A Systematic Evaluation of Active Learning Strategies for LLM-based Generative Tasks

Chun-Heng Chien **Thuy-Trang Vu**
Department of Data Science & AI, Monash University
cchi0115@student.monash.edu
trang.vu1@monash.edu

Abstract

Active learning (AL) reduces annotation costs by prioritising the labeling of the most informative unlabeled samples. While AL has been extensively studied for discriminative classifiers, its effectiveness in generative settings, where large language models (LLMs) produce output labels directly rather than predicting a distribution over a fixed label set, remains underexplored, particularly for decoder-only LLMs. Prior research has primarily explored LLMs as annotators or query selectors, leaving it unclear whether established AL methods are effective when applied to generative learners. To address this gap, we systematically evaluate representative uncertainty-based, diversity-based, and hybrid acquisition strategies across three text-generation tasks: reading comprehension, multi-step reasoning, and summarisation. Our experimental results reveal that the effectiveness of AL is highly task-dependent, with different acquisition strategies exhibiting distinct strengths across tasks. These findings demonstrate that no single AL strategy is universally effective and offer practical guidance for designing task-specific hybrid acquisition strategies for generative learners.

1 Introduction

Active learning (AL) is a well-established paradigm for improving model performance under limited annotation budgets, by prioritising the most informative unlabelled data points for labeling and training (Settles, 2009; Zhang et al., 2022). Recently, large language models (LLMs) have been increasingly integrated into the AL loop (Xia et al., 2025a), serving either as generators of synthetic training instances (Qian et al., 2024; Yuan et al., 2024) or as cost-effective alternatives to human annotators in training downstream classifiers (i.e., *active learners*) (Xiao et al., 2023; Qi et al., 2026). Despite this growing adoption of LLMs within AL, the effect of established AL query strategies on LLM-based active learners remains insufficiently understood.

Prior AL research has mainly focused on discriminative tasks such as text classification where the label space is discrete and finite (Vu et al., 2019; Zhang and Plank, 2021). Within this setting, acquisition strategies generally fall into three broad categories: (i) uncertainty-based methods which prioritise the most uncertain instances judged by the model outputs (Culotta and McCallum, 2005; Margatina et al., 2021; Schröder et al., 2022), (ii) diversity-based methods which select representative instances to maximise coverage of the input space (Erdmann et al., 2019; Yuan et al., 2020) and (iii) hybrid methods which combine both uncertainty and diversity criteria (Ash et al., 2020; Shi et al., 2021). While these strategies have proven certain effectiveness for discriminative tasks (Lowell et al., 2019), their applicability to generative settings, which are becoming increasingly prevalent with the rise of LLMs, has received limited systematic study.

Unlike discriminative tasks, generative tasks involve a large open-ended output space in which a single input may correspond to multiple semantically equivalent responses with different surface forms (Kuhn et al., 2023; Xia et al., 2025b). As a result, conventional notions of uncertainty and informativeness become less well-defined, raising fundamental questions about whether established AL strategies remain effective in generative settings. To address this question, we conduct a systematic and extensive analysis of three uncertainty-based acquisition strategies, including verbal uncertainty, self-consistency, and perplexity, and two diversity-based strategies based on semantic embeddings and structural tags extracted from model-generated responses. Finally, we examine hybrid acquisition strategies that integrate uncertainty and diversity signals to determine whether such combinations yield more robust strategies for generative learners.

We evaluate these strategies across three repre-

sentative generative tasks: reading comprehension, commonsense reasoning, and summarisation. Our experiment results demonstrate that AL strategies developed for discriminative models do not transfer uniformly to generative tasks, and highlight the importance of designing task-aware acquisition strategies for LLM-based AL. Specifically, we made the following contributions and findings:

- The effectiveness of AL is highly task-dependent. Self-Self-Consistency is the strongest individual strategy, significantly outperforming random sampling on reading comprehension and at several budgets on commonsense reasoning. In contrast, AL offers little or no improvement over random sampling for summarisation, suggesting that the strong prior knowledge encoded in pretrained LLMs may already provide sufficiently informative representations for this task, thereby limiting the benefits of selective sampling.
- Hybrid acquisition strategies are beneficial only when uncertainty and diversity provide complementary information.
- The ordering of hybrid acquisition pipelines is a critical design choice. Commonsense reasoning benefits from a *diversity-first* pipeline that promotes broad exploration of the input space before uncertainty estimation, whereas reading comprehension achieves better performance with an *uncertainty-first* pipeline when uncertainty estimates are sufficiently reliable.

2 Preliminary

Active learning (AL) aims to maximise model performance under a limited annotation budget by iteratively selecting the most informative unlabeled instances for annotation. Let $\mathcal{D}_U = \{x_1, \ldots, x_N\}$ denote an unlabeled pool and $\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^{n}$ a labeled set, where $y_i \in \mathcal{Y}$. At each iteration, a model f_θ selects instances for annotation according to a query strategy $Q(\cdot)$:

$$x^* = \arg\max_{x \in \mathcal{D}_U} Q(x; f_\theta) \tag{1}$$

after which the newly labeled instances are added to $\mathcal{D}_L$, removed from $\mathcal{D}_U$, and the model is updated. The process repeats until the annotation budget is exhausted.

