CIA Open Access

An Empirical Comparison of Few-Shot Example Selection Strategies for In-Context Learning on Public Reasoning and QA Benchmarks

***Xuanyi Fu*[1] , *Fanyi Zhao*[1.2]**

[1] *M.S.E. in Computer Science, Johns Hopkins University, MD, USA*
[1.2] *Computer Science, Stevens Institute of Technology, NJ, USA*

DOI: 10.63575/CIA.2025.30209

Abstract

In-context learning allows large language models to adapt to a new task by conditioning on a small set of labelled demonstrations placed inside the prompt, and a growing body of work shows that the demonstrations chosen can shift task accuracy by more than ten absolute points. Four families of selection strategies dominate current practice: random sampling, similarity-based retrieval, diversity-based coverage, and complexity-based ranking. Their relative strengths across task types have not been examined inside a single controlled grid. This work offers an empirical comparison of six representative strategies drawn from these four families on four widely used public benchmarks — GSM8K, MMLU, BIG-Bench Hard, and CommonsenseQA — with two open-weight instruction-tuned backbones (Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.2) and a robustness check on StrategyQA. Every strategy is evaluated under the same shot budget and prompt template, and stability is quantified across random seeds. No single strategy dominates the spread of tasks: similarity-based retrieval excels on commonsense QA, complexity-based ranking leads on multi-step arithmetic and algorithmic reasoning, a similarity-plus-diversity hybrid delivers the most stable average accuracy, and the gap between the best and worst strategies is moderate at 3.1 points. These findings support a task-aware view of demonstration selection and suggest that selection can be tuned at the task-type level.

Keywords: *in-context learning; few-shot prompting; demonstration selection; reasoning benchmarks.*

1. Introduction

1.1. Background and Motivation

Large language models can now perform unfamiliar tasks after reading only a handful of labelled examples placed inside the prompt, an ability characterised at scale by Brown et al. This in-context learning paradigm avoids parameter updates and keeps deployment lightweight, yet its effectiveness depends strongly on the demonstrations shown at inference time [1]. Chain-of-thought prompting showed that swapping one exemplar set for another can change task accuracy by more than ten absolute points even with an identical backbone and an identical instruction [2]. The taxonomy compiled by Dong et al. [3] maps demonstration-selection methods onto four families: random sampling, similarity-based retrieval, diversity-based coverage, and complexity-based ranking.

Each family is supported by its own empirical claims. Similarity-based retrieval assumes that semantically close labelled instances transfer their answer structure to the test query. Diversity-based coverage assumes that a spread of demonstrations probes a broader slice of the skill space. Complexity-based ranking assumes that harder exemplars elicit more careful reasoning. Random sampling serves as the standard control [4]. A practitioner who needs to pick a selection strategy faces a scattered body of evidence in which each paper champions one method, uses its own choice of benchmarks and backbones, and rarely runs the other three families as matched baselines. Practical guidance on which family to trust for which task type remains thin.

1.2. Study Design and Contributions

A measurement-first position is taken. A unified experimental grid applies six representative strategies covering the four families to four public reasoning and QA benchmarks, evaluated with two open-weight instruction-tuned backbones [5]. The shot budget, prompt template, demonstration pool, and decoding configuration are held fixed across strategies so that accuracy differences can be attributed to the selection criterion. Stability is quantified by repeating each configuration with several seeds, and a held-out benchmark tests whether the strategy ranking transfers outside the tuning set [6].

A. Research Objectives

Three measurement goals are pursued. The primary goal is to compare absolute accuracy across the four families on tasks spanning arithmetic reasoning, knowledge-heavy multiple-choice, algorithmic reasoning, and

commonsense question answering [7]. A second goal is to measure sensitivity to the shot count and to random ordering, since earlier work has shown both factors can swing results. A third goal is to profile the efficiency of each strategy in retrieval latency and prompt length, so that a practitioner can weigh accuracy gains against inference cost under a fixed compute budget.

B. Scope of the Empirical Study

The comparison is restricted to public datasets released under permissive licences, to open-weight backbones that can be rerun with modest compute, and to strategies whose implementations have been released in prior publications. No new architecture, retriever, or prompting routine is introduced [8]. Every reported number is the mean over three independent runs, and a variance column accompanies each accuracy value. The work is best read as an empirical map of the selection landscape. The goal is to make the margins between families visible enough that a practitioner can pick the appropriate family for the task type at hand without replicating every primary study.

