Beyond Grey-Box Assumptions: Uncertainty-Guided Example Selection for Black-Box Language Models

Egor Krasheninnikov[1], **Zainab Afolabi**[1], **Giuseppe Mascellaro**[1], **Salvatore Radosta**[2]

[1]Amazon Web Services, [2]Prometeia

egorkr@amazon.co.uk, zafolabi@amazon.co.uk, giusmasc@amazon.it, salvatore.radosta@prometeia.com

Abstract

In-context learning (ICL) with Large Language Models has been historically effective, but performance depends heavily on demonstration quality while annotation budgets remain constrained. Existing uncertainty-based selection methods like Cover-ICL achieve strong performance through logit-based uncertainty estimation, but most production LLMs operate as black-box APIs where internal states are inaccessible. This paper investigates whether effective uncertainty guided example selection can be maintained under black-box constraints by developing a consistency-based uncertainty estimation using only output observations. We evaluate five active learning methods (random, hardest, VoteK, fast-VoteK, and Cover-ICL) across six benchmark datasets under both grey-box and black-box settings. Experiments reveal paradigm-dependent strategies: grey-box achieves best performance with Cover-ICL (62.34% average accuracy), while black-box favors hardest selection (68.71% average accuracy), but no single method dominates across all datasets. Our framework enables selecting appropriate uncertainty estimation strategies based on model accessibility constraints in practical deployment scenarios.

Introduction

Large language models (LLMs) have demonstrated remarkable capabilities through in-context learning (Brown et al. 2020; Wei et al. 2023), where model performance heavily depends on the quality and diversity of demonstration examples (Liu et al. 2022; Min et al. 2022). However, collecting high-quality labeled examples is expensive and time-consuming, creating a critical need for efficient example selection strategies that maximize performance gains under limited annotation budgets.

Recent advances in few-shot example selection have shown significant promise through uncertainty-guided approaches in grey-box settings, where researchers assume access to model logits and loss information (Su et al. 2022; Zhang, Feng, and Tan 2022). Notable work like Cover-ICL (Mavromatis et al. 2023, 2024) has demonstrated effective uncertainty quantification by leveraging logit-based uncertainty estimation to identify the most informative examples for annotation. However, most production LLM applications rely exclusively on black-box APIs (GPT-4, Claude, Gemini) where internal model states are completely inaccessible. This accessibility constraint forces practitioners to abandon logit-based uncertainty methods, despite their proven effectiveness in research settings.

The fundamental question becomes: *Can effective uncertainty-based example selection be achieved when we can only observe model outputs, not internal representations or logits?* This is not a design choice but an imposed limitation as API providers restrict access to model internals for competitive and safety reasons (Shorinwa et al. 2025).

We systematically evaluate whether effective demonstration selection remains feasible under black-box constraints by extending the Cover-ICL (Mavromatis et al. 2024) framework with consistency-based uncertainty estimation methods that quantify model confidence through response disagreement across multiple samples. Our systematic comparison between black-box and grey-box versions of identical model architectures reveals important insights about uncertainty estimation reliability across access paradigms. We provide the first comprehensive evaluation of how access constraints affect uncertainty estimation quality and downstream in-context learning performance across six datasets spanning three NLP tasks.

Our contributions include: (1) Consistency-based uncertainty estimation methods enabling effective example selection without logit access (Huang et al. 2025), extending Cover-ICL principles to production-ready black-box settings; (2) First systematic evaluation of how model accessibility affects uncertainty estimation quality across multiple tasks and architectures, quantifying when black-box approaches can match grey-box performance and when additional model access becomes essential.

Related Work

Traditional active learning optimizes model performance through parameter updates (Lewis and Gale 1994; Settles 2009), but fine-tuning large language models is computationally expensive. This has driven a shift toward in-context learning (ICL), where models adapt through demonstration examples rather than parameter modifications (Zhang, Feng, and Tan 2022; Liu et al. 2022). ICL selection strategies include uncertainty-based methods prioritizing low-confidence examples (Lewis and Gale 1994; Huang et al. 2024; Gal and Ghahramani 2016), diversity-based approaches ensuring

broad input coverage (Su et al. 2022), and hybrid methods like VoteK combining both principles (Su et al. 2022). Cover-ICL (Mavromatis et al. 2024) unifies these through a graph-based framework applying Maximum Coverage for selection, but requires grey-box logit access, limiting applicability to proprietary APIs.