In discriminative tasks, the output space $\mathcal{Y}$ is typically a finite set of class labels, allowing query strategies to estimate sample informativeness using predictive uncertainty. In contrast, generative tasks define $\mathcal{Y}$ as an open-ended space of token sequences, where a single input may admit multiple valid outputs. Therefore,the notions of uncertainty become less well-defined, motivating the generative query strategies investigated in this work.

3 AL for Text Generation Tasks

We present the AL query strategies evaluated in this study, organising into three categories: uncertainty-based, diversity-based and hybrid methods that integrate both signals. Table 1 provides a comparison of the pros and cons of each method.

3.1 Uncertainty-Based Methods

Uncertainty-based methods aim to select instances for which the current model is most uncertain. In discriminative tasks, uncertainty is typically estimated from prediction probabilities over a finite label space using metrics such as entropy, least confidence, or margin sampling (Zhang et al., 2022). Extending these measures to generative models is challenging due to their open-ended output space and the existence of multiple semantically equivalent responses. Therefore, uncertainty for LLMs is better estimated using sequence- or semantic-level signals rather than token-level probabilities. In this work, we investigate three uncertainty estimation strategies: *Verbal Uncertainty*, *Self-Consistency*, and *Perplexity*.

Verbal Uncertainty estimates confidence by prompting the model f_θ to explicitly report its confidence for a generated response $y = f_\theta(x)$ as a scalar score $c(x) \in [0, 1]$ (Yang et al., 2024). The acquisition score is defined as

$$Q_{\mathrm{VU}}(x) = 1 - c(x), \tag{2}$$

where larger values indicate greater uncertainty. The prompt templates used for each task are provided in section A.2. This approach is broadly generalizable and computationally cheap, but is heavily dependent on the model's calibration and prompt sensitivity.

Perplexity estimates uncertainty from the conditional likelihood assigned by f_θ to its generated response. For a generated sequence $y = (y_1, \ldots, y_N)$, the acquisition score is

$$Q_{\mathrm{PPL}}(x) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \log P(y_i \mid x, y_{<i})\right), \tag{3}$$

where higher perplexity indicates higher uncertainty. Although widely used as a likelihood-based uncertainty measure, perplexity primarily captures uncertainty over surface forms and may assign high scores to correct but uncommon patterns.

Self-Consistency estimates uncertainty from the agreement among multiple responses sampled from the same model (Wang et al., 2022). Given an input x, we generate K responses $\{y^{(1)}, \ldots, y^{(K)}\}$ and extract the corresponding answers $A = \{a^{(1)}, \ldots, a^{(K)}\}$. The acquisition score is defined based on the proportion of the majority vote:

$$Q_{\mathrm{SC}}(x) = 1 - \frac{\max_{a \in A} \mathrm{count}(a)}{K}, \qquad (4)$$

where instances with lower agreement are considered more informative.

3.2 Diversity-Based Methods

Diversity-based methods aim to maximise the coverage of the unlabeled pool by selecting representative instances from different regions of the input space. Formally, the acquisition function $Q(x; f_\theta)$ favors samples that maximise diversity with respect to the current query set. In this work, we investigate two complementary approaches for representing diversity: semantic embeddings and LLM-generated structural tags.

Semantic Embeddings represent each unlabeled instance $x \in \mathcal{D}_U$ as a dense embedding vector $\mathbf{e}(x)$ extracted from a pretrained language model. K-Means clustering is then applied to partition the embedding space into B clusters, where B is the query budget, and the instance closest to each cluster centroid is selected. The resulting acquisition function promotes semantic coverage of the input space while remaining task-agnostic.

LLM-Tags uses an auxiliary language model to generate structured tags describing the underlying concepts or reasoning patterns of each unlabeled instance (Lu et al., 2024). The tags are clustered using Agglomerative Clustering to construct a concept space, after which a greedy maximum-coverage algorithm selects a diverse set of concepts and representative instances nearest to each cluster centroid. Compared with semantic embeddings, LLM-Tags better captures problem-level diversity beyond surface semantics, but incurs higher computational cost and depends on the quality of the generated tags. Detailed algorithms are provided in Algorithm 1.

Algorithm 1 Diversity Sampling via LLM-Tags

Require: Unlabeled data pool $\mathcal{U}$, Query budget B, Tagger LLM M_{tag}, Prompt Template P
Ensure: Selected query set $\mathcal{Q}$

$\mathcal{T} \leftarrow \emptyset$
for each $u_i \in \mathcal{U}$ **do**
  $prompt_i \leftarrow P(u_i)$
  $tag_i \leftarrow M_{tag}(prompt_i)$
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{(u_i, tag_i)\}$
end for
$\mathcal{V}_{tag} \leftarrow \text{Vectorize}(\{tag_i\}_{i=1}^{|\mathcal{U}|})$
$\mathcal{C} \leftarrow \text{Clustering}(\mathcal{V}_{tag}, \text{num_clusters} = B)$
$\mathcal{Q} \leftarrow \emptyset$
for each cluster $c_j \in \mathcal{C}$ **do**
  Sample one representative u^* based on tag centroid
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{u^*\}$
end for
return $\mathcal{Q}$

3.3 Hybrid Methods

Uncertainty and diversity capture complementary notions of sample informativeness. While uncertainty-based methods prioritise samples on which the model is most uncertain, diversity-based methods promote broad coverage of the input space. Hybrid methods therefore combine both criteria within a single acquisition pipeline. We investigate two pipeline variants that differ in the order in which the acquisition functions are applied.

