unclear, and their use in this context may not reflect practical usage scenarios. Whether different prompting strategies lead to different behavior and relative performance gains for token-level and semantic methods is an interesting topic for future analysis.

	Qwen2.5-7B	Qwen2.5-3B	Qwen2.5-1.5B	Qwen2.5-0.5B
Parameters	7.62B	3.09B	1.54B	494M

Table 2. Number of Qwen2.5 model parameters, as reported by HuggingFace.

	Llama-3.1-8B	Llama-3.2-3B	Llama-3.2-1B
Parameters	8.03B	3.21B	1.24B

Table 3. Number of Llama model parameters, as reported by HuggingFace.

	Gemma3-1B	FLAN-T5	mT0	BERT Base (multilingual, cased)
Parameters	1.00B	783M	1.23B	179M

Table 4. Number of model parameters, as reported by HuggingFace.

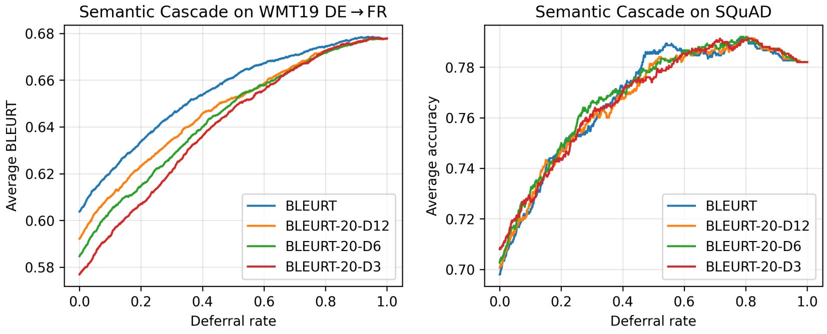
Ensemble Deferral In Section 3.1 it is described how, in the context of ensemble deferral, we set $s = \max_i o_i$. We also tested setting $s = \text{mean}_i o_i$, and observe nearly identical results. This is also true for token-level ensemble cascades.

For token-level cascades, we analyze Chow-Quantile for quantiles $q \in \{0.0, 0.1, ..., 1.0\}$. This provides a balance between expressivity and avoiding spuriously high "Best Quantile" results due to noise.

We use the BLEU implementation as provided by SacreBLEU (Post, 2018).

B. BLEURT Model Sizes

There are several different official BLEURT model sizes available. We use BLEURT-20 in our experiments, which has 30 layers and 579M parameters, but there are also three lossily compressed versions available: BLEURT-20-D12 (12 layers, 167M parameters), BLEURT-20-D6 (6 layers, 45M parameters), and BLEURT-20-D3 (3 layers, 30M parameters). We find that on domains with longer output lengths, the size (quality) of the BLEURT model becomes important, but is largely irrelevant at very small scales (Figure 4).

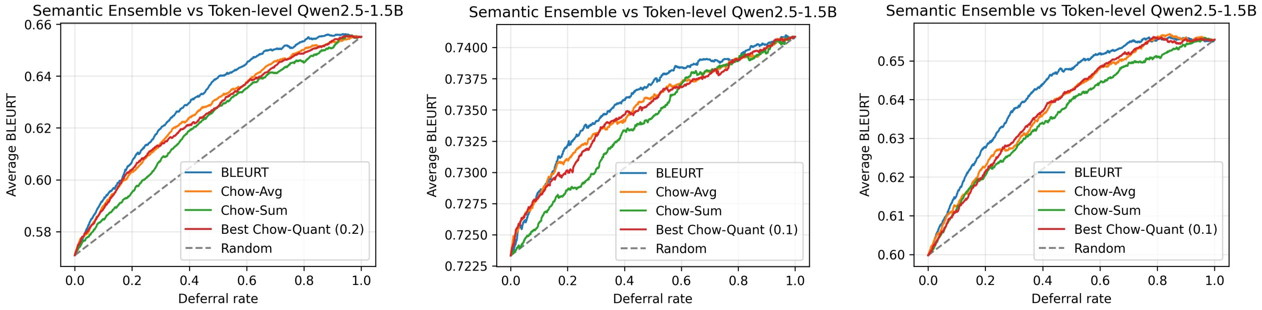


(a)

(b)

Figure 4. **(a)** Comparison of deferral curves for different BLEURT sizes for a semantic cascade of Qwen2.5-1.5B, Gemma3-1B, mT0 Large, deferring to large model Llama-3.1-8B. Smaller sizes of BLEURT lead to worse performance. **(b)** A similar comparison on SQuAD1.1 of the same semantic ensemble, deferring to Qwen2.5-7B. Smaller sizes of BLEURT do not impact cascade performance due to the short nature of responses.

