Optimal Demonstration Selection Techniques in In-context Learning

Yamini Preethi Kamisetty
Jonathan Tong
Satvik Praveen
Vinay Chandra Bandi

Abstract

In-Context Learning (ICL) leverages the inherent knowledge of Large Language Models (LLMs) to enhance performance across a wide range of tasks without requiring task-specific fine-tuning. However, the effectiveness of ICL is significantly influenced by factors such as the choice of demonstration selection methods and the compatibility of the model. This study provides a comprehensive comparative analysis of diverse demonstration selection techniques, examining their impact on ICL performance through multiple lenses: similarity (how closely demonstrations align with test samples), diversity (the variety of demonstrations used), and model dependency (how different models respond to various demonstrations). By exploring these dimensions, this research aims to identify optimal demonstration selection strategies that maximize ICL performance across different tasks and model scales. The study evaluated several state-of-the-art methods like Iterative Demonstration Selection, TopK + ConE, Sequential Example selection, Influence based example selection and Reinforcement Learning-based sample selection, to provide insights into their strengths and limitations. Ultimately, this work seeks to contribute to the development of more efficient demonstration selection in ICL systems.

1 Introduction

Large Language Models have become increasingly larger in the hopes that new emergent abilities will be learned by the models. One of the most recent advances has been In-context Learning (ICL), which utilizes a few output-input examples to better align the model's predictions on a task (Dong et al., 2024).

In-Context Learning has emerged as a promising approach to enhance reasoning in LLMs, particularly in few-shot learning scenarios (An et al., 2023). ICL utilizes LLMs to perform reasoning (Wu et al., 2024) by providing a carefully curated set of demonstrations as context, rather than relying solely on extensive model retraining. This methodology allows LLMs to leverage their inherent capabilities for understanding and processing text, making them particularly suitable for tasks with limited labeled data, without need for additional model training (Kossen et al., 2024).

Despite ICL's advantages, its performance heavily hinges on optimal demonstration selection, balancing relevance and diversity. While highly similar samples may yield good model performance, this approach often fails to generalize well to complex real-world data (Mueller et al., 2024).

Further, selecting an optimal sequence of examples for ICL remains a challenging problem, as the effectiveness of ICL is highly dependent on both the order and contextual relationships between examples (Rubin et al., 2022; Cheng et al., 2023). Traditional selection methods follow a "select then organize" paradigm (Liu et al., 2021; Su et al., 2022), which overlooks the intricate dependencies between examples, potentially leading to inconsistencies between training and inference. To address this, recent studies have explored sequence-aware selection approaches that dynamically construct example sequences by modeling their conditional probability in the given context.(Lu et al., 2022; Wu et al., 2023)

In this paper, we focus on studying existing approaches for demonstration selection focusing on relevance and diversity of these techniques: Reinforcement Learning-based Selection, Iterative Demonstration Selection (IDS), Model-dependent Methods, Sequential Example Selection for In-Context Learning (Se^2), In-context Example Selection with Influences.

2 Background

Unlike traditional machine learning approaches that require fine-tuning on large labeled datasets, ICL leverages prompts containing input-output ex-

emplars to guide model predictions. This capability was first observed in transformer-based architectures, particularly in models scaled to billions of parameters (Brown et al., 2020).

ICL performance is highly sensitive to the selection and ordering of demonstration examples included in the prompt. Different sets or sequences of examples can lead to significant variability in model performance, underscoring that careful example choice and arrangement are crucial.

ICL is rooted in the broader field of meta-learning, where models are trained to learn new tasks efficiently with minimal supervision (Vanschoren, 2018). By conditioning on a set of demonstrations, LLMs infer underlying patterns and apply them to unseen inputs, effectively performing a form of implicit task adaptation. This adaptability is particularly valuable for scenarios where labeled data is scarce or when rapid deployment across diverse tasks is required (Xie et al., 2022).

2.1 Related Work

We reviewed a number of previous works to inform our experiments. Dong's survey (Dong et al., 2024) defined ICL, explored its foundations, and discussed various theoretical perspectives, including Bayesian and Gradient Descent views. (Peng et al., 2024) addressed inconsistencies in demonstration selection strategies across models and varying example counts by evaluating how each example influences model understanding. They proposed a selection strategy that achieved consistent improvements.

Iterative Demonstration Selection (IDS) (Qin et al., 2024) refines in-context demonstration selection using Zero-Shot Chain-of-Thought (Zero-Shot-CoT) reasoning to iteratively optimize similarity and diversity, enhancing model performance. Through iterative refinement, IDS optimizes demonstration selection by capturing both task-specific nuances and reasoning consistency, ultimately leading to improved model performance.