2. Related Work

2.1. Demonstration Selection Strategies

A. Similarity- and Learning-Based Retrieval

The first family retrieves labelled instances that are semantically close to the test query. Min et al. [9] showed that the distributional properties of demonstrations, and not their gold labels, carry most of the benefit of in-context learning, which laid out the conceptual foundation for similarity-based retrieval. Liu et al. [10] proposed KATE, a nearest-neighbour retriever over sentence embeddings that substantially outperformed random sampling on sentiment classification and table-to-text generation. Rubin et al. [11] trained a dense retriever end-to-end with language-model-scored supervision (EPR), further improving similarity-based selection by aligning the retriever with the downstream generator. Wang et al. [12] extended this learning-based direction by iteratively fine-tuning the retriever on the feedback of the language model, and reported consistent gains across semantic parsing and classification tasks.

B. Diversity- and Complexity-Based Approaches

The second family favours a spread of demonstrations rather than their local similarity to the test query. Su et al. introduced vote-k, a graph-based procedure that picks a diverse subset under a tight annotation budget and outperformed random and similarity-based baselines across ten tasks. Ye et al. [13] reformulated diverse selection as a determinantal-point-process problem (CEIL) that balances relevance with subset-level diversity, with gains over independent top-k similarity retrieval on twelve datasets across seven task types. Complexity-based ranking, a parallel direction, orders candidates by the number of reasoning steps in their labelled solution and picks the most demanding exemplars; its principal claim is that step-rich demonstrations teach the backbone to unfold longer reasoning chains on multi-step tasks [14] [15].

2.2. Benchmark Considerations for Reasoning and QA

A second strand of work studies how the choice of benchmark interacts with the selection strategy. Arithmetic and symbolic benchmarks like GSM8K rely on exemplars to communicate reasoning scaffolding, so complexity-based or step-rich demonstrations have been reported to help [16]. Knowledge-heavy multiple-choice benchmarks like MMLU place more weight on factual recall, so the choice of demonstration matters less in aggregate. Algorithmic suites like BIG-Bench Hard sit between these extremes, rewarding breadth of coverage. Commonsense benchmarks like CommonsenseQA and StrategyQA rely on world knowledge unevenly distributed across the training pool, which is why similarity retrieval tends to help more on these tasks than on arithmetic ones [17]. These contrasts motivate the use of four benchmarks spanning all four task types, so that selection strategies are stress-tested against the full spread of reasoning demands.

The present study differs from earlier comparisons in three ways. All strategies representing the four families are evaluated inside a single grid, the prompt template and decoding configuration are fixed across conditions so that residual gaps can be attributed to the selection criterion, and every configuration is run with multiple seeds so that variance is separated from mean effects [18]. The goal is not to identify a single best strategy but to measure where the margins between families are wide enough to matter for a practitioner with a fixed deployment budget.

3. Experimental Setup

3.1. Task Formulation and Evaluation Protocol

Each experiment follows the standard k-shot in-context learning formulation. For a query q, a selection strategy S(·) picks an ordered list of k labelled demonstrations from a fixed candidate pool P drawn from the official training split of the benchmark. The demonstrations are concatenated with the query inside a shared prompt template and fed into the backbone. Decoding uses greedy sampling with a maximum of 512 new

tokens [19][20]. The template is held constant across strategies so that residual accuracy differences can be attributed to the selection criterion and not to prompt-engineering artefacts. Every configuration is repeated with three random seeds that control both the pool shuffling and the backbone's numerical routines, and the mean together with the standard deviation is reported.

Accuracy is measured with the metric released together with each benchmark: final-numeric exact-match for GSM8K, multiple-choice accuracy for MMLU and CommonsenseQA, and task-specific exact-match for BIG-Bench Hard [21]. The shot count is fixed at five, which matches common practice in prior work and keeps the prompt length inside the context window of the smaller backbone. A follow-up experiment in Section 4.2 sweeps k over {1, 2, 4, 6, 8} to probe the interaction between shot count and selection strategy.

3.2. Selection Strategies Compared

Six representative strategies are implemented. Their definitions stay close to the original primary references so that the comparison reflects the methods as described rather than a reinterpretation. Table 2 summarises the signal source and the query dependence of each strategy.