Uncertainty-First ranks the unlabeled pool $\mathcal{D}_U$ using an uncertainty-based acquisition function $Q_U(\cdot)$ and retains the top-M most uncertain instances ($M >> B$). A diversity-based acquisition function $Q_D(\cdot)$ is then applied to this reduced candidate set to select the final query set of size B. This approach focuses computation on highly uncertain samples, but may overlook informative regions of the input space if uncertainty estimates are unreliable.

Diversity-First partitions the entire unlabeled pool using a diversity-based acquisition function $Q_D(\cdot)$ to obtain a representative set of clusters. The uncertainty acquisition function $Q_U(\cdot)$ is then applied independently within each cluster, selecting the most uncertain instance as the query. This preserves global coverage of the input space while prioritising uncertain samples within each region.

AL Method	Pros	Cons
	Uncertainty-Based Methods	
Verbal	• Broad generalizability • Exceptionally low computational cost	• Dependent on model calibration • Highly sensitive to prompt design • Black-box mechanism
Perplexity	• Mathematically rigorous and objective • Directly leverages internal logits	• Captures surface-form uncertainty • Unfairly penalizes atypical phrasing
Self-Consistency	• Evaluates output variance directly	• Unsuitable for open-ended tasks • Requires strict logical convergence
	Diversity-Based Methods	
Semantic Embeddings	• Task-agnostic and computationally efficient • No task-specific prompting required	• Captures surface-level textual similarity • Misses genuine algorithmic diversity
LLM-Tags	• Projects into problem-oriented space • Captures deeper, authentic diversity	• High computational overhead • Bottlenecked by tagging LLM's capabilities
	Hybrid Methods	
Uncertainty-First	• Strictly guarantees high uncertainty • Reduces downstream clustering cost	• Risk of early information loss • May operate on a fundamentally biased pool
Diversity-First	• Preserves global structural topology • Guarantees broad conceptual coverage	• May waste budget on trivial data points • Significantly higher computational cost

Table 1: Comparison of active learning query strategies

3.4 Comparison of Query Strategies

The evaluated acquisition strategies exhibit complementary strengths and limitations (Table 1). Uncertainty-based methods directly exploit model confidence, but their effectiveness depends on the quality of uncertainty estimation and may not generalize across different generative tasks. Diversity-based methods improve coverage of the input space and are largely model-agnostic, yet they often fail to capture task-specific informativeness. Hybrid methods combine both signals to balance exploration and exploitation, although their effectiveness depends on how the two criteria are integrated. These trade-offs motivate our systematic comparison across representative generative tasks.

4 Experimental Settings

This section describes the experimental setup used to evaluate different AL strategies with respect to the following research questions (RQs):

- **RQ1**: How does AL generalise across generative tasks?
- **RQ2**: Do hybrid strategies improve query selection?
- **RQ3**: How does the ordering of uncertainty and diversity acquisition affect the performance of hybrid active learning strategies?

Tasks and Dataset We evaluate all AL methods on three representative generative tasks: **CommonSenseQA** (Talmor et al., 2019) for commonsense reasoning, **SQuAD_v2** (Rajpurkar et al., 2018) for reading comprehension, and **TL;DR** (Völske et al., 2017) for abstractive summarisation. These tasks span a range of generative settings, from structured prediction with well-defined answers to open-ended text generation, enabling us to assess whether the effectiveness of AL strategies generalises across different types of generative tasks.

Parameter	CSQA	SQuAD	TL;DR
LoRA Config			
Rank (r)	16	16	16
Alpha (α)	32	32	32
Batch Size	16	8	16
Learning Rate	1e-4	5e-6	1e-5
AL Setup			
Total Train Pool	9714	15,000*	10,000*
Initial Seed Size	100	100	10
Query Budget	300	300	50
Total Iterations	15	15	10
Total Budget	4300	4300	460
Trials	3	3	3

* Downsampled from the original training samples.

Table 2: Hyperparameters and Active Learning Loop Configurations across Datasets

Models All experiments use `LLaMA-3.2-1B` (Grattafiori et al., 2024) as the active learner to ensure a consistent evaluation across query strategies. We select the 1B model because it presents a more challenging learning setting than larger, more capable models, making the impact of AL under varying annotation budgets easier to observe.