C. Worse-on-Average Models Provide Useful Signal for Deferral



(a) WMT19 DE→FR

(b) WMT14 FR→EN

(c) WMT14 EN→FR

Figure 5. Deferral curve on WMT19 DE→FR, WMT14 FR→EN, and WMT14 EN→FR for a simplest semantic ensemble of Qwen2.5-1.5B and Qwen2.5-0.5B, always using the outputs of Qwen2.5-1.5B when not deferring, plotted with deferral curves for token-level Qwen2.5-1.5B. Both cascades defer to Qwen2.5-7B. In all cases, the deferral curve from this semantic cascade improves over single-model token-level deferral signals, improving for instance the AUC to .6307 on DE→FR over single-model token-level deferral signals (Best (Chow-Avg) AUC: .6263), and significantly over random deferral (AUC: .6130). This demonstrates how, even in a very simple ensemble, semantic similarity with a substantially *worse* model can still provide a strong indication for when deferral is appropriate.

D. Difficulties with Assessing Summarization

Summarization is difficult to fairly assess as a result of the fact that several small models outperform the "target" large models by a large degree (Table 6). In particular, we observe FLAN-T5 Large outperform large models on XLSum and CNN/DailyMail, and mT0 Large outperform large models on XLSum. For this reason, we did not include them in our main analysis. The small models which perform well here typically perform less well on other datasets, and as semantic cascading is typically able to account for their weak performance on other datasets and come close to matching their performance on summarization when they are included in the ensemble, it arguably remains a solid choice.

However, it should be noted that ROUGE scores are highly noisy, imperfect measures of summarization quality. Further, there are often many equally satisfactory ways of summarizing a given article, but there exists only one ground-truth summary per example in these datasets. Additionally, ROUGE does not account for factual inaccuracies. Together, these factors mean that ROUGE scores alone cannot reliably determine summarization quality.

Based on a qualitative analysis we performed on a small sample of model responses, the differences in ROUGE performance appear mostly to be a result of Qwen, Llama, and Gemma models giving more detail (and using complete sentences in the case of CNN/DailyMail, where target summaries are often comprised of multiple sentence fragments), rather than these models outputting strictly "worse" summaries. In other words, there is a stylistic mismatch between their outputs and the targets, but not necessarily a semantic one. Furthermore, we observed more factual inaccuracies in the summaries generated by FLAN-T5 Large and mT0 Large than in those generated by the large models (Table 5). Different prompt styles, particularly few-shot prompting or prompts which otherwise give examples and explicit descriptions of the target style, might better force the models to adhere to the desired style and be able to close this performance gap.

The Dallas Mavericks player accused Ted Kritza of taking the money . He believes Kritza took it from his bank credit line without his permission . In a recorded phone call with Jefferson, Kritza 'confesses to wrongdoing' Recording of the conversation is now in the hands of FBI .	Dallas Mavericks player Richard Jefferson helped FBI find $2million dollars that was taken from his bank credit line without his permission. Jefferson, 34, had reported the crime before the investigation began. In a recorded phone call with Kritza, Jefferson 'confesses to wrongdoing' The recording is now in the hands of the FBI.
NBA player Richard Jefferson helped the FBI recover $2 million after accusing his former business manager, Ted Kritza, of stealing the money from his bank credit line without permission. Jefferson recorded a phone call with Kritza, in which Kritza allegedly confessed to the crime, and the recording is now in the hands of the FBI. Jefferson is seeking to put a hold on a lawsuit from a bank that is seeking part of the $2 million until the federal investigation into Kritza is complete.	Richard Jefferson, a Dallas Mavericks basketball player, assisted the FBI in recovering $2 million that his former business manager, Ted Kritza, allegedly stole from Jefferson's bank credit line. The case is currently under federal investigation, and Jefferson is seeking to delay legal action from a bank that has sued him for a portion of the stolen funds until the investigation is complete.