Similarly, sequential-aware strategies like Se² explicitly model selection as an order-sensitive process, ensuring that chosen examples align with the LLM's inference-time behavior (Zhang et al., 2022). Traditional selection methods, such as heuristic scoring (e.g., entropy, uncertainty) or embedding-based similarity searches, often neglect example interactions, leading to inconsistencies. Se² explicitly models the selection process as sequential, allowing each example choice to depend on the current prompt context. By doing so, it addresses the gap left by static "select-then-organize" approaches

If selection prioritizes only similarity, models risk biased representations that fail to generalize to real-world data. Relevance-Diversity Enhanced Selection (RDES) mitigates this by leveraging Q-learning to dynamically balance diversity and relevance, improving classification performance (Wang et al., 2024) employs reinforcement learning, specifically a Q-learning framework, to dynamically select demonstrations that maximize both diversity and relevance to the classification task. The authors propose two variants: RDES/B (base version) and RDES/C (enhanced with Chain-of-Thought reasoning). The research highlights the importance of balancing diversity and relevance in demonstration selection and showcases the potential of integrating reinforcement learning with advanced reasoning techniques to improve LLM performance in various NLP tasks.

(Nguyen and Wong, 2023) proposed an influence-based example selection method to enhance ICL performance by identifying both positive and negative examples. This approach leverages in-context influences to measure the impact of examples on task performance, demonstrating significant improvements over other baselines across various SuperGLUE tasks. The method involves calculating the influence of each example by comparing the average performance of subsets with and without the example, allowing for the selection of the most impactful examples for k-shot prompting. This work not only showcases the efficacy of influence-based selection but also provides insights into phenomena like recency bias in example ordering, underscoring the importance of careful example selection in ICL.

3 Novelty and Challenges

Though there has been much ICL research, many problems and questions remain unanswered. Among the greatest of these is explaining the inconsistency of selection strategies. (Peng et al., 2024) theorize that ICL strategies differ in their performance due to the model they work with, and not by how similar or varied the demonstrations they provide to the model are. Similarly, (Qin et al., 2024) found that both similarity and variability are equally important in sample selection. As such,

research needs to be done to pinpoint the optimal settings for each model and dataset in order to improve ICL performance.

Further, the key novelty of Se² lies in reframing example selection as an order-sensitive process, which helps mitigate the inconsistencies found in prior selection strategies. Traditional ICL demonstration selection did not account for how examples interact, often leading to a disconnect between the selection procedure and the model's actual inference behavior.

Another problem is that of efficiency and scalability. (Shu and Du, 2024) found that higher accuracy models ran far slower than their less accurate counterparts, thus posing problems for live applications.

4 Research Goals

We aim to map ICL performance along a number of different strategies, models, and datasets. This will further the field by guiding future researchers towards the optimal parameters for each, thus leading to better ICL techniques tailored to specific tasks. As such, our research will:

- Examine a number of different ICL selection strategies and implement them.
- Test how LLM performance varies based on task and selection strategy.
- Map ICL performance on different datasets.

5 Evaluation Metrics

While we observed a variety of methods to evaluate selection performance (F1, Rouge-L, RDES, etc.), we decide to use accuracy to evaluate all our methods due to its simplicity.

6 Methodology

6.1 Datasets

Our comparative study includes diverse datasets, particularly those where sequential dependencies in example selection are crucial. Previous works covered a large selection of datasets which informed the data for our own experiments:
(Peng et al., 2024) utilize seven datasets spanning sentiment analysis (SST-2, SST-5, CR, Subj) and natural language inference (MNLI, QNLI). (Qin et al., 2024) incorporate five datasets covering mathematical reasoning (GSM8K, MATH), Commonsense reasoning, logical reasoning, and question answering . RDES evaluates ten baseline methods, including prompt engineering and demonstration selection techniques, across four benchmark datasets (Wang et al., 2024).

(Nguyen and Wong, 2023) evaluates the influence-based example selection method using datasets from the SuperGLUE benchmark and additional tasks like PIQA, Hellaswag, and OpenBookQA. These datasets were chosen for their diversity, covering a wide range of natural language tasks including textual entailment, question-answering, and multi-choice tasks. They vary in complexity and task type, allowing for a comprehensive assessment of the influence-based method's performance on both binary classification and multi-choice tasks. Additionally, these datasets are well-documented and accessible, facilitating reproducibility and comparison with other methods.