Query Strategy	Reasoning - CommonSenseQA		Reading Comp. - SQuAD_v2		Summarisation - TL;DR	
	AULC	Final EM	AULC	Final EM	AULC	Final R-1
Baseline						
Random	1136.40±11.55	37.50±1.12	1563.85±25.49	43.03±0.54	73.77± 0.81	**20.73±0.25**
Uncertainty-Based						
Perplexity (PPL)	1121.20±37.54	36.23±0.62	1366.30±67.07	39.47±0.62	69.13± 0.15	18.90±0.37
Verbal	1122.55±29.13	36.80±2.51	1651.50±18.57⋆	44.80±0.65	**74.00±0.41**	20.17±0.78
Self-Consistency (SC)	1187.53±33.14	**37.65±0.80**	1658.40±37.00⋆	**47.27±0.74**⋆	N/A†	N/A†
Diversity-Based						
Semantic Embedding (SE)	1048.45±23.64	36.87±1.13	1386.05±21.01	42.57±1.97	73.70± 0.29	20.70±0.78
LLM-Tags	1155.95±29.72	37.20±1.49	1543.00± 2.16	42.87±0.37	72.47± 1.16	20.17±0.91
Hybrid: Uncertainty-First						
Self-Consistency + SE	1145.25±45.85	36.60±0.45	**1668.20±46.32**⋆	47.00±0.83⋆	N/A†	N/A†
Self-Consistency + LLM-Tags	1129.05±52.25	35.27±1.03	1583.95±95.91	44.57±1.73	N/A†	N/A†
Verbal + SE	1134.90±65.87	35.90±1.02	1618.40±32.95	43.87±0.65	73.27± 0.87	19.70±0.29
Verbal + LLM-Tags	1128.95±40.51	34.73±0.75	1602.10±18.55	44.67±1.04⋆	73.81± 1.11	19.20±0.37
Perplexity + SE	1124.25±29.74	35.13±1.04	1441.65±16.37	41.43±0.90	70.56± 0.55	19.07±0.50
Perplexity + LLM-Tags	1095.60±48.65	34.43±1.80	1432.05±35.17	43.23±1.25	71.92± 0.93	18.23±0.70
Hybrid: Diversity-First						
SE + Self-Consistency	**1196.05±38.49**	37.33±0.53	1627.85±66.24	44.37±0.66	N/A†	N/A†
LLM-Tags + Self-Consistency	1150.70±17.36	35.90±0.75	1601.20± 8.89	44.80±0.51	N/A†	N/A†
SE + Verbal	1154.75±68.67	35.60±1.21	1628.65±24.99	45.97±0.68⋆	72.67± 0.40	19.50±0.14
LLM-Tags + Verbal	1137.50±16.96	35.63±2.13	1576.90±16.45	43.67±0.39	72.79± 0.22	20.20±0.71
SE + Perplexity	1121.40±33.80	35.97±0.41	1429.90±24.08	41.07±1.10	69.77± 0.66	18.97±0.17
LLM-Tags + Perplexity	1134.00±11.88	35.30±0.99	1457.60±35.28	41.40±1.93	72.05± 1.59	18.97±0.45

† Note: Self-Consistency (SC) is not applicable to the open-ended summarisation task (TL;DR).
⋆ Note: denotes a statistically significant improvement over random sampling (one-tailed t-test, $p < 0.05$).

Table 3: Overall performance of AL query strategies and the random baseline across three generative tasks. For each dataset, we report the Area Under the Learning Curve (AULC) and the Final Metric. **Color coding**: Green indicates performance exceeding the random baseline, while red indicates underperformance. Underlined values denote the best result within each specific category, and **bold** values highlight the absolute best performance across all evaluated methods.

Active Learning Setup We adopt a task-specific AL configuration for each benchmark. CommonSenseQA and SQuAD_v2 use an initial seed set of 100 labeled instances, followed by 15 acquisition rounds with a query budget of 300 samples per round, yielding a total annotation budget of 4,300 instances. For TL;DR, the learner is initialized with 10 labeled examples and performs 10 acquisition rounds with a query budget of 50 samples, resulting in a total budget of 460 labeled instances. We repeat each experiment over three independent trials. Table 2 summarises the full configuration.

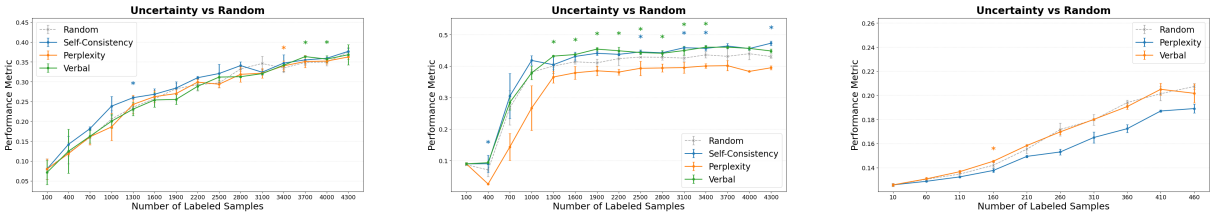
Implementation The verbal uncertainty prompts are provided in Appendix Section A.2. For hybrid methods, we introduce an intermediate candidate pool of size βB, where B is the query budget and $\beta = 5$ by default. In the Diversity-First pipeline, we first partition the unlabeled pool into βB clusters, then select and rank the most uncertain instance from each cluster to obtain the final B queries. Therefore, the Uncertainty-First pipeline first selects the βB most uncertain instances before applying diversity-based clustering to produce the final query set. All experiments are conducted on NVIDIA A100 GPUs and fine-tuned using LoRA (Hu et al., 2022). Complete implementation details, including training hyperparameters, are provided in Table 2.