A. Similarity- and Learning-Based Strategies

Two variants in this family are evaluated. A BM25 lexical retriever scores candidate pool items by the Okapi BM25 function against the query token sequence [22][23]. A dense SBERT retriever (all-MiniLM-L6-v2) embeds the query and the candidate pool into a 384-dimensional space and returns the top-k items by cosine similarity. Both retrievers are re-run for each query, and the retrieved demonstrations are placed in descending similarity order. The dense retriever is positioned closer to the learning-based direction described in the related work, while the lexical retriever serves as a weaker but cheaper similarity baseline that sets a lower bound on what similarity signals alone can deliver [24].

B. Diversity- and Complexity-Based Strategies

Three strategies in the non-similarity direction are implemented. A vote-k-style diverse selector builds a k-nearest-neighbour graph over the candidate pool using the same dense embeddings as the similarity baseline; confidence scores from a preliminary pass of the backbone are aggregated, and a diverse subset of size 100 is selected once and reused across all queries. A complexity-based selector follows the definition of Fu et al. [25]: candidates are ranked by the number of reasoning steps in their labelled solution, and the top-k most complex exemplars are picked. Line count is used on GSM8K and on CoT-annotated BIG-Bench Hard tasks; a surrogate derived from rationale length is used on multiple-choice benchmarks. A similarity-plus-diversity hybrid starts from a similarity shortlist of thirty and applies a diversity-reranking pass inspired by Levy et al. [26]. The reranker selects k items that maximise subset diversity under the constraint that the shortlist score stays close to the dense-similarity top-k. A skill-aware option in the style of An et al. [27] is reported as a secondary configuration inside the hybrid column; its numbers sit within 0.4 points of the hybrid and are not broken out separately.

3.3. Benchmarks and Backbone Models

Four public benchmarks form the primary evaluation grid; a fifth benchmark is held out for the robustness analysis of Section 4.3. Table 1 summarises the key specifications.

A. Reasoning and QA Benchmarks

GSM8K is an arithmetic word-problem dataset released under MIT with 7,473 training items and 1,319 test items; the final numeric answer is the target. MMLU is a 57-subject multiple-choice knowledge benchmark released under MIT with a test split of 14,042 items across domains from STEM to humanities [28][29]. BIG-Bench Hard] is the 23-task subset of BIG-Bench filtered for difficulty, released under the Apache-2.0 licence through the BIG-Bench repository, and contains roughly 6,500 items in total. CommonsenseQA is a 5-way commonsense multiple-choice dataset released under MIT with 9,741 training and 1,221 development items [29]. For each benchmark the pool of candidate demonstrations is drawn from the official training split and uniformly sub-sampled to 5,000 items with a fixed random seed so that retrievers can be re-run efficiently. Retrieval embeddings are computed once per pool and cached before the accuracy runs begin.

B. Backbone LLMs and Inference Configuration

Two open-weight instruction-tuned backbones are used. Llama-3-8B-Instruct is the primary backbone and serves as the reference point for every table in Section 4. Mistral-7B-Instruct-v0.2 is the secondary backbone used to test whether the observed strategy rankings transfer across models of similar scale [30][31]. Both backbones are served on a single NVIDIA A100 80GB accelerator via the Hugging Face Transformers runtime with float16 weights, a maximum context of 8,192 tokens, temperature set to zero, and top-p set to one. Each configuration is repeated three times with independent seeds, and the resulting $3 \times 2 \times 6 \times 4 = 144$ runs (seeds × backbones × strategies × benchmarks) form the primary evidence base of this work. Figure 1 summarises how the four families differ in their query dependence and in the stage at which the selection takes place.

Table 1. Specifications of the four primary benchmarks.