When internal states are inaccessible, uncertainty estimation requires measuring response consistency across multiple invocations (Huang et al. 2024; Wei et al. 2023), creating computational overhead compared to single-query grey-box approaches. LLM deployment encompasses distinct accessibility levels: grey-box models provide logit access for direct uncertainty quantification (Ma et al. 2025), while black-box models restrict access to text outputs only (Huang et al. 2024). No systematic comparison exists between grey-box and black-box uncertainty estimation using identical architectures for ICL selection. Our work addresses this gap by providing the first systematic evaluation comparing logit-based and consistency-based uncertainty estimation across six datasets spanning three NLP tasks using multiple model families.

Experimental Setup

Research Questions

Our experimental analysis addresses two key research questions examining uncertainty estimation across model accessibility paradigms:

RQ1: How do uncertainty estimation capabilities differ between grey-box models (with logit access) and black-box models (using consistency measures) when applying identical sample selection methods?

RQ2: Which uncertainty quantification approach – logit-based confidence or consistency-based measures – provides more reliable sample selection for few-shot learning?

Framework and Methodology

We extend the Cover-ICL experimental framework (Mavromatis et al. 2024) to enable systematic comparison between black-box and grey-box versions of identical underlying models. Our approach simulates an active learning scenario where, given a pool of unlabeled data $\mathcal{U}$ and a limited annotation budget B, we select B examples to annotate such that few-shot performance on test data is maximized.

We treat the training set as an unlabeled pool $\mathcal{U}$, hiding labels during selection and revealing them only after annotation. Our experimental pipeline follows four sequential stages: **(i) Uncertainty estimation** - we compute uncertainty scores for all examples in $\mathcal{U}$ using either logit-based (grey-box) or consistency-based (black-box) methods; **(ii) Active selection** - each method selects $B = 20$ examples from $\mathcal{U}$ based on uncertainty and/or diversity criteria; **(iii) Few-shot retrieval** - for each test example, we retrieve the top-5 most similar annotated examples using sentence-transformers/all-mpnet-base-v2 (Song et al. 2020); **(iv) Evaluation** - we measure accuracy on the test set using these 5-shot demonstrations.

We compare five selection strategies: random sampling as baseline, uncertainty-based methods (Hardest (Lewis and Gale 1994)), diversity-based approaches (VoteK and Fast-VoteK (Su et al. 2022)), and uncertainty+diversity methods (Cover-ICL (Mavromatis et al. 2024)).

Active sampling methods We evaluate five selection strategies across both paradigms: **Random**, **Hardest** (Lewis and Gale 1994), **VoteK** (Su et al. 2022), **Fast-VoteK** (Su et al. 2022), and **Cover-ICL** (Mavromatis et al. 2024). Random provides baseline sampling from the training set. Hardest selects examples with highest uncertainty. VoteK combines uncertainty binning with diversity sampling. Fast-VoteK removes uncertainty computation for improved efficiency. Cover-ICL applies graph-based Maximum Coverage optimization to select examples whose neighborhoods cover the most hard examples while avoiding redundancy.

Grey-box paradigm In the grey-box setting, we leverage base models with full logit accessibility. Since base models require demonstration examples to understand task format, we construct 5-shot ICL prompts using randomly selected training examples with known labels. For uncertainty estimation, we use different random 5-shot examples for each unlabeled example x_i to generate prompts and obtain model predictions. We quantify uncertainty using the negative log-likelihood (NLL) across all possible classification labels, where higher NLL values correspond to greater uncertainty and lower model confidence.

We rank examples by uncertainty and designate the top θN examples as "hard examples" for graph-based selection, where $\theta \in [0, 1]$ denotes the portion of examples considered as hard ones (default $\theta = 0.5$, meaning the top 50% most uncertain examples). Uncertainty is computed as:

$$\text{uncertainty}(x_i) = \min_{j \in \mathcal{Y}} \text{NLL}(y_j | x_i, \text{5-shot context})$$

where $\mathcal{Y}$ represents the label space and $\text{NLL}(y_j | x_i, \text{5-shot context})$ denotes the negative log-likelihood of label y_j given example x_i and the 5-shot demonstration context.