Evaluation Metrics. Performance is evaluated using task-specific metrics: exact match (EM) for CommonSenseQA and SQuAD_v2, and ROUGE-1 (Lin, 2004) for TL;DR. We assess statistical significance ($\alpha = 0.05$) using one-tailed t-tests from two perspectives: (i) *point-wise comparison*, which compares methods at a specific annotation budget, and (ii) *global efficiency*, which compares the Area Under the Learning Curve (AULC) to capture performance throughout the entire active learning process. All reported results are averaged over three independent trials, with error bars representing one standard deviation.

5 Results

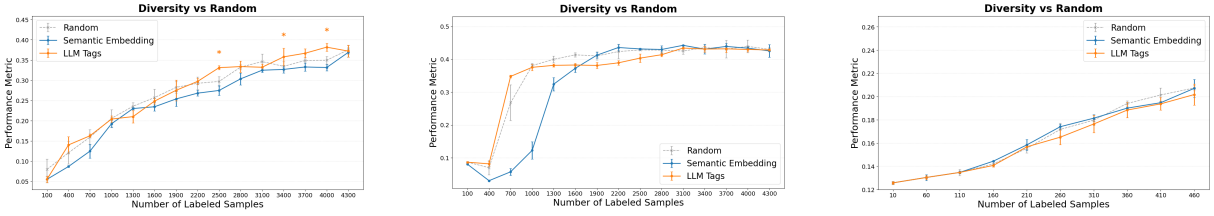
5.1 Overall Performance

Table 3 summarises the performance of all 17 active learning (AL) query strategies across the three



(a) Reasoning (EM) (b) Reading Comprehension (EM) (c) Summarisation (ROUGE-1)

Figure 1: Performance of uncertainty-based strategies vs. Random across annotation budgets. Markers indicate a statistically significant improvement over random sampling (one-tailed t-test, $p < 0.05$).



(a) Reasoning (EM) (b) Reading Comprehension (EM) (c) Summarisation (ROUGE-1)

Figure 2: Performance of Diversity Strategies vs. Random across Tasks. Markers indicate a statistically significant improvement over random sampling (one-tailed t-test, $p < 0.05$).

benchmark tasks. Overall, no single acquisition strategy consistently achieves the best performance, indicating that the effectiveness of AL is highly task-dependent. While uncertainty-based methods generally outperform diversity-based methods on structured tasks, hybrid strategies provide additional gains only in specific settings, and no evaluated strategy consistently improves over random sampling on the open-ended summarisation task.

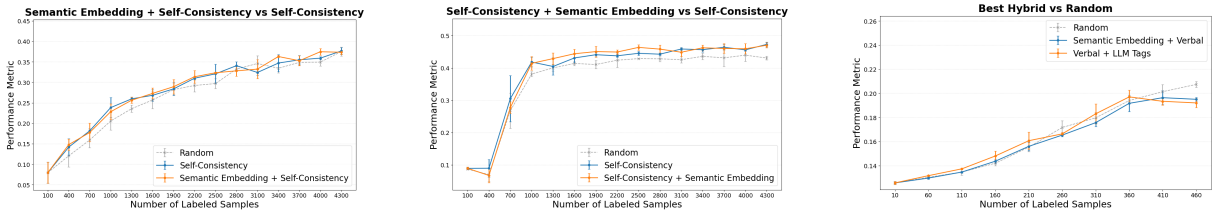
Among the uncertainty-based methods, Self-Consistency (SC) is the strongest individual strategy, achieving the best performance on both CommonSenseQA and SQuAD_v2. In contrast, Perplexity consistently underperforms the random baseline across all tasks, while Verbal Uncertainty exhibits mixed behaviour, producing competitive results on SQuAD_v2 but offering limited benefit for reasoning. Diversity-based methods are similarly task-sensitive: LLM-Tags consistently outperforms Semantic Embeddings on CommonSenseQA, whereas both methods are less effective than uncertainty-based acquisition on SQuAD_v2 and provide little benefit for summarisation.

Hybrid strategies further reveal that combining uncertainty and diversity does not universally improve performance. The Diversity-First SE+SC pipeline achieves the highest overall AULC on CommonSenseQA, whereas the Uncertainty-First SC+SE pipeline performs best on SQuAD_v2. However, neither hybrid configuration consistently surpasses its strongest individual counterpart across tasks, and no hybrid method improves performance on TL;DR. These observations suggest that both the choice of acquisition signal and the design of hybrid pipelines should be adapted to the target generative task rather than applied uniformly. The following subsections examine these findings with respect to the three research questions.

5.2 Generalisation Across Generative Tasks (RQ1)

Figures 1 and 2 compare the performance of different active learning (AL) strategies across varying annotation budgets. To improve readability, we separately visualize uncertainty-based and diversity-based methods against the random baseline, avoiding excessive clutter from displaying all 17 strategies simultaneously.