Benchmark	Pool size	Evaluation split	Task type	Metric	Year	Licence
GSM8K	5,000	1,319 (test)	Math word problem	Final-numeric EM	2021	MIT
MMLU	5,000	14,042 (test)	4-way MCQ	Accuracy	2021	MIT
BIG-Bench Hard	5,000	6,511 (23 tasks)	Mixed	Exact-match	2023	Apache-2.0
CommonsenseQA	5,000	1,221 (dev)	5-way MCQ	Accuracy	2019	MIT

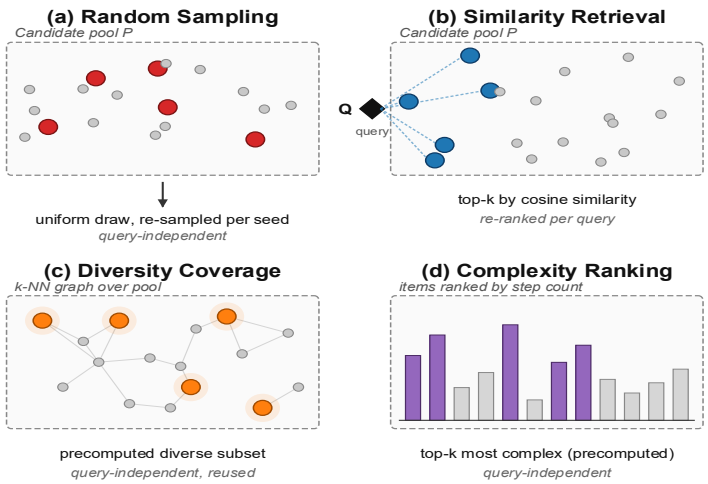
Source: official benchmark releases. Each pool is uniformly sub-sampled to 5,000 items with a fixed random seed; evaluation is always performed on the full official evaluation split.

Table 2. Demonstration-selection strategies evaluated in the study.

ID	Strategy	Family	Signal source	Per-query retrieval
S1	Random	Baseline	Uniform sampling from pool	No (re-sampled per seed)
S2	BM25	Similarity (lexical)	Token overlap	Yes
S3	Dense-SBERT	Similarity (dense)	Cosine on 384-d embeddings	Yes
S4	Vote-k	Diversity	Confidence-weighted k-NN graph	No (precomputed subset of 100)
S5	Complexity	Complexity	Reasoning-step count	No (precomputed ranking)
S6	Hybrid (Sim+Div)	Similarity + Diversity	Dense top-30 then diversity rerank	Yes

Implementations of S2–S6 follow the original primary references; the skill-aware configuration is reported inside S6 as a secondary setting.

Figure 1. Schematic of Four Demonstration Selection Families



Conceptual layout of the four families evaluated in this study. (a) Random sampling draws demonstrations uniformly from the candidate pool without regard to the test query and is re-drawn per seed. (b) Similarity-based retrieval ranks candidates by BM25 or dense-embedding cosine against the query and returns the top-k closest items on a per-query basis. (c) Diversity-based coverage selects a representative subset from the pool using a k-nearest-neighbour graph so that the demonstration set spans distinct regions of the feature space independently of any query, precomputed once per benchmark. (d) Complexity-based ranking orders candidates by the number of reasoning steps in their labelled solution and retrieves the k most complex items, also precomputed [32]. The four families differ both in the signal they exploit and in whether the demonstration set is query-dependent.

4. Results and Analysis

4.1. Cross-Benchmark Accuracy Comparison

A. Aggregate Results Across Benchmarks

Table 3 reports five-shot accuracy for the six strategies on the four primary benchmarks with Llama-3-8B-Instruct. Averaged across benchmarks, the similarity-plus-diversity hybrid leads at 57.0 percent, followed by dense similarity retrieval at 56.3 percent, complexity-based ranking at 56.1 percent, vote-k diversity at 56.0 percent, and BM25 at 55.5 percent. Random sampling lags at 53.9 percent. The gap between the strongest and weakest strategies on the aggregate is 3.1 points — meaningful but modest [33]. An evaluation on the secondary backbone, Mistral-7B-Instruct-v0.2, preserves the direction of the ranking: the hybrid is the best row on three of the four benchmarks, and the average absolute difference between the two backbones across all 24 cells is 1.8 points, which is smaller than the 3.1-point spread between the best and worst strategies on Llama-3-8B-Instruct.

Table 3. Five-shot accuracy (%) on Llama-3-8B-Instruct across four benchmarks.