Table 5. Ground-truth and predicted summaries on CNN/DailyMail. Top left: ground-truth. Top right: FLAN-T5 Large (ROUGE-2: .4583). Bottom left: Llama-3.1-8B (ROUGE-2: .2656). Bottom right: Qwen2.5-7B (ROUGE-2: .09346). Factual inaccuracies are highlighted in red, and superfluous information to the ground truth are highlighted in orange. Ground truth summary is comprised of sentence fragments, whereas models output longer full sentences and sometimes unnecessary detail. Qualitatively, FLAN-T5 exhibits better stylistic matching but higher rates of factual errors.

Ensemble deferral strategies further struggled on these datasets due to large relative differences in baseline model performance. For instance, on CNN/DailyMail, the difference in baseline performance between Qwen2.5-1.5B and Llama-3.2-1B is almost three times as large as the difference in baseline performance between Llama-3.2-1B and Llama-3.2-8B. It should be noted that, on XLSum, *no deferral strategy* significantly improves on random deferral, leading to flat deferral curves and baseline performance becoming the dominating factor (Figure 6).

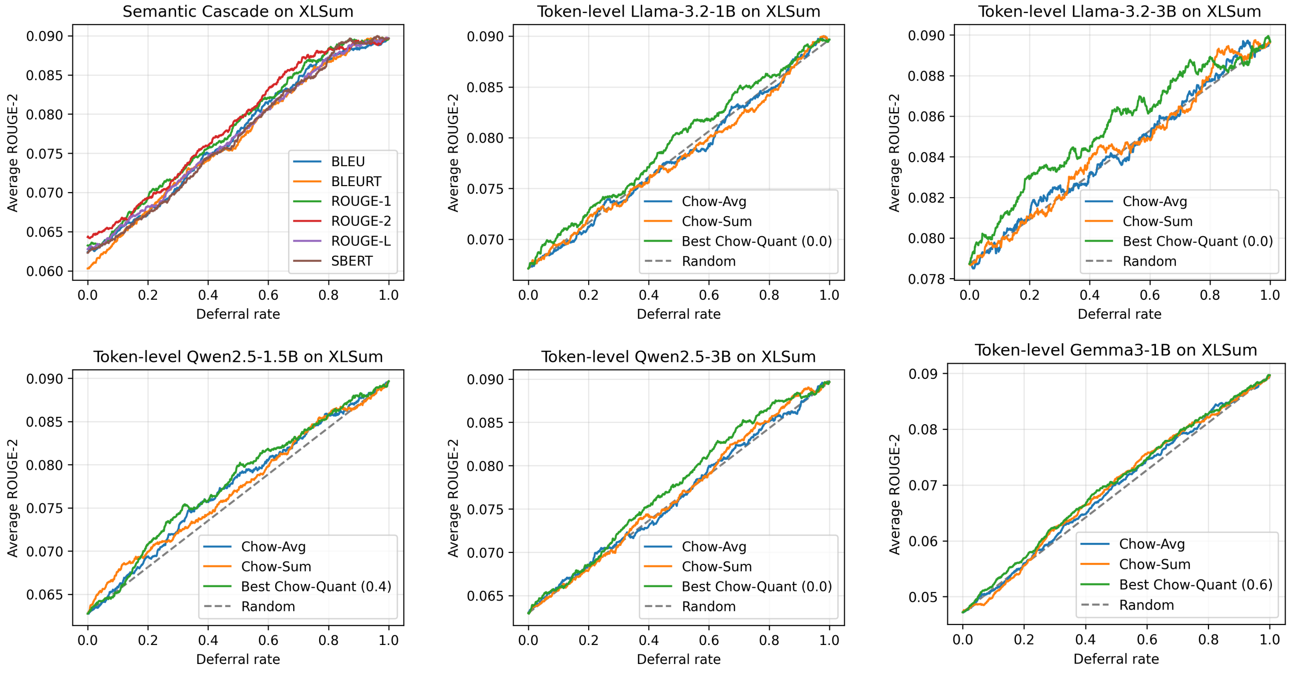


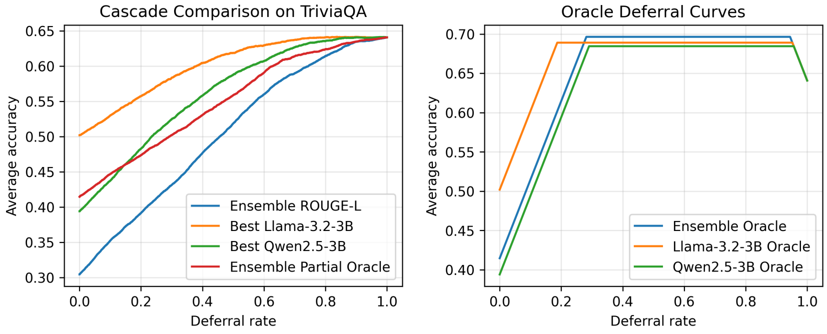
Figure 6. Deferral curves on XLSum. Semantic cascade's ensemble is comprised of Llama-3.2-1B, Qwen2.5-1.5B, Gemma3-1B. All cascades defer to Llama-8B. No cascades improve significantly on random deferral from their single-best ensemble model (only ensemble model, for token-level cascades), hence deferral curves are nearly flat. The semantic cascade has a slightly better curve shape than the token-level cascades, but begins at a lower baseline than its single-best ensemble model (Llama-3.2-1B). This is enough for it to achieve similar AUC (.0786 versus 0.0795), and match or exceed performance at the 60% deferral rate and above.

E. TriviaQA: Performance Gap and Oracle Insights

As noted, closed-book QA poses a challenge for semantic similarity: 3B models outperform all ensembles of the same total size by wide margins. For example, a token-level cascade with Llama-3.2-3B deferring to Llama-3.1-8B achieves an AUC of .6016, compared to an AUC of just .5044 for a semantic ensemble of the three best 1B models. Even still, semantic cascades typically nearly match, or exceed, their single-best ensemble model's performance.