(Wang et al., 2024) evaluated the RDES framework by conducting extensive experiments across four benchmark datasets: BANKING77, CLINC150, HWU64, and LIU54 using the classification accuracy metric. The performance was compared against ten baselines, including prompt engineering methods and demonstration selection metods.

Based off these works and their findings, we chose three datasets across three tasks: SST-5 for sentiment analysis (Socher et al., 2013), CommonsenseQA for topic classification (Talmor et al., 2019), and the AGNews dataset for news topic classification (Zhang et al., 2016). These datasets were chosen for their variety, thus testing the generalization ability of the models and methods selected.

6.2 Models

Since example selection strategies can have varying impacts depending on model scale, we include different sizes to examine how selection strategies generalize across architectures. Our model selection was guided by open-source availability and efficiency. For this reason we choose Llama3.1-8B, GPT-2, and Deepseek-llm-7B-chat. Llama and GPT are popular, open-source options used in most publications. We follow the literature in comparing these two, as well as adding a new model, Deepseek, to our comparison. Deepseek and Llama3.2 were chosen for their comparable sizes and because previous literature had not re-

viewed these in depth due to their relatively recent releases. Unfortunately, GPT-2 is much smaller than modern models, but is the only GPT model publically and freely available. As such, we hope that our findings will identify general trends within each of these model families by exhaustively comparing them.

6.3 Approaches

We will experiment with the following ICL selection strategies:

- The TopK + ConE approach (Peng et al., 2024) is a data- and model-dependent strategy designed to enhance in-context learning (ICL) performance by selecting demonstrations that maximize the model's understanding of test inputs. Initially, it employs the TopK method to narrow down demonstration candidates based on their similarity to the test input. These candidates are then ranked using conditional entropy (ConE), which measures how much each demonstration reduces the uncertainty of the model about the test input. By minimizing conditional entropy, the method ensures that selected demonstrations effectively enhance the model's understanding. This approach has been shown to yield consistent improvements across various model scales and tasks, providing a unified explanation for the effectiveness of previous demonstration selection methods. The strategy's effectiveness is further validated through experiments that verify its performance and explore correlations with other selection methods.

- The Iterative Demonstration Selection (IDS) approach, proposed by (Qin et al., 2024), aims to enhance in-context learning (ICL) for large language models by iteratively selecting optimal demonstration examples. IDS leverages zero-shot chain-of-thought reasoning (Zero-shot-CoT) to generate a reasoning path for the test sample, which is then used to select semantically similar training examples as demonstrations. These demonstrations are prepended to the test sample for inference, and the process is repeated with the new reasoning path until a final result is obtained through majority voting. By balancing diversity and similarity, IDS outperforms existing ICL demonstration selection methods across various tasks, including reasoning, question answering, and topic classification

- To deal with selecting diverse examples, RDES utilizes Q-learning framework to optimize the selection of diverse reference demonstrations for text classification tasks using Large Language Models (LLMs). The framework represents the state as a tuple containing the input text, selected demonstrations, predicted label, and diversity score. Actions involve selecting demonstrations from a knowledge base. The Q-learning algorithm iteratively updates Q-values based on rewards, which are determined by classification accuracy. RDES uses cosine similarity with TF-IDF vectorization to select initially relevant demonstrations, then calculates a diversity score based on unique labels. If diversity is below a threshold, it incorporates less similar demonstrations to increase diversity. The framework uses an -greedy strategy for action selection, balancing exploration and exploitation. RDES can be integrated with Chain-of-Thought (CoT) reasoning, where the LLM provides explanations for its classifications, further enhancing performance.

- To examine the effect of selection order in in-context learning (ICL), we leverage **Se^2** (Sequential Example Selection), which treats example selection as a sequential decision-making process rather than a static optimization problem. Unlike "select-then-organize" paradigms that ignore inter-example dependencies and disrupt consistency between training and inference, **Se^2** dynamically constructs example sequences by conditioning each selection on the evolving context and prior examples. At each step, **Se^2** uses feedback from a frozen large language model (LLM) to score candidate examples, modeling contextual relevance and coherence. This iterative approach captures semantic relationships among examples while enabling prompt construction aligned with K-shot inference settings. During training, a bi-encoder architecture is optimized using InfoNCE loss to discriminate informative examples under varying contexts. At inference, beam search expands the candidate space and improves sequence quality and diversity. Empirical results across 23 NLP

tasks show that **Se^2** significantly outperforms competitive baselines such as BM25, SBERT, UPRISE, and AES, achieving a 42% relative gain over random selection. Additionally, **Se^2** enhances stability, generalization across tasks, and transferability across LLM scales. These findings underscore the importance of order-aware, feedback-driven example selection for robust and adaptive ICL.