This single-pass approach makes grey-box estimation computationally efficient but requires model deployments that expose internal states, limiting applicability to proprietary APIs.

Black-box paradigm For black-box models, we employ instruction-tuned models accessed via message-based templates without internal probability access. Since instruction-tuned models are aligned to follow task instructions without requiring demonstration examples, we adapt prompts with explicit task clarification for 0-shot inference. We estimate uncertainty through consistency analysis across repeated queries. For each unlabeled example x_i, we generate $K = 5$ independent predictions using temperature $T = 0.7$ with identical instruction-based prompts. Unlike grey-box models where we compute losses across all possible labels, we let the model generate freely and apply postprocessing to extract labels from the generated text when present, otherwise randomly assigning labels as predictions. Uncertainty is quantified as the normalized disagreement rate:

Method	AGNews	TREC	SST2	RTE	MRPC	MNLI	Avg
Random	46.17	**43.64**	67.41	59.21	63.39	45.14	54.82
Hardest	57.03	39.23	65.29	58.48	61.33	43.97	54.22
VoteK	57.53	42.97	66.86	**59.43**	62.28	42.63	55.28
Fast VoteK	53.24	33.54	67.08	57.37	62.78	43.58	52.93
Cover-ICL	**59.26**	33.77	**70.42**	58.09	**63.39**	**47.66**	**56.31**

Table 1: Grey-box method performance using logit-based uncertainty measures. Results are averaged across Gemma, Llama, and Qwen model families.

Method	AGNews	TREC	SST2	RTE	MRPC	MNLI	Avg
Random	66.40	23.44	81.25	71.04	**55.25**	**55.30**	58.78
Hardest	**69.42**	22.38	81.64	**72.60**	54.74	54.52	**59.22**
VoteK	67.80	22.43	80.69	70.20	53.63	54.85	58.27
Fast VoteK	67.80	**24.22**	80.86	71.26	53.13	52.68	58.33
Cover-ICL	66.52	21.82	**81.98**	70.87	53.85	54.46	58.25

Table 2: Black-box method performance using consistency-based uncertainty measures with instruct-tuned models. Results are averaged across Gemma, Llama, and Qwen model families.

$$\text{uncertainty}(x_i) = \frac{|\text{unique responses}| - 1}{\max(1, K - 1)}$$

This metric ranges from 0 (complete consensus across all K samples) to 1 (maximum disagreement with all distinct responses). The approach captures epistemic uncertainty via response inconsistency, enabling uncertainty estimation for any generative model accessible through text-based APIs without requiring logit access.

Datasets and Models

We evaluate our framework across six diverse datasets, including Topic Classification (AGNews (Zhang, Zhao, and LeCun 2016), TREC (Hovy et al. 2001)), Sentiment Analysis (SST2 (Socher et al. 2013)), Natural Language Inference (RTE (Giampiccolo et al. 2008), MRPC (Dolan, Quirk, and Brockett 2004), MNLI (Williams, Nangia, and Bowman 2018)), ensuring comprehensive coverage of different NLP task complexities and domain characteristics. We use the standard train/test splits for each dataset, where the training set serves as the unlabeled pool $\mathcal{U}$ for uncertainty estimation and active selection, and we evaluate performance on the standard test sets using accuracy as the primary metric.

Our model selection encompasses three prominent LLM families in both base and instruction-tuned variants: (1) **Gemma-3 series:** 1B and 4B parameter models (google/gemma-3-) (Team 2025a). (2) **Qwen3 series:** 0.6B, 8B and 14B parameter models (Qwen/Qwen3–Base) (Team 2025b). (3) **Llama series:** Llama-3.2 (1B, 3B) models (meta-llama/Llama-3.2) (Aaron Grattafiori 2024).

This experimental design enables systematic comparison of confidence-based and consistency-based uncertainty estimation across model scales and accessibility paradigms while maintaining controlled conditions for robust analysis.

Experimental Results and Analysis

We evaluated uncertainty estimation techniques across different active learning methods and model accessibility paradigms using seven representative models: gemma-3-1b, gemma-3-4b, qwen3-0.6b, qwen3-8b, qwen3-14b, llama-3.2-1b, and llama-3.2-3b. Our analysis extends the Cover-ICL framework to compare grey box (confidence) and black box (consistency) uncertainty estimation across all datasets (tables 1 and 2). While TREC results are included in dataset-level analysis, instruct prompt compatibility issues led to its exclusion from model-family comparisons.