It is tempting to assume the performance gap between 3B parameter models and ensembles of 1B parameter models arises because the 3B parameter models simply encode more information than the smaller ensemble models. Surprisingly, however, the oracle deferral curves for this task suggest that this explanation is incomplete and that other dynamics are at play (Figure 7). We hypothesize that the gap in performance originates not simply from a lack of encoded information, but also due to the difficulty of the ensemble deferral task, which requires not only identifying when to defer to a large model, but also selecting the best answer from the ensemble's multiple outputs when not deferring. This challenge is particularly pronounced in TriviaQA, where answers are short and baseline ensemble models perform poorly—leading to frequent disagreement from which semantic similarity struggles to extract meaningful signal.

To illustrate this, we plot oracle curves representing the optimal performance achievable with perfect knowledge of each model's output quality on every example. For cascades involving a single baseline model, the oracle curve is computed by deferring based on the score difference between the large and small models. In the case of ensemble-based cascades, the oracle defers based on the score difference between the large model and the best-performing output among the small models in the ensemble; when not deferring, it selects the best small-model output. Additionally, we define a partial oracle, which always chooses the best output from the small models but relies on semantic similarity to decide whether to defer.



(a)

(b)

Figure 7. **(a)** Deferral curves on TriviaQA for token-level cascades and a semantic cascade with ensemble models Qwen2.5-1.5B, Llama1B, Gemma3-1B. The partial oracle always selects the best output among the ensembles but uses semantic similarity as a deferral rule. All cascades defer to Llama-3.1-8B. **(b)** Oracle deferral curves. Half of the initial performance gap between the ensemble method and Llama-3.2-3B is closed simply by selecting the best ensemble outputs, and the ensemble oracle lies strictly above the Qwe2.5-3B oracle, demonstrating that differences in model-encoded knowledge cannot by themselves adequately explain the observed differences in baseline performance.