- Another approach is to utilize the in-context influences to enhance in-context learning (ICL) performance in large language models (LLMs). (Nguyen and Wong, 2023) propose an influence-based example selection method. This method involves randomly selecting subsets of examples from a training set and evaluating the model's performance on a validation set for each subset. The influence of each example is then calculated by comparing the average performance of subsets that include the example versus those that do not. Examples are ranked based on their influence scores, and the top influential ones are selected for k-shot prompting to maximize performance. This approach outperforms many other baselines in both positive and negative example selection across nine SuperGLUE tasks, revealing a significant performance gap between the most positive and negative examples. The framework is computationally efficient, requiring only forward passes through the model, and can be applied to various tasks and models to study phenomena like recency bias in example ordering.

6.4 Experimental Settings

All experiments were run on either Google Colab or Texas A&M HPRC resources. A100 GPUs were utilized for almost every experiment, with T4 GPUs being used on a handful of experiments.

As TopK + CoNE required no training, we followed in the creator's footsteps and sampled three batches of 100 prompts from the train sets, taking the average as our final score.

For evaluating the IDS approach, we utilized 5,000 training samples and 1,000 test samples. For the Influence-based selection method, we used the same training and test samples as IDS, with an additional 1,000 validation examples.

For RDES approach we used the entire dataset because of the training involved.

All the code used to train and test the models can be found in the following github repository: GitHub Link.

7 Results, Findings and Insights

7.1 TopK + CoNE

Generally, bigger models performed better for TopK + CoNE, although LLAMA struggled on AGnews. Gemma matched or exceeded LLAMA on the first two tasks, but struggled on CommonsenseQA. GPT-2, being far smaller than the other two models, performed the worst, but had surprisingly high performance on AGnews for a model of its size.

Model	AGnews	SST5	CommonsenseQA
LLAMA-3.2-3b	0.786	0.567	0.640
Gemma-2-2b	0.790	0.530	0.496
GPT-2 (100M)	0.603	0.370	0.260

Table 1: TopK + CoNE comparison across AGnews, SST5, and CommonsenseQA datasets. Metric is accuracy.

7.2 Iterative Demonstration Selection (IDS)

This study evaluates the Iterative Demonstration Selection (IDS) approach as described by (Qin et al., 2024). The results in fig.1 show that GPT-4o-mini consistently outperforms other models across tasks, particularly in commonsense reasoning with an accuracy of 82.9% on CommonsenseQA. This highlights its strong capability in complex reasoning tasks, likely due to its architecture and training strategy. Meanwhile, Gemma-2b-it outperforms LLaMA-3b-instruct on certain tasks despite having fewer parameters, suggesting that instruction tuning and dataset alignment are crucial.

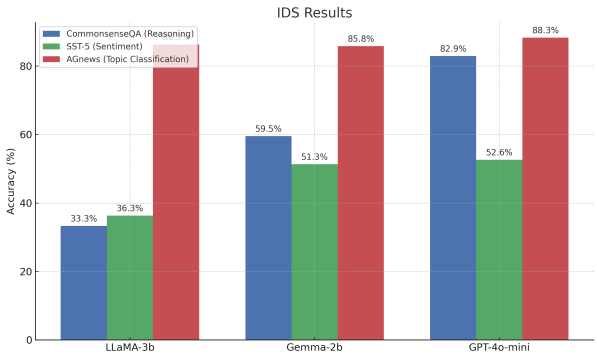

Figure 1: Performance of IDS on different models and dataset combinations.

Across all models, the AG News dataset shows high performance (85–88%), indicating that topic classification is relatively easier and relies on

surface-level understanding. In contrast, fine-grained sentiment classification (SST-5) remains challenging. The results emphasize that larger models are not always superior; instead, factors like instruction tuning, data selection, and model-task alignment are critical for success, especially in tasks requiring deeper reasoning or nuanced understanding.

7.3 Reinforcement Demonstration Selection (RDES)

This study assesses the effectiveness of using reinforcement learning (RL) for demonstration selection in in-context learning, as proposed by (Wang et al., 2024). The evaluation involves comparing the performance of three models-Gemma 2B, Llama 3.2 3B, and GPT-2-across four datasets: binary sentiment classification (SST2), multi-class sentiment analysis (SST5), news categorization (AG News), and commonsense reasoning (CSQA) as shown in Figure 2. The results show significant performance variations across different model-dataset combinations. For instance, Gemma 2B excels in binary sentiment classification with a high accuracy of 91%, outperforming both GPT-2 and Llama 3.2 3B.