Grey Box Performance (Confidence Metric)

Table 1 presents dataset-level performance across six benchmarks. Cover-ICL achieves the highest average accuracy (56.31%), demonstrating particular strength on sentiment analysis tasks (SST2: 70.42%).

Table 3 presents accuracy across all evaluated methods and models. Cover-ICL achieves the highest overall average accuracy of 62.34%, though results vary by model. Most notably, Cover-ICL shows substantial improvements on qwen3-0.6b (71.39% vs. 59.53-62.66% for other methods), representing a 9-12% performance gain. This finding is particularly significant as it demonstrates that advanced uncertainty estimation techniques can enable smaller models to achieve performance levels comparable to larger instruct-tuned variants.

For larger models (qwen3-8b, qwen3-14b), Cover-ICL maintains competitive performance (79.69% and 79.92% respectively) but shows diminishing relative improvements compared to baseline methods. The gemma family shows mixed results, with random selection slightly outperforming Cover-ICL on gemma-3-4b (50.55% vs. 50.16%), suggesting that uncertainty-based methods provide less consistent gains for this architecture.

Black Box Performance (Consistency Metric)

Table 2 presents dataset-level results across six benchmarks. Hardest selection achieves optimal performance (59.22% average), outperforming Cover-ICL (58.25%). Cover-ICL demonstrates strong results on sentiment analysis tasks (SST2: 81.98%).

Method	gemma-3-1b	gemma-3-4b	qwen3-0.6b	qwen3-8b	qwen3-14b	llama-3.2-1b	llama-3.2-3b	Average*
Random	51.56	**50.55**	60.06	77.81	78.90	47.50	51.87	59.75
Hardest	**53.44**	48.28	59.53	79.14	78.59	47.11	53.51	59.94
VoteK	53.28	49.61	62.03	**79.76**	79.30	**48.05**	49.45	60.21
Fast VoteK	49.61	48.28	62.66	79.06	78.75	46.56	48.83	59.11
Cover-ICL	53.13	50.16	**71.39**	79.69	**79.92**	47.34	**54.77**	**62.34**

* Average excludes TREC dataset

Table 3: Grey-box performance using logit-based uncertainty measures across Gemma, Llama, and Qwen model families.

Method	gemma-3-1b	gemma-3-4b	qwen3-0.6b	qwen3-8b	qwen3-14b	llama-3.2-1b	llama-3.2-3b	Average*
Random	63.75	76.56	67.03	81.64	83.20	39.84	65.08	68.16
Hardest	**65.16**	75.86	**68.44**	**82.66**	83.28	40.31	65.24	**68.71**
VoteK	60.70	77.11	66.25	82.34	83.05	39.69	65.70	67.83
Fast VoteK	63.21	75.78	65.24	82.50	**83.98**	38.52	65.55	67.82
Cover-ICL	63.52	**77.34**	63.59	82.58	**83.98**	**41.80**	**66.33**	68.45

* Average excludes TREC dataset

Table 4: Black-box performance using consistency-based uncertainty measures across Gemma, Llama, and Qwen model families.

Table 4 reveals distinct patterns in black box uncertainty estimation, with hardest selection achieving the highest overall performance at 68.71% average accuracy. Unlike grey box settings, the optimal method varies significantly by model family. For larger models (qwen3-8b, qwen3-14b), hardest selection achieves strong performance (82.66% and 83.28% respectively), though Cover-ICL matches or exceeds it on qwen3-14b (83.98%).

However, Cover-ICL shows selective effectiveness in the llama family, consistently outperforming baseline methods: llama-3.2-1b (41.80% vs. 39.69-40.31%) and llama-3.2-3b (66.33% vs. 65.08-65.70%). Cover-ICL also achieves the best performance on gemma-3-4b (77.34%), with VoteK close behind (77.11%), suggesting that diversity-aware methods may be particularly effective for certain architectural families.