F. Additional Experimental Results

F.1. Baseline Model Performances

Model	WMT19 DE→FR (BLEURT)	WMT14 FR→EN (BLEURT)	WMT14 EN→FR (BLEURT)	CNN/DailyMail (ROUGE-2)	XLSum (ROUGE-2)	SQuAD1.1 (Accuracy)	TriviaQA (Accuracy)
Qwen2.5-7B	.6552	.7409	.6555	.1222	.0787	.7820	.5075
Qwen2.5-3B	.6109	.7298	.6293	.1188	.0630	.6920	.3940
Qwen2.5-1.5B	.5708	.7233	.5998	.1033	.0628	.6967	.3022
Qwen2.5-0.5B	.4026	.6831	.4835	.1016	.0478	.4653	.1069
Llama-3.1-8B	.6778	.7400	.6817	.1443	.0897	.7427	.6407
Llama-3.2-3B	.6168	.7270	.6480	.1377	.0787	.6887	.5019
Llama-3.2-1B	.4549	.6694	.5652	.1329	.0671	.3747	.2280
mT0 Large	.4467	.6636	.5120	.1325	.1060	.7767	.0608
FLAN-T5 Large	.3963	.6883	.4534	.1915	.1597	.5973	.1449
Gemma3-1B	.5670	.7090	.6234	.0862	.0472	.5247	.1982

Table 6. Baseline model performances. At each model tier (roughly 1B, 3B, 7-8B), no single model dominates across datasets.

F.2. AUC Values

F.2.1. WMT19 DE→FR

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.6444	0.6514	**0.6563**	0.6557	0.2
Qwen2.5-1.5B	0.6243	0.6342	**0.6397**	0.6391	0.3
Qwen2.5-0.5B	0.5402	0.5544	**0.5588**	0.5576	0.4
Llama-3.2-3B	0.6473	0.6542	**0.6587**	0.6580	0.4
Llama-3.2-1B	0.5663	0.5786	**0.5881**	0.5860	0.2
mT0 Large	0.5622	0.5792	**0.5827**	0.5814	0.4
FLAN-T5 Large	0.5370	0.5606	**0.5619**	0.5609	0.2
Gemma3-1B	0.6224	0.6325	**0.6388**	0.6386	0.3

Table 7. AUC values on WMT19 DE→FR for token-level cascades. All cascades defer to Llama-3.1-8B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Gemma3-1B, Qwen2.5-1.5B	0.6445	**0.6455**	0.6448	0.2
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B	**0.6330**	0.6315	0.6328	0.2
Gemma3-1B, Qwen2.5-1.5B, mT0 Large	**0.6345**	0.6320	0.6326	0.2
Gemma3-1B, Qwen2.5-1.5B, FLAN-T5 Large	**0.6297**	0.6247	0.6247	0.1
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	**0.6181**	0.6147	0.6164	0.2

Table 8. AUC values on WMT19 DE→FR for token-level ensemble cascades. All cascades defer to Llama-3.1-8B.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B	0.6331	0.6330	0.6342	0.6354	**0.6508**	0.6403
Gemma3-1B, Qwen2.5-1.5B, mT0 Large	0.6357	0.6381	0.6375	0.6387	**0.6548**	0.6418
Gemma3-1B, Qwen2.5-1.5B, FLAN-T5 Large	0.6337	0.6362	0.6369	0.6371	**0.6554**	0.6426
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.6167	0.6203	0.6192	0.6207	**0.6368**	0.6272
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.6339	0.6337	0.6349	0.6350	**0.6552**	0.6428
Gemma3-1B, Qwen2.5-1.5B, mT0 Large, FLAN-T5 Large	0.6331	0.6339	0.6344	0.6354	**0.6572**	0.6417

Table 9. AUC values on WMT19 DE→FR for various semantic cascades. All cascades defer to Llama-3.1-8B.

F.2.2. WMT14 FR→EN

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.7353	0.7357	0.7365	**0.7367**	0.9
Qwen2.5-1.5B	0.7321	0.7339	**0.7350**	0.7348	0.1
Qwen2.5-0.5B	0.7120	0.7157	**0.7191**	0.7185	0.2
Llama-3.2-3B	0.7339	0.7344	0.7361	**0.7365**	0.5
Llama-3.2-1B	0.7051	0.7046	0.7115	**0.7119**	1.0
mT0 Large	0.7023	0.7070	**0.7121**	0.7115	0.1
FLAN-T5 Large	0.7146	0.7186	**0.7233**	0.7230	0.1
Gemma3-1B	0.7250	0.7264	0.7284	**0.7292**	0.5