On CommonSenseQA, SC achieves the highest AULC among uncertainty-based strategies and significantly outperforms random sampling at several budgets, whereas Perplexity (PPL) and Verbal Uncertainty offer little or no improvement (Fig. 1a). Among diversity-based methods, LLM-Tags consistently outperforms Semantic Embeddings (Fig. 2a). A similar pattern is observed on SQuAD_v2, where SC achieves the strongest overall performance and substantially outperforms both diversity-based strategies (Figs. 1b and 2b). In contrast, none of the evaluated strategies consistently surpass random sampling on TL;DR (Figs. 1c and 2c). Overall, these results demonstrate that the effectiveness of active learning is highly task-dependent: uncertainty-based acquisition is particu-



(a) Reasoning (EM) (b) Reading Comprehension (EM) (c) Summarisation (ROUGE-1)

Figure 3: Performance of Best Hybrid vs Best Single AL strategy across Tasks. Markers indicate a statistically significant improvement over random sampling (one-tailed t-test, $p < 0.05$).

larly effective for structured reasoning and reading comprehension, whereas conventional AL strategies provide little benefit for open-ended summarisation.

5.3 Hybrid Strategies Performance (RQ2)

Figure 3 compares the best hybrid strategy against the strongest diversity-based and uncertainty-based acquisition method for each task. As shown in Table 3, for CommonSenseQA, the Diversity-First SE+SC hybrid achieves the highest AULC, slightly outperforming SC alone despite Semantic Embeddings performing poorly individually (Figure 3a). On SQuAD_v2, hybrid strategies closely overlap with SC throughout the acquisition process, indicating only marginal improvement over uncertainty sampling alone (Figure 3b). For TL;DR, hybrid methods remain statistically indistinguishable from random sampling and fail to recover performance beyond the best individual strategy (Figure 3c). These results suggest that hybrid acquisition is beneficial only when uncertainty and diversity capture complementary information. Otherwise, combining multiple acquisition criteria yields little additional benefit.

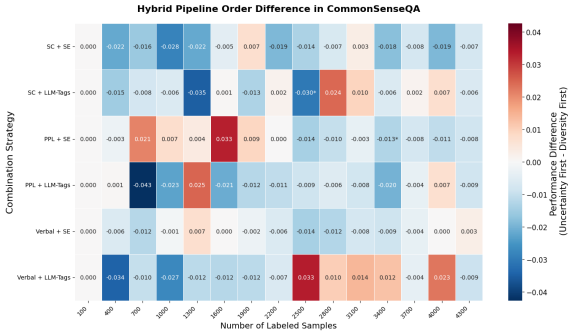
5.4 Hybrid Pipeline Ordering (RQ3)

Figure 4 compares Uncertainty-First and Diversity-First hybrid pipelines. CommonSenseQA consistently favors Diversity-First, with negative performance differences across most hybrid combinations (Figure 4a). Conversely, SQuAD_v2 exhibits a modest but consistent preference for Uncertainty-First, particularly for hybrids built on Self-Consistency (Figure 4b), particularly for hybrids built on Self-Consistency, where Uncertainty-First significantly outperforms Diversity-First at 1,900, 2,200, and 4,300 labeled samples. For TL;DR, performance differences between the two orderings remain close to zero across all hybrid configurations, indicating that pipeline ordering has little influence (Figure 4c). These results demonstrate that pipeline ordering is task dependent rather than universally optimal.

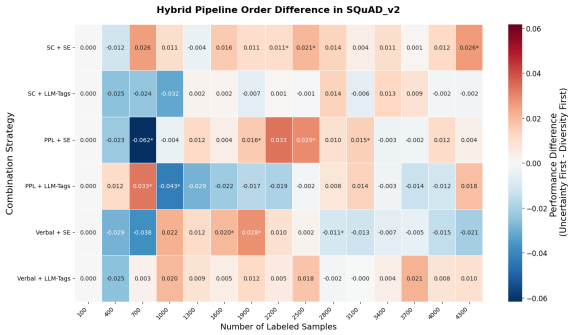
6 Discussion

Our findings demonstrate that AL strategies developed for discriminative models do not transfer uniformly to generative settings. Rather than exhibiting a universally optimal acquisition strategy, AL effectiveness is strongly task-dependent and governed by the interaction between uncertainty estimation and task characteristics. Self-Consistency consistently emerges as the most reliable uncertainty measure for structured generative tasks, while token-level confidence measures such as Perplexity and Verbal Uncertainty provide limited benefit. More importantly, diversity is not inherently beneficial: it improves query selection only when it contributes information that is complementary to uncertainty. This explains why hybrid acquisition yields a small gain on commonsense reasoning but offers only marginal gains for reading comprehension and no measurable benefit for open-ended summarisation.