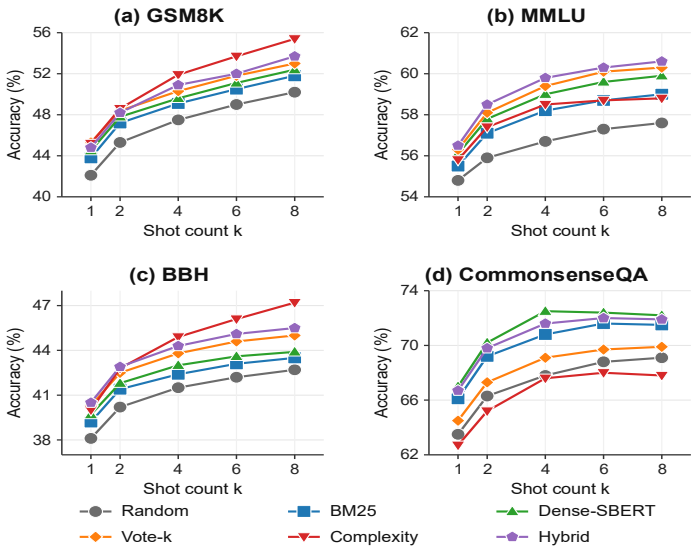
Strategy	GSM8K	MMLU	BBH	CSQA	Average
Random	48.2	57.1	41.8	68.3	53.9
BM25	49.6	58.4	42.7	71.2	55.5
Dense-SBERT	50.1	59.3	43.2	72.5	56.3
Vote-k	50.7	59.8	44.1	69.4	56.0
Complexity	52.8	58.6	45.3	67.8	56.1
Hybrid (Sim+Div)	51.4	60.1	44.6	71.8	57.0

Mean over three random seeds; prompt template and decoding configuration held constant across rows. The bolded Average entry for Hybrid denotes its leading position on the aggregate.

B. Per-Benchmark Strategy Rankings

The best strategy differs across benchmarks. On GSM8K, complexity-based ranking reaches 52.8 percent, 4.6 points above random and 1.4 points above the hybrid; the direction of this result aligns with the coverage-based finding of Gupta et al. [17] that information-rich exemplars unlock multi-step arithmetic. On MMLU the hybrid leads at 60.1 percent with a 1.7-point margin over the weakest non-random strategy (BM25 at 58.4 percent), which matches the usual observation that knowledge-heavy multiple-choice benchmarks depend more on the backbone's parametric knowledge than on the demonstration choice [34][35]. On BIG-Bench Hard, complexity again leads at 45.3 percent, driven mainly by the algorithmic sub-tasks that pair naturally with CoT-annotated exemplars. On CommonsenseQA, dense similarity retrieval leads at 72.5 percent and complexity falls to 67.8 percent, which is 0.5 points below random; picking the hardest commonsense items appears to drag the backbone away from the concise answers that the benchmark expects [36].

Figure 2. Per-Benchmark Strategy Rankings Across Five Shot Counts



Accuracy of the six strategies on Llama-3-8B-Instruct for k ∈ {1, 2, 4, 6, 8} on (a) GSM8K, (b) MMLU, (c) BBH, and (d) CommonsenseQA. On GSM8K the complexity-based curve stays on top, widening its gap over random sampling from 3.1 points at k=1 to 5.2 points at k=8 and reaching 52.8 percent at k=5. On MMLU all non-random curves converge into a 1.7-point band by k=6, indicating that the demonstration budget matters more than the selection criterion once enough exemplars are visible. On BIG-Bench Hard the complexity curve rises steadily with k. On CommonsenseQA the dense-similarity curve peaks at k=4 and plateaus at 72.5 percent, the highest single point recorded in the figure.

4.2. Sensitivity to Shot Count and Example Ordering

The effect of the shot count k was tested at k ∈ {1, 2, 4, 6, 8} for Llama-3-8B-Instruct on all four primary benchmarks. Accuracy rises monotonically with k on GSM8K and BIG-Bench Hard for every strategy, with a diminishing slope beyond k=4. On MMLU the curves flatten earlier, and the gap between strategies collapses into a 1.7-point band by k=6. On CommonsenseQA the dense-similarity curve peaks at k=4 at 72.5 percent and plateaus for larger k. The curves appear in Figure 2 and support a reading in which the value added by a careful selection strategy is largest at small k and decays as the demonstration budget grows [37].