Table 10. AUC values on WMT14 FR→EN for token-level cascades. All cascades defer to Qwen2.5-7B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Gemma3-1B, Qwen2.5-1.5B	0.7312	0.7323	**0.7344**	0.7
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B	0.7247	0.7259	**0.7265**	0.0
Gemma3-1B, Qwen2.5-1.5B, mT0 Large	0.7268	0.7267	**0.7280**	0.2
Gemma3-1B, Qwen2.5-1.5B, FLAN-T5 Large	0.7305	0.7311	**0.7312**	0.2
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.7218	**0.7222**	0.7219	0.0

Table 11. AUC values on WMT14 FR→EN for token-level ensemble cascades. All cascades defer to Qwen2.5-7B.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B	0.7322	0.7325	0.7317	0.7323	**0.7350**	0.7339
Gemma3-1B, Qwen2.5-1.5B, mT0 Large	0.7339	0.7337	0.7323	0.7334	**0.7369**	0.7348
Gemma3-1B, Qwen2.5-1.5B, FLAN-T5 Large	0.7343	0.7344	0.7336	0.7345	**0.7380**	0.7356
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.7287	0.7284	0.7284	0.7283	**0.7327**	0.7306
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.7325	0.7328	0.7318	0.7328	**0.7359**	0.7337
Gemma3-1B, Qwen2.5-1.5B, mT0 Large, FLAN-T5 Large	0.7330	0.7337	0.7323	0.7335	**0.7375**	0.7346

Table 12. AUC values on WMT14 FR→EN for various semantic cascades. All cascades defer to Qwen2.5-7B.

F.2.3. WMT14 EN→FR

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.6555	0.6606	**0.6636**	0.6633	0.2
Qwen2.5-1.5B	0.6407	0.6514	**0.6535**	0.6528	0.2
Qwen2.5-0.5B	0.5826	0.5996	**0.6034**	0.6022	0.3
Llama-3.2-3B	0.6648	0.6693	**0.6722**	0.6719	0.4
Llama-3.2-1B	0.6234	0.6345	**0.6415**	0.6405	0.5
mT0 Large	0.5969	0.6144	**0.6170**	0.6157	0.4
FLAN-T5 Large	0.5676	0.5889	**0.5906**	0.5899	0.3
Gemma3-1B	0.6526	0.6602	**0.6626**	0.6624	0.1

Table 13. AUC values on WMT14 EN→FR for token-level cascades. All cascades defer to Llama-3.1-8B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Gemma3-1B, Qwen2.5-1.5B	**0.6637**	0.6621	0.6635	0.1
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B	**0.6621**	0.6600	0.6614	0.1
Gemma3-1B, Qwen2.5-1.5B, mT0 Large	**0.6574**	0.6535	0.6531	0.2
Gemma3-1B, Qwen2.5-1.5B, FLAN-T5 Large	**0.6486**	0.6467	0.6458	0.1
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	**0.6500**	0.6467	0.6458	0.1

Table 14. AUC values on WMT14 EN→FR for token-level ensemble cascades. All cascades defer to Llama-3.1-8B.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B	0.6569	0.6574	0.6570	0.6582	**0.6685**	0.6602
Gemma3-1B, Qwen2.5-1.5B, mT0 Large	0.6571	0.6579	0.6569	0.6589	**0.6707**	0.6618
Gemma3-1B, Qwen2.5-1.5B, FLAN-T5 Large	0.6550	0.6556	0.6550	0.6558	**0.6714**	0.6588
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.6454	0.6466	0.6464	0.6473	**0.6605**	0.6499
Gemma3-1B, Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.6588	0.6592	0.6587	0.6591	**0.6723**	0.6618
Gemma3-1B, Qwen2.5-1.5B, mT0 Large, FLAN-T5 Large	0.6541	0.6548	0.6537	0.6556	**0.6719**	0.6593

Table 15. AUC values on WMT14 EN→FR for various semantic cascades. All cascades defer to Llama-3.1-8B.