The performance analysis highlights that each model has distinct strengths in specific tasks. Gemma 2B performs well in binary classification and commonsense reasoning, while Llama 3.2 3B excels in multi-class categorization tasks. GPT-2 consistently underperforms across all datasets, particularly struggling with multi-class sentiment analysis. These findings suggest that the effectiveness of RL-based demonstration selection is influenced by model size and task complexity. Larger models seem to benefit more from this approach, indicating that they can better utilize informative demonstrations. The study concludes that demonstration selection strategies should be tailored to both the specific task characteristics and the target model architecture to maximize performance benefits.

7.4 Se²

We first evaluated our Sequential Example Selection (Se²) method on the same three benchmarks and three model architectures, namely GPT-Neo-1.3B, GEMMA-2B, and GPT-2-medium. Table 2 and Figure 3 report the average accuracy over three random 200-example splits.

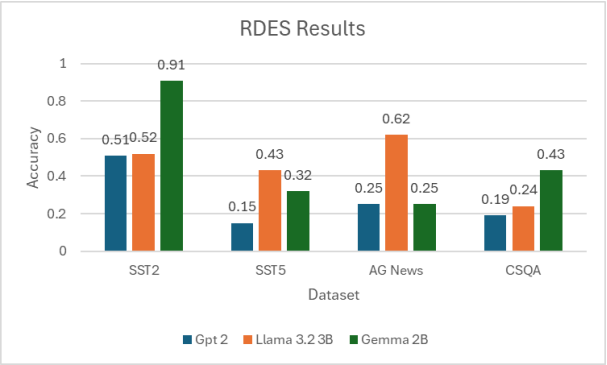

Figure 2: Performance of RDES across different models and datasets

Model	CommonsenseQA	AG News	SST-5
GPT-Neo-1.3B	0.223	0.698	0.394
GEMMA-2B	0.211	0.825	0.258
GPT-2-medium	0.196	0.581	0.263

Table 2: Se² accuracy (averaged over 3 splits) on AG-news, SST-5, and CommonsenseQA.

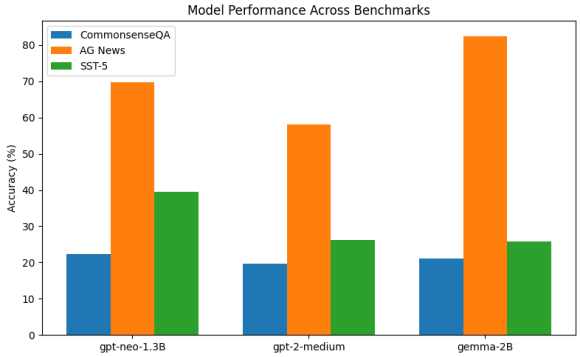

Figure 3: Se² performance across models and datasets.

Shot & Beam	1	2	3
1-shot	0.108	0.098	0.091
2-shot	0.108	0.140	0.179
3-shot	0.079	0.060	0.080

Table 3: Llama-3.2-3B few-shot accuracy on CommonsenseQA (averaged over 3 splits).

The above results were obtained by running the models under Se² approach with 3-shot setting under a beam size 3. Additionally, we worked with Llama-3.2 3B model by varying the number of shots and beam size hyperparameter to obtain the results as shown in Table 3.

Tables 2 and Figure 3 report aggregate Se² accuracy (averaged over three independent 100-example splits) on three benchmarks. On AG News, all three backbones achieve their highest scores—GPT-Neo-1.3B: 69.8 %, Gemma-2B: 82.5 %, GPT-2-medium: 58.1 %—reflecting that topic classification is the easiest of our three tasks un-

der Se². Sentiment (SST-5) comes next (25.8–39.4 %), while CommonsenseQA remains the hardest (< 22.3 %). The relative ordering of models (Gemma > GPT-Neo > GPT-2) is consistent across tasks, which suggests Se²'s sequential formulation scales predictably with model capacity.