Key insights: Grey-box settings favor Cover-ICL on average across model families, achieving 62.34% average performance with particularly strong gains on qwen3-0.6b. In contrast, black-box settings exhibit method-architecture interactions, with hardest selection achieving 68.71% average performance but Cover-ICL proving superior for llama models and gemma-3-4b. Both paradigms show diminishing returns for larger models (8B+ parameters), with black-box settings exhibiting even tighter method convergence.

Conclusion and Limitations

This study systematically analyzes uncertainty estimation across black-box and grey-box paradigms, extending the Cover-ICL framework across seven diverse language models. Cover-ICL's graph-based Maximum Coverage approach excels with precise logit-based uncertainty enabling reliable neighborhood construction, while consistency-based uncertainty provides less precise signals that degrade graph quality. Hardest selection proves more robust to uncertainty noise through simple ranking. Models with 8B+ parameters show minimal sensitivity to method choice (differences under 2%), suggesting well-calibrated representations make method selection less critical for larger models. Importantly, practitioners cannot simply translate grey-box findings to black-box settings due to fundamental differences in uncertainty signal quality, necessitating paradigm-specific optimization strategies.

Model accessibility drives paradigm selection, with black-box access commoditized through major APIs while grey-box requires custom deployments. Black-box consistency estimation requires fixed probes (K=5) regardless of task complexity, while grey-box scales with class numbers. Production systems should prioritize black-box for scalability and cost predictability; grey-box remains valuable for research requiring fine-grained uncertainty control.

Future directions: LLM-as-a-judge applications represent a particularly promising area, as judge performance depends critically on demonstration quality (Li et al. 2024). The proposed framework could identify illustrative examples that improve accuracy and human-LLM alignment, while consistency-based measures could enable adaptive protocols with uncertainty-dependent confidence thresholds. Additionally, investigating architectural factors that determine method-model family interactions could enable more sophisticated uncertainty estimation strategies, as different model families (Gemma, Qwen, Llama) exhibit distinct uncertainty patterns.

Multi-modal extensions present opportunities for vision-language models and code generation tasks, where visual uncertainty compounds textual uncertainty. Task generalization beyond the three evaluated NLP tasks to generation tasks (summarization, dialogue, creative writing) would validate broader framework applicability and reveal task-specific uncertainty patterns.

Limitations

Our analysis focuses on three NLP task types (classification, sentiment, NLI); patterns may differ for generation tasks. Our model selection, while diverse across three families, represents a subset of available architectures. Fixed hyperparameters (B=20, K=5, T=0.7) may not be optimal across all tasks. Additionally, the consistency-based uncertainty metric assumes reliable label extraction from outputs, which may degrade for ambiguous specifications or format violations.

References

Aaron Grattafiori, A. J. e. a., Abhimanyu Dubey. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Dolan, B.; Quirk, C.; and Brockett, C. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 350–356. Geneva, Switzerland: COLING.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. arXiv:1506.02142.

Giampiccolo, D.; Dang, H. T.; Magnini, B.; Dagan, I.; Cabrio, E.; and Dolan, W. B. 2008. The Fourth PASCAL Recognizing Textual Entailment Challenge. In *Text Analysis Conference*.

Hovy, E.; Gerber, L.; Hermjakob, U.; Lin, C.-Y.; and Ravichandran, D. 2001. Toward Semantics-Based Answer Pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Huang, H.-Y.; Wu, Z.; Yang, Y.; Zhang, J.; and Wu, Y. 2025. Unlocking the Power of LLM Uncertainty for Active In-Context Example Selection. arXiv:2408.09172.

Huang, H.-Y.; Yang, Y.; Zhang, Z.; Lee, S.; and Wu, Y. 2024. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326*.

Lewis, D. D.; and Gale, W. A. 1994. A Sequential Algorithm for Training Text Classifiers. arXiv:cmp-lg/9407020.

Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Liu, Y.; Hu, J.; Wan, X.; and Chang, T.-H. 2022. A Simple yet Effective Relation Information Guided Approach for Few-Shot Relation Extraction. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 757–763. Dublin, Ireland: Association for Computational Linguistics.

Ma, H.; Chen, J.; Wang, G.; and Zhang, C. 2025. Estimating llm uncertainty with logits. *arXiv e-prints*, arXiv–2502.

Mavromatis, C.; Srinivasan, B.; Shen, Z.; Zhang, J.; Rangwala, H.; Faloutsos, C.; and Karypis, G. 2023. Which Examples to Annotate for In-Context Learning? Towards Effective and Efficient Selection. arXiv:2310.20046.