F.2.4. CNN/DailyMail

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.1315	0.1325	0.1322	**0.1326**	0.3
Qwen2.5-1.5B	0.1238	0.1255	0.1259	**0.1265**	0.6
Qwen2.5-0.5B	0.1230	0.1234	0.1262	**0.1270**	0.6
Llama-3.2-3B	0.1410	**0.1422**	0.1418	0.1420	0.4
Llama-3.2-1B	0.1386	0.1408	0.1406	**0.1409**	0.1
mT0 Large	0.1384	0.1408	0.1420	**0.1423**	0.2
Gemma3-1B	**0.1152**	0.1143	0.1140	0.1146	0.8

Table 16. AUC values on CNN/DailyMail for token-level cascades. All cascades defer to Llama-3.1-8B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Llama-3.2-1B, mT0 Large	**0.1454**	0.1423	0.1447	0.1
Llama-3.2-1B, Qwen2.5-1.5B	**0.1383**	0.1361	0.1373	0.5
Qwen2.5-1.5B, Gemma3-1B	0.1239	0.1226	**0.1251**	0.8
Llama-3.2-1B, mT0 Large, Qwen2.5-1.5B	**0.1428**	0.1392	0.1415	0.2

Table 17. AUC values on CNN/DailyMail for token-level ensemble cascades. All cascades defer to Llama-3.1-8B.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
Llama-3.2-1B, mT0 Large, Qwen2.5-1.5B	0.1382	0.1367	**0.1390**	0.1378	0.1384	0.1365
Llama-3.2-1B, mT0 Large, Qwen2.5-0.5B	0.1380	**0.1388**	0.1383	0.1386	0.1376	0.1367
Qwen2.5-1.5B, Qwen2.5-0.5B, Gemma3-1B	0.1248	0.1231	**0.1262**	0.1244	0.1222	0.1243

Table 18. AUC values on CNN/DailyMail for various semantic cascades. All cascades defer to Llama-3.1-8B.

F.2.5. XLSum

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.0763	0.0767	0.0765	**0.0777**	0.0
Qwen2.5-1.5B	0.0762	0.0773	0.0775	**0.0780**	0.4
Qwen2.5-0.5B	0.0687	0.0688	0.0689	**0.0692**	0.4
Llama-3.2-3B	0.0842	0.0844	0.0843	**0.0854**	0.0
Llama-3.2-1B	0.0784	0.0781	0.0782	**0.0795**	0.0
Gemma3-1B	0.0684	0.0695	0.0692	**0.0701**	0.6

Table 19. AUC values on XLSum for token-level cascades. All cascades defer to Llama-3.1-8B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Llama-3.2-1B, Qwen2.5-1.5B	0.0783	0.0781	**0.0786**	0.0
Llama-3.2-1B, Qwen2.5-1.5B, Gemma3-1B	0.0745	0.0761	**0.0767**	1.0
Llama-3.2-1B, Qwen2.5-0.5B, Gemma3-1B	0.0721	0.0730	**0.0735**	0.1

Table 20. AUC values on XLSum for token-level ensemble cascades. All cascades defer to Llama-3.1-8B.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
Llama-3.2-1B, Qwen2.5-1.5B, Gemma3-1B	0.0773	0.0783	**0.0786**	0.0773	0.0769	0.0770
Llama-3.2-1B, Qwen2.5-0.5B, Gemma3-1B	0.0734	0.0745	**0.0748**	0.0742	0.0739	0.0747
Qwen2.5-1.5B, Qwen2.5-0.5B, Gemma3-1B	0.0730	0.0729	0.0740	0.0736	0.0745	**0.0746**

Table 21. AUC values on XLSum for various semantic cascades. All cascades defer to Llama-3.1-8B.

F.2.6. SQuAD1.1

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.7370	0.7583	**0.7592**	0.7584	0.1
Qwen2.5-1.5B	0.7393	**0.7616**	0.7574	0.7580	0.0
Qwen2.5-0.5B	0.6237	**0.6650**	0.6603	0.6590	0.2
Llama-3.2-3B	0.7353	0.7534	0.7558	**0.7572**	0.1
Llama-3.2-1B	0.5783	0.6154	**0.6269**	0.6251	0.3
mT0 Large	0.7793	0.7935	0.7949	**0.7961**	0.1
FLAN-T5 Large	0.6897	**0.7330**	0.6671	0.7183	0.0
Gemma3-1B	0.6533	**0.6711**	0.6669	0.6695	0.0

Table 22. AUC values on SQuAD1.1 for token-level cascades. All cascades defer to Qwen2.5-7B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
mT0 Large, Qwen2.5-1.5B	**0.7881**	0.7852	0.7851	0.0
mT0 Large, Qwen2.5-1.5B, FLAN-T5 Large	0.7520	**0.7787**	0.7775	0.0
Qwen2.5-1.5B, FLAN-T5 Large, Gemma3-1B	0.7090	0.7229	**0.7348**	0.0