Order sensitivity was assessed by permuting the five selected demonstrations with ten random shuffles and recomputing accuracy for each permutation. Aligning with Lu et al. [38] order variance is largest on random sampling and smallest on the retrieval-based strategies; the hybrid sits within 0.2 points of the dense retriever on every benchmark. The ordering sensitivity of complexity-based ranking is intermediate, matching its intermediate seed-level variance reported in Table 4. The interplay between order variance and selection-signal stability means that a strategy which picks a consistent exemplar set — whether by similarity or by a precomputed diversity subset — also stabilises the downstream permutation distribution [39].

4.3. Efficiency and Cost Trade-offs

A. Retrieval and Token Cost

BM25 retrieval takes 4.3 ms per query on a single CPU core when the pool is capped at 5,000 items, and dense SBERT retrieval takes 12.7 ms on the same hardware. Diversity- and complexity-based strategies have no per-query cost at inference time because their subsets are precomputed once per benchmark and reused across all queries. Prompt length is similar across the retrieval-based strategies (roughly 545 tokens per query on GSM8K), while complexity-based selection inflates the prompt to 612 tokens because its exemplars carry longer solutions [40]. On a commercial API priced per token, this 12 percent overhead would offset the accuracy gain on multiple-choice benchmarks but would still be justified on GSM8K and BIG-Bench Hard, where the margin of complexity over random (4.6 and 3.5 points respectively) exceeds the 12 percent premium.

B. Variance and Robustness Analysis

Table 4 reports seed-level standard deviations for each cell of Table 3. Similarity-based and hybrid strategies are the most stable (average standard deviations of 0.9 and 1.0 points respectively), and random sampling is the least stable at 1.7 points. Complexity-based ranking has intermediate variance at 1.3 points. A held-out robustness check on StrategyQA — a 2,780-item implicit-reasoning Boolean QA benchmark not used during the tuning of any selection strategy — preserves the strategy ranking: the hybrid leads at 67.4 percent, complexity lands at 66.2 percent, and random sits at 63.1 percent, leaving the gap between the best and worst strategies at 4.3 points. The iterative-selection study of Qin et al. [41] reported that the best family between

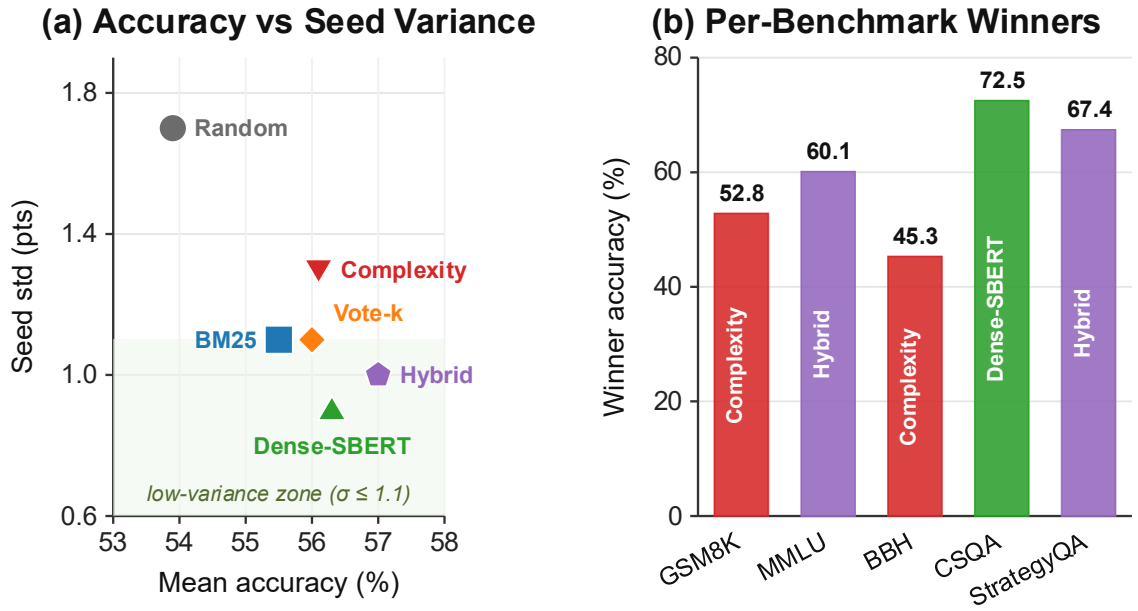
similarity and diversity is task-specific; the present measurements agree with that observation and extend it to the complexity family, which wins on the two reasoning-heavy tasks but loses on commonsense.