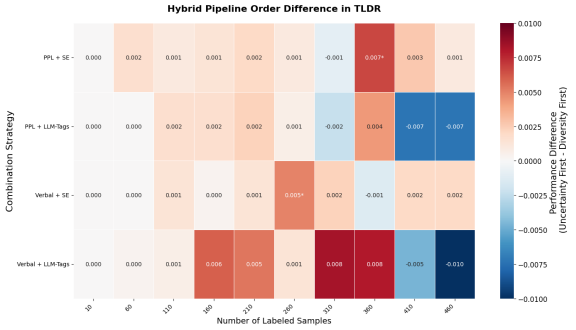
Our results further show that hybrid pipeline ordering should be regarded as a design choice rather than a simple implementation detail. For reasoning tasks, applying diversity before uncertainty preserves broader coverage of the problem space and improves overall learning efficiency. In contrast, reading comprehension benefits from prioritising uncertainty when the uncertainty estimate is sufficiently reliable, whereas pipeline ordering has negligible influence on open-ended summarisation where conventional uncertainty and diversity signals are weak. These findings suggest that future AL frameworks for LLMs should adopt task-aware acquisition strategies, selecting both the acquisition criteria and their composition according to the characteristics of the target generative task rather than relying on a single strategy across all settings.



(a) Reasoning (EM)



(b) Reading Comprehension (EM)



(c) Summarisation (ROUGE-1)

Figure 4: Impact of Pipeline Ordering. * indicates a statistically significant difference ($p < 0.05$)

7 Related Work

LLM-based Active Learning LLMs are increasingly integrated into the AL loop as query selectors or annotation oracles (Xia et al., 2025a). Selection-side methods such as PriorAL (Zhang et al., 2026) and ActivePrune (Azeemi et al., 2026) leverage LLM-derived signals to improve acquisition, while FreeAL (Xiao et al., 2023), NoiseAL (Yuan et al., 2024), and MoLLIA (Qi et al., 2026) replace or refine human annotations with LLMs. These methods target discriminative downstream learners, whereas we study the LLM itself as the active learner.

Active Learning for Generation Tasks AL for generation is more challenging because uncertainty cannot be inferred from a finite label distribution, and sequence-level probabilities often conflate genuine uncertainty with linguistic variation (Shelmanov et al., 2025). Conventional AL strategies frequently yield limited or inconsistent gains over random sampling for generation, summarisation, and reasoning (Perlitz et al., 2023; Muldrew et al., 2024; Stylianos Giouroukis et al., 2025; Zhang et al., 2024a), although task-specific diversity signals, such as hallucination types, can help for summarisation (Xia et al., 2024). Semantic uncertainty signals based on multiple generations, such as explanation disagreement or self-consistency, have shown promise in specific settings (Zhang et al., 2024b; Xiao et al., 2026). Hybrid strategies combining uncertainty and diversity have also emerged (Stylianos Giouroukis et al., 2025; Chen et al., 2026), but their pipeline ordering is typically fixed and evaluated on a single task.

The closest work to ours is Perlitz et al. (2023), who evaluate uncertainty- and diversity-based AL on open-ended NLG tasks with FLAN-T5-Large and find that no strategy consistently outperforms random selection. Despite differences in architecture, adaptation method, and acquisition signals, our results agree with theirs on TL;DR, the setting shared by both studies. Beyond their study, we analyse hybrid pipelines: although hybrids do not significantly outperform the strongest single strategy, they help only when the two signals are complementary, and their ordering can affect performance (Appendix B.2). To the best of our knowledge, we are the first to examine how the effectiveness of hybrid ordering varies with task characteristics.

8 Conclusion

We presented a systematic evaluation of AL query strategies for LLM-based generative learning across three representative tasks: commonsense reasoning, reading comprehension, and abstractive summarisation. Our results show that no single acquisition strategy is universally optimal. Instead, the effectiveness of AL is highly task-dependent: Self-Consistency provides the strongest uncertainty signal for structured tasks, hybrid strategies are beneficial only when uncertainty and diversity offer complementary information, and the optimal ordering of hybrid pipelines varies across tasks. Finally, the limited gains observed on summarisation suggest that conventional uncertainty- and diversity-based acquisition strategies are insufficient for open-ended generation, motivating the development of task-aware query strategies tailored to the characteristics of generative learning.

Limitations

Our study has several limitations. First, due to computational constraints, experiments on SQuAD_v2 and TL;DR were conducted on downsampled training sets and a single 1B-parameter LLM, leaving open the question of whether the observed trends generalize to larger datasets and more capable foundation models. Second, the diversity-based methods rely on external semantic representations, including pretrained embedding models and LLM-generated tags, whose quality directly influences query selection. Misaligned embeddings or noisy structural tags may propagate errors into the acquisition process, particularly for open-ended tasks such as summarisation. Future work should investigate larger-scale models, full-dataset evaluations, and adaptive diversity representations that better align with the underlying task structure. Third, we focus on three representative uncertainty estimators; more advanced uncertainty quantification methods for generation (Vashurin et al., 2025) may provide stronger acquisition signals, especially for summarisation where self-consistency is not applicable. Fourth, summarisation is evaluated with ROUGE-1 only, and complementary metrics may reveal differences not captured here.

Acknowledgments

The authors are grateful to the anonymous reviewers for their helpful comments. This research/work was supported by Monash eResearch capabilities, including M3, Machine Learning eResearch Platform (MLeRP).