Table 23. AUC values on SQuAD1.1 for various token-level ensemble cascades. All cascades defer to Qwen2.5-7B. Heterogeneous baseline performance significantly harms AUC for token-level ensemble cascades.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
mT0 Large, Qwen2.5-1.5B, FLAN-T5 Large	0.7947	0.7959	0.7879	0.7966	0.7942	**0.7969**
mT0 Large, Qwen2.5-1.5B, FLAN-T5 Large, Llama-3.2-1B	0.7852	0.7927	0.7913	**0.7944**	0.7847	0.7872
Qwen2.5-1.5B, FLAN-T5 Large, Gemma3-1B	0.7654	0.7678	0.7569	**0.7685**	0.7678	0.7682
Qwen2.5-1.5B, FLAN-T5 Large, Llama-3.2-1B	0.7589	0.7622	0.7526	0.7624	0.7606	**0.7635**

Table 24. AUC values on SQuAD1.1 for various semantic cascades. All cascades defer to Qwen2.5-7B. ROUGE-L and SBERT perform slightly better than other measures of semantic similarity. In SQuAD, there can be multiple "correct" answers, typically excerpts of different length from the same passage with equivalent semantic meaning. As such, these metrics capture when models output different answers which are all still correct. Additionally, it can be observed that adding even the significantly weaker Llama3.2-1B to an ensemble has a very minor impact on the AUC, suggesting resilience to heterogeneous baseline model performance.

F.2.7. TRIVIAQA

Model	Random	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-3B	0.5174	**0.5624**	0.5600	0.5603	0.1
Qwen2.5-1.5B	0.4715	**0.5113**	0.5099	0.5096	0.1
Llama-3.2-3B	0.5713	**0.6016**	0.5991	0.5998	0.1
Llama-3.2-1B	0.4344	**0.4589**	0.4544	0.4545	0.1
mT0 Large	0.3504	0.3540	0.3528	**0.3560**	0.2
FLAN-T5 Large	0.3928	0.4146	0.4093	**0.4149**	0.0
Gemma3-1B	0.4194	**0.4448**	0.4441	0.4435	0.1

Table 25. AUC values on TriviaQA for token-level cascades. All cascades defer to Llama-3.1-8B. The Random column represents the expected AUC if queries are deferred uniformly at random.

Ensemble	Chow-Sum	Chow-Avg	Best Chow-Quantile	Best Quantile
Qwen2.5-1.5B, Llama-3.2-1B	0.4996	**0.5031**	0.5008	0.1
Llama-3.2-1B, Gemma3-1B	0.4634	**0.4676**	0.4649	0.0
Qwen2.5-1.5B, Llama-3.2-1B, Gemma3-1B	0.4980	0.4961	**0.4989**	0.1
Qwen2.5-1.5B, Llama-3.2-1B, FLAN-T5 Large	0.4851	0.4876	**0.4895**	0.0
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.4630	0.4529	**0.4708**	0.0

Table 26. AUC values on TriviaQA for token-level ensemble cascades. All cascades defer to Llama-3.1-8B.

Ensemble	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	SBERT
Qwen2.5-1.5B, Llama-3.2-1B, Gemma3-1B	0.4993	0.5043	0.4803	**0.5044**	0.4957	0.4887
Qwen2.5-1.5B, Llama-3.2-1B, FLAN-T5 Large	0.4911	0.4966	0.4782	**0.4966**	0.4838	0.4824
Qwen2.5-1.5B, Llama-3.2-1B, mT0 Large	0.4862	0.4891	0.4772	**0.4891**	0.4708	0.4692
Llama-3.2-1B, Gemma3-1B, FLAN-T5 Large	0.4598	**0.4659**	0.4400	0.4657	0.4600	0.4549
Qwen2.5-1.5B, Llama-3.2-1B, Gemma3-1B, FLAN-T5 Large	0.4970	**0.5014**	0.4781	0.5011	0.4950	0.4870

Table 27. AUC values on TriviaQA for various semantic cascades. All cascades defer to Llama-3.1-8B.