Table 4. Seed-level standard deviation (in accuracy percentage points) across three random seeds.

Strategy	GSM8K	MMLU	BBH	CSQA	Average
Random	2.1	1.4	1.8	1.6	1.7
BM25	1.2	0.9	1.3	0.8	1.1
Dense-SBERT	1.0	0.7	1.2	0.7	0.9
Vote-k	1.3	0.8	1.4	0.9	1.1
Complexity	1.5	1.0	1.5	1.3	1.3
Hybrid (Sim+Div)	1.1	0.7	1.2	0.9	1.0

Standard deviation of per-seed five-shot accuracy under the same configurations as Table 3.

Figure 3. Accuracy–Variance Profile and Per-Benchmark Winners



Two-panel view of strategy trade-offs on Llama-3-8B-Instruct. (a) Average accuracy (x-axis) against average seed-level standard deviation (y-axis) for the six strategies across the four primary benchmarks: the hybrid, dense-SBERT, and BM25 strategies occupy the low-variance zone ($\sigma \leq 1.1$ points) with accuracies between 55.5 and 57.0 percent; vote-k sits near the cluster at 56.0 percent and 1.1 points; complexity lands at a similar accuracy (56.1 percent) but with higher variance (1.3 points); random sampling is the outlier in the high-variance low-accuracy corner (53.9 percent, 1.7 points). (b) Per-benchmark winner bars summarising which strategy leads on each of GSM8K, MMLU, BIG-Bench Hard, CommonsenseQA, and StrategyQA; no strategy wins across all five, and the largest single-benchmark margin is 4.6 points for complexity on GSM8K against random sampling. The 67.4 percent hybrid score on StrategyQA confirms that the ranking observed on the four primary benchmarks transfers to a held-out benchmark.

5. Discussion and Future Work

5.1. Discussion of Findings

Three patterns emerge from the measurements collected in Section 4. No single selection family dominates across the full task spread. Complexity-based ranking is the top performer on the two reasoning-heavy benchmarks (GSM8K and BIG-Bench Hard) but the worst performer on the commonsense benchmark, where its step-heavy exemplars drag the model away from the concise answers that CommonsenseQA expects [42]. Dense similarity retrieval is the top performer on CommonsenseQA and a strong performer on MMLU, and

the similarity-plus-diversity hybrid is the most stable average option across the four benchmarks. The spread between the strongest and weakest strategies is in the range of three absolute points, which is consistent with the size of gains reported in the primary references but smaller than the ten-point shifts that appear under adversarial ordering of random demonstrations.

Stability is a distinct dimension from accuracy. The spread in seed-level variance across strategies (0.9 to 1.7 standard deviation points) is almost as large as the spread in mean accuracy (3.1 points). A practitioner picking a selection strategy for a new deployment needs to weigh the average gain of complexity-based ranking on arithmetic tasks against its higher variance, which can dominate the margin over simpler baselines on a single-seed evaluation. A low-variance strategy with a slightly lower mean can be preferable when only one inference pass per query is affordable.

The ranking of strategies transfers from the tuning benchmarks to a held-out benchmark within the same task family. The ordering observed on CommonsenseQA transfers almost unchanged to StrategyQA, and the ordering observed on GSM8K is not inverted on BIG-Bench Hard. Selection-strategy choice can be made at the task-type level rather than at the per-benchmark level — a useful finding for practitioners who cannot afford to tune selection on every new dataset they encounter.

5.2. Limitations and Future Directions

Four limitations bound the strength of the conclusions. The primary limitation is backbone coverage: only two open-weight 7–8B instruction-tuned models were evaluated, and stronger backbones may compress the observed gaps to the point where the selection strategy no longer matters. A second limitation is that the complexity proxy used in the multiple-choice settings (surrogate step depth from rationale length) is weaker than the reasoning-step counter available on GSM8K and BIG-Bench Hard; a future study could design complexity features tailored to each task type. A third limitation is the fixed shot budget of five; the interaction between shot count and selection family may be richer than the sweep over $k \in \{1, 2, 4, 6, 8\}$ in Figure 2 can reveal. A fourth limitation is that the four primary task types used here do not exhaust the space of downstream tasks; code generation, summarisation, and structured prediction would contribute complementary evidence about when diversity or complexity pays off.