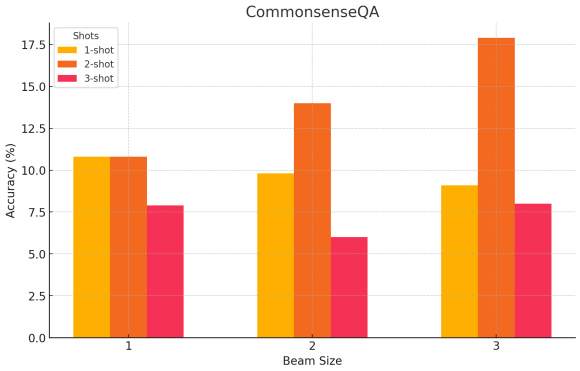

Figure 4: Few-shot accuracy of LLaMA-3.2-3B under Se² on CommonsenseQA, for 1–3 shots and beam sizes 1–3.

Shot & Beam	1	2	3
1-shot	0.748	0.729	0.724
2-shot	0.731	0.748	0.759
3-shot	0.727	0.765	0.786

Table 4: Llama-3.2-3B few-shot accuracy on AG News (averaged over 3 splits).

Figures 4–6 (and Tables 3–5) then fix the model to LLaMA-3.2-3B and sweep shot-count (1–3) and beam-size (1–3). On CommonsenseQA (Table 3, Figure 4), two-shot prompts deliver the best accuracy—peaking at 18.0 % for beam = 3—whereas one- and three-shot stay below 11 %. For AG News (Table 4, Figure 5), three-shot accuracy climbs steadily from 72.4 % (beam 1) to 78.6 % (beam 3), and two-shot likewise rises from 72.9 % to 75.9 %; one-shot shows a slight downward drift with larger beams. On SST-5 (Table 5, Figure 6), two-shot peaks at 38.7 % (beam 2), three-shot at 39.5 % (beam 1), and one-shot varies between 34.0–35.8 %. These results confirm that (a) Se²'s beam-search expansion and (b) the number of in-context examples both have a task-dependent sweet spot, so hyperparameter tuning is key for each new dataset. Overall, these findings underscore the importance of carefully optimizing model parameters to achieve optimal performance across diverse tasks.

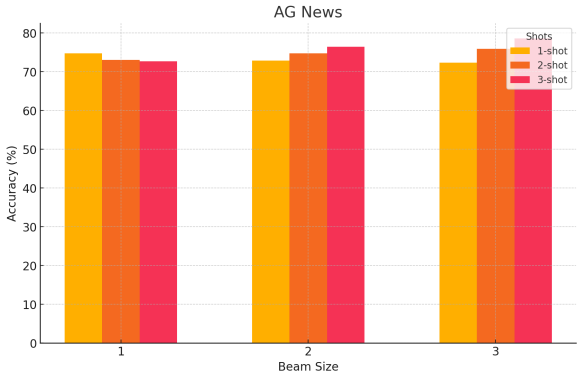

Figure 5: Few-shot accuracy of LLaMA-3.2-3B under Se² on AG News, for 1–3 shots and beam sizes 1–3.

Shot & Beam	1	2	3
1-shot	0.339	0.363	0.358
2-shot	0.361	0.387	0.382
3-shot	0.394	0.352	0.396

Table 5: Llama-3.2-3B few-shot accuracy on SST-5 (averaged over 3 splits).

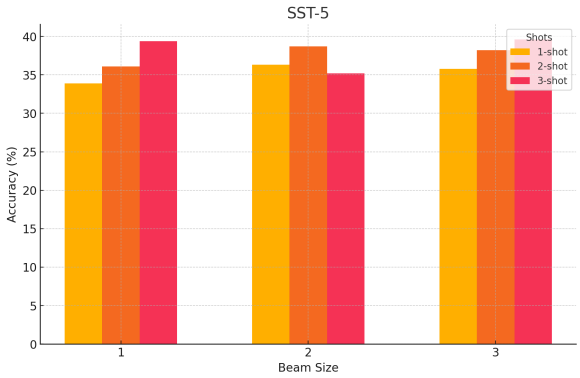

Figure 6: Few-shot accuracy of LLaMA-3.2-3B under Se² on SST-5, for 1–3 shots and beam sizes 1–3.

7.5 Influence based Selection

The experimental results for influence-based demonstration selection across three language models and three diverse datasets reveal several notable patterns in performance effectiveness. Figure 7 illustrates the accuracy results achieved by applying influence-based demonstration selection across three models (LLAMA-3.2 3b-instruct, gemma-2-2b-it, and GPT-2) on three standard NLP datasets (AGnews, SST5, and CommonsenseQA).