Mavromatis, C.; Srinivasan, B.; Shen, Z.; Zhang, J.; Rangwala, H.; Faloutsos, C.; and Karypis, G. 2024. CoverICL: Selective Annotation for In-Context Learning via Active Graph Coverage. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21268–21286. Miami, Florida, USA: Association for Computational Linguistics.

Min, S.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Noisy Channel Language Model Prompting for Few-Shot Text Classification. arXiv:2108.04106.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shorinwa, O.; Mei, Z.; Lidard, J.; Ren, A. Z.; and Majumdar, A. 2025. A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions. arXiv:2412.05563.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Yarowsky, D.; Baldwin, T.; Korhonen, A.; Livescu, K.; and Bethard, S., eds., *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Su, H.; Kasai, J.; Wu, C. H.; Shi, W.; Wang, T.; Xin, J.; Zhang, R.; Ostendorf, M.; Zettlemoyer, L.; Smith, N. A.; and Yu, T. 2022. Selective Annotation Makes Language Models Better Few-Shot Learners. arXiv:2209.01975.

Team, G. 2025a. Gemma 3.

Team, Q. 2025b. Qwen3 Technical Report. arXiv:2505.09388.

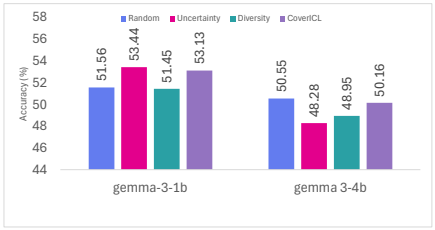
Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. New Orleans, Louisiana: Association for Computational Linguistics.

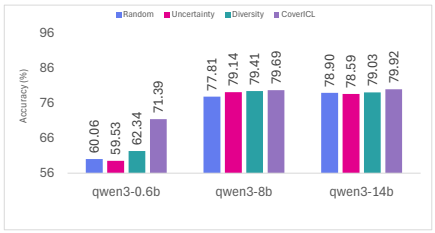
Zhang, X.; Zhao, J.; and LeCun, Y. 2016. Character-level Convolutional Networks for Text Classification. arXiv:1509.01626.

Zhang, Y.; Feng, S.; and Tan, C. 2022. Active Example Selection for In-Context Learning. arXiv:2211.04486.

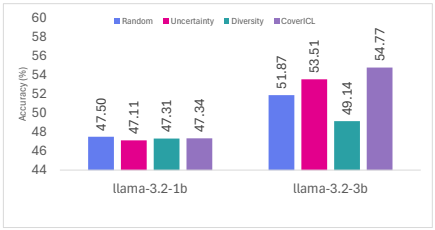
Appendix: additional charts



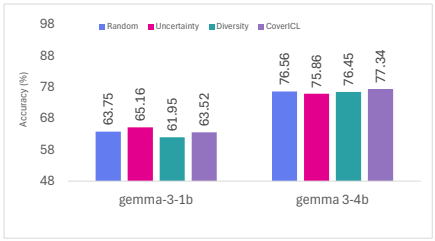
(a) Gemma models (confidence)



(b) Qwen models (confidence)

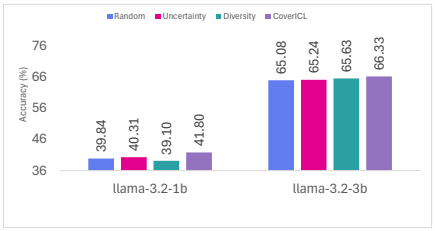


(c) Llama models (confidence)



(d) Gemma models (consistency)

(e) Qwen models (consistency)



(f) Llama models (consistency)

Figure 1: Performance comparison of active learning methods across model families using confidence-based (top row) and consistency-based (bottom row) uncertainty estimation. Results show accuracy (%) for four selection strategies: Random, Uncertainty (hardest), Diversity, and Cover-ICL. Grey-box confidence estimation enables Cover-ICL to achieve optimal performance (62.34% average), while black-box consistency measures favor hardest selection (68.71% average), demonstrating paradigm-dependent optimal strategies for uncertainty-guided example selection in in-context learning.