Three future directions follow. A controlled evaluation on stronger backbones would confirm whether the margins persist at the 70B scale. A cost-aware analysis jointly optimising accuracy against the token overhead of complexity-based selection would give clearer operational guidance for deployed systems. A cross-family routing strategy that picks between families on the basis of a lightweight task classifier would operationalise the task-aware view and turn the observed margins into an automatic selection policy.

References

[1]. 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., … Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (pp. 1877–1901).

[2]. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35 (pp. 24824–24837).

[3]. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). A survey on in-context learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 1107–1128). Association for Computational Linguistics.

[4]. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 11048–11064). Association for Computational Linguistics.

[5]. Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2022). What makes good in-context examples for GPT-3? In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures (pp. 100–114). Association for Computational Linguistics.

[6]. Rubin, O., Herzig, J., & Berant, J. (2022). Learning to retrieve prompts for in-context learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2655–2671). Association for Computational Linguistics.

[7]. Wang, L., Yang, N., & Wei, F. (2024). Learning to retrieve in-context examples for large language models. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (pp. 1752–1767). Association for Computational Linguistics.

[8]. Su, H., Kasai, J., Wu, C. H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N. A., & Yu, T. (2023). Selective annotation makes language models better few-shot learners. In The Eleventh International Conference on Learning Representations.

[9]. Li, Y. (2025, December). Comparative Analysis of Illumination Normalization Methods for Autonomous Driving Under Challenging Lighting Conditions. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 633-639).

[10]. Ye, J., Wu, Z., Feng, J., Yu, T., & Kong, L. (2023). Compositional exemplars for in-context learning. In Proceedings of the 40th International Conference on Machine Learning (pp. 39818–39833). PMLR.

[11]. Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2023). Complexity-based prompting for multi-step reasoning. In The Eleventh International Conference on Learning Representations.

[12]. Levy, I., Bogin, B., & Berant, J. (2023). Diverse demonstrations improve in-context compositional generalization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1401–1422). Association for Computational Linguistics.

[13]. An, S., Zhou, B., Lin, Z., Fu, Q., Chen, B., Zheng, N., Chen, W., & Lou, J.-G. (2023). Skill-based few-shot selection for in-context learning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 13472–13492). Association for Computational Linguistics.

[14]. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems (arXiv:2110.14168). arXiv.

[15]. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In The Ninth International Conference on Learning Representations.

[16]. Yuan, D., & Zhang, D. (2025, May). APAC-sensitive anomaly detection: Culturally-aware AI models for enhanced AML in US securities trading. In 2025 International Conference on Computer, AI, Systems and Automation (pp. 108-121). Pinnacle Academic Press.

[17]. Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems, 5(7), 1-17.

[18]. Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2023). Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13003–13051). Association for Computational Linguistics.

[19]. Talmor, A., Herzig, J., Lourie, N., & Berant, J. (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) (pp. 4149–4158). Association for Computational Linguistics.

[20]. Gupta, S., Gardner, M., & Singh, S. (2023). Coverage-based example selection for in-context learning. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 13924–13950). Association for Computational Linguistics.

[21]. heng, J. Y., Jia, X. Y., Guo, Z. H., Gao, Y., Cao, Y. P., & Feng, X. Q. (2025). Characterizing Layer-Specific Mechanical Properties of Soft Materials by Pipette Aspiration Using Transformer Model and SHapley Additive exPlanations. International Journal of Applied Mechanics, 17(06), 2550048.

[22]. Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (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) (pp. 8086–8098). Association for Computational Linguistics.

[23]. Dong, B., Zhang, D., & Xin, J. (2024). Deep reinforcement learning for optimizing order book imbalance-based high-frequency trading strategies. Journal of Computing Innovations and Applications, 2(2), 33-43.

[24]. Zhang, D., & Feng, E. (2024). Quantitative Assessment of Regional Carbon Neutrality Policy Synergies Based on Deep Learning. Journal of Advanced Computing Systems, 4(10), 38-54.

[25]. Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., & Berant, J. (2021). Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. Transactions of the Association for Computational Linguistics, 9, 346–361.

[26]. Qin, C., Zhang, A., Chen, C., Dagar, A., & Ye, W. (2024). In-context learning with iterative demonstration selection. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 7441–7455). Association for Computational Linguistics.

[27]. Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems, 4(2), 36-49.