The results indicate that gemma-2-2b-it consistently achieves superior performance with influence-based demonstration selection on classification tasks (AGnews: 52.4%, SST5: 51.1%). This suggests that Gemma's architecture may be particularly well-suited for leveraging influential

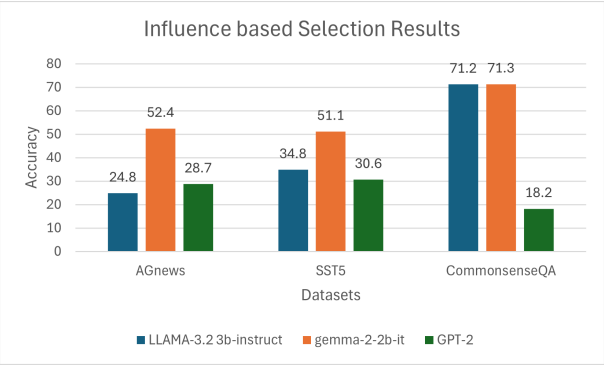

Figure 7: Performance of Influence based selection technique

examples in text classification scenarios.

For the CommonsenseQA dataset, both LLAMA-3.2 and gemma-2-2b-it achieve comparable high performance (71%), substantially outperforming GPT-2 (18.2%). This indicates that influence-based selection is especially effective for question-answering tasks when paired with more recent model architectures.

GPT-2 consistently underperforms compared to the other models across all datasets, suggesting that influence-based selection may be less effective for older or smaller language models with more limited contextual understanding capabilities

8 Discussion

Our comparative study evaluated five demonstration selection techniques across three language models (LLAMA 3.2 3B instruct, GEMMA 2-2B-it, and GPT2) on three diverse datasets (AGNews, SST-5, and CommonsenseQA). Figure 8 presents the accuracy results for each approach-model-dataset combination, revealing significant performance variations and interesting patterns.

Iterative Demonstration Selection (IDS) emerges as the most consistently effective approach, achieving the highest overall accuracy across most configurations. Particularly noteworthy is IDS's exceptional performance on GPT-AGNews (89%), LLAMA-AGNews (85%), and GPT-CQA (83%). This suggests that iterative refinement of demonstrations based on model feedback provides robust performance across diverse architectures and tasks.

Sequential Example Selection (Se²) demonstrates strong performance specifically with LLAMA models, achieving 82% accuracy on SST-5 and 78% on AGNews. However, its effectiveness drops dramatically with GPT models (17-21%), indicating strong model-dependency in its selection criteria.

TopK+ConE maintains consistent performance across most model-dataset combinations, particularly excelling on LLAMA-SST-5 (79%) and LLAMA-AGNews (78%). This approach demonstrates robust cross-model generalization without extreme performance fluctuations.

In-context Example Selection with Influences (ICINF) exhibits the most variable performance pattern, achieving remarkable results on GPT-SST-5 (70%) and GPT-AGNews (70%) while performing poorly on LLAMA-AGNews (25%). This stark contrast suggests that influence-based selection interacts uniquely with different model architectures.

Reinforcement Learning-based Selection (RDES) generally performs moderately, with its peak performance on LLAMA-AGNews (62%). The approach shows more consistent behavior on LLAMA models but struggles significantly with GPT models.

The experimental results show distinct interactions between models and different selection techniques. LLAMA models work well with Se² and TK+CE approaches, indicating they effectively use sequentially selected demonstrations. This is evident in the significant performance difference on AGNews, where IDS outperforms other techniques by a large margin.

In contrast, GEMMA models are more robust and show less variation across different selection methods, with IDS generally performing best. GPT models, however, are highly sensitive to selection methods, with significant performance differences between techniques. IDS and ICINF outperform other approaches on certain tasks, while RDES and Se² consistently underperforms, suggesting GPT's internal representations align better with specific selection criteria.

9 Future Work

Building upon our comparative analysis, several promising research directions emerge for enhancing demonstration selection in in-context learning. First, we aim to develop an adaptive meta-selection framework that automatically identifies the optimal selection technique based on both model architecture and task characteristics, potentially eliminating the need for manual technique selection. Second, investigating the architectural factors within language models that influence selection technique effectiveness would provide valuable insights for designing model-specific selection strate-

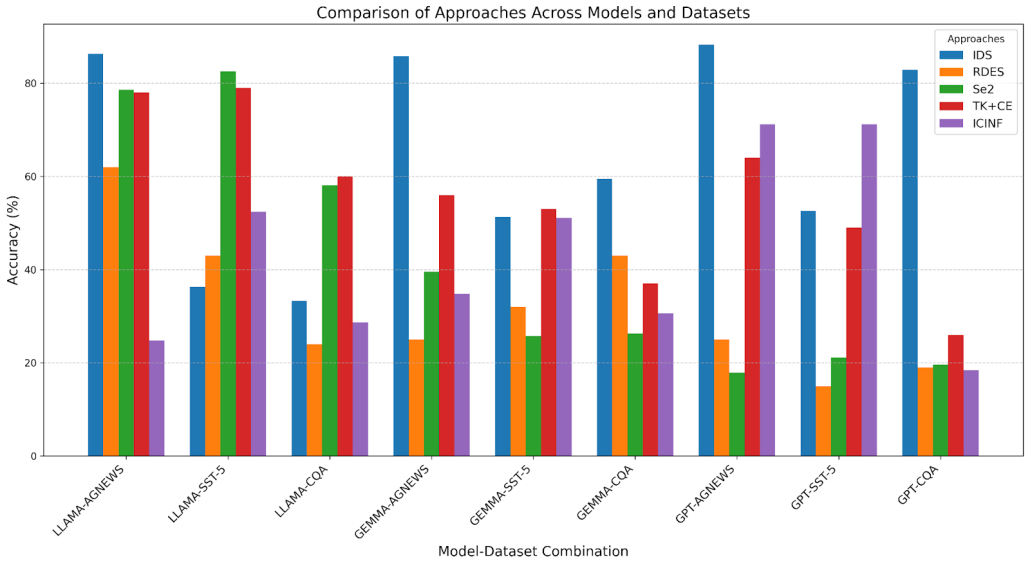

Figure 8: Comparison of all methods by model-task.

gies. Third, exploring computationally efficient approximations of high-performing techniques like IDS could make these approaches more practical for resource-constrained environments. Fourth, we plan to extend our analysis to multimodal tasks and models, examining whether the observed patterns generalize beyond text-only scenarios. Fifth, developing hybrid approaches that combine the strengths of multiple techniques (e.g., the consistency of TK+CE with the peak performance of IDS) represents a promising direction for robust cross-model performance. Finally, investigating how demonstration selection techniques scale with increasingly larger language models would provide insights into future-proofing selection strategies as model capabilities continue to advance. These directions collectively aim to establish more principled, theoretically-grounded approaches to demonstration selection that can adapt to the evolving landscape of language models and tasks.

10 Conclusion

Our comparative analysis of demonstration selection techniques reveals several critical insights for in-context learning optimization. The experimental results across three models and three datasets demonstrate that selection strategy effectiveness is highly dependent on the specific model-task combination, with IDS consistently delivering superior performance (averaging 70-90% accuracy) across diverse scenarios while maintaining exceptional robustness. The dramatic performance disparities observed when applying identical techniques to different models-exemplified by Se²'s 82% accuracy on LLAMA-SST-5 versus merely 21% on GPT-SST-5-underscore the critical importance of aligning selection strategies with specific model architectures. Additionally, the substantial variation in method effectiveness across different tasks highlights how optimal technique selection must simultaneously consider both model architecture and task characteristics. While approaches like TK+CE demonstrate valuable cross-model consistency (maintaining 50-70% accuracy across most configurations), potentially benefiting deployment scenarios requiring model flexibility, they generally underperform compared to specialized techniques like IDS. These findings reveal an important trade-off between computational requirements and performance gains, suggesting that investment in sophisticated selection methods typically yields substantial accuracy improvements despite higher computational costs. Collectively, these insights illuminate the complex interplay between demonstration selection techniques, model architectures, and task characteristics, emphasizing that demonstration selection for in-context learning requires thoughtful consideration rather than a universal approach.

References

Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen, and Jian-Guang Lou. 2023. Skill-based few-shot selection for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13472–13492, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337, Singapore. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning.

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. In-context learning learns label relationships but is not conventional learning.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3?

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Aaron Mueller, Albert Webson, Jackson Petty, and Tal Linzen. 2024. In-context learning generalizes, but not always robustly: The case of syntax.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2024. In-context learning with iterative demonstration selection.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning.

Dong Shu and Mengnan Du. 2024. Comparative analysis of demonstration selection algorithms for llm in-context learning.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Selective annotation makes language models better few-shot learners.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Joaquin Vanschoren. 2018. Meta-learning: A survey.

Xubin Wang, Jianfei Wu, Yichen Yuan, Mingzhe Li, Deyu Cai, and Weijia Jia. 2024. Demonstration selection for in-context learning via reinforcement learning.

Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, and Jianhua Tao. 2024. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.