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2026. Language model-driven data pruning enables efficient active learning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4373–4392, Rabat, Morocco. Association for Computational Linguistics.

Zhuowei Chen, Liwei Chen, Christian Schunn, Raquel Coelho, and Xiang Lorraine Li. 2026. Neuron-aware active few-shot learning for LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33047–33062, San Diego, California, United States. Association for Computational Linguistics.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. 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 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 21–30, Hong Kong, China. Association for Computational Linguistics.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *International Conference on Learning Representations*, volume 2024, pages 36456–36474.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for

large language models. In *Proceedings of the 41st International Conference on Machine Learning*.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein Dor. 2023. Active learning for natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9862–9877.

Yuanyuan Qi, Xiaohao Yang, Jueqing Lu, Guoxiang Guo, Joanne Enticott, Gang Liu, and Lan Du. 2026. Next generation active learning: Mixture of llms in the loop. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 24909–24917.

Kun Qian, Yisi Sang, Farima Bayat, Anton Belyi, Xianqi Chu, Yash Govind, Samira Khorshidi, Rahul Khot, Katherine Luna, Azadeh Nikfarjam, and 1 others. 2024. Ape: Active learning-based tooling for finding informative few-shot examples for llm-based entity matching. In *Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop (DaSH 2024)*, pages 1–3.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey.

Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. 2025. Uncertainty quantification for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Petros Stylianos Giouroukis, Alexios Gidiotis, and Grigorios Tsoumakas. 2025. Dual: Diversity and uncertainty active learning for text summarization. *arXiv e-prints*, pages arXiv–2503.

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.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the workshop on new frontiers in summarization*, pages 59–63.

Thuy-Trang Vu, Ming Liu, Dinh Phung, and Gholamreza Haffari. 2019. Learning how to active learn by dreaming. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4091–4101, Florence, Italy. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yu Xia, Xu Liu, Tong Yu, Sungchul Kim, Ryan Rossi, Anup Rao, Tung Mai, and Shuai Li. 2024. Hallucination diversity-aware active learning for text summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8665–8677, Mexico City, Mexico. Association for Computational Linguistics.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, and 15 others. 2025a. From selection to generation: A survey of LLM-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, Vienna, Austria. Association for Computational Linguistics.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025b. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396, Vienna, Austria. Association for Computational Linguistics.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023.

FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Zhiping Xiao, Yusheng Zhao, Qixin Zhang, Jiaye Xie, Wanjia Zhao, Weizhi Zhang, Xiao Luo, Philip S. Yu, and Ming Zhang. 2026. Sample lottery: Unsupervised discovery of critical instances for LLM reasoning. In *The Fourteenth International Conference on Learning Representations*.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. Hide and seek in noise labels: Noise-robust collaborative active learning with LLMs-powered assistance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10977–11011, Bangkok, Thailand. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Jiancheng Zhang, Meiqing Li, Qi Zhang, and Yinglun Zhu. 2026. Active learning with foundation model priors: Efficient learning under class imbalance. In *Forty-third International Conference on Machine Learning*.

Linhai Zhang, Jialong Wu, Deyu Zhou, and Guoqiang Xu. 2024a. STAR: Constraint LoRA with dynamic active learning for data-efficient fine-tuning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3519–3532, Bangkok, Thailand. Association for Computational Linguistics.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 395–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yifei Zhang, Bo Pan, Chen Ling, Yuntong Hu, and Liang Zhao. 2024b. ELAD: Explanation-guided large language models active distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4463–4475, Bangkok, Thailand. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Diversity Sampling Algorithms

Algorithm 2 Diversity Sampling via Semantic Embeddings

Require: Unlabeled data pool $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$, Query budget B, Embedding Model E
Ensure: Selected query set $\mathcal{Q}$

$\mathcal{V} \leftarrow \emptyset$
for each $u_i \in \mathcal{U}$ **do**
  $v_i \leftarrow E(u_i)$
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_i\}$
end for
$\mathcal{C} \leftarrow$ K-Means$(\mathcal{V}, \text{num_clusters} = B)$
$\mathcal{Q} \leftarrow \emptyset$
for each cluster $c_j \in \mathcal{C}$ **do**
  $v^* \leftarrow$ Find closest sample to the centroid of c_j
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{u^*\}$
end for
return $\mathcal{Q}$

A.2 Verbal Uncertainty Prompts

1. CommonSenseQA (CSQA)

> Read the following text and answer the question. After answering, provide the probability between 0.0 and 1.0 that your answer is correct. Use the following format to respond:
> Answer: [Your answer text]
> Probability: [Write your probability between 0.0 and 1.0 here.]
>
> Question: {question}
> Options: {options}

2. SQuAD_v2

> Read the following text and answer the question. If the answer is not in the text, reply with 'unanswerable'. After answering, provide the probability between 0.0 and 1.0 that your answer is correct. Use the following format to respond:
> Answer: [Your answer text]
> Probability: [Write your probability between 0.0 and 1.0 here.]
>
> Text: {context}
> Question: {question}

3. TL;DR (Summarisation)

> Read the following post and summarize it. After answering, provide the probability between 0.0 and 1.0 that your answer is correct. Use the following format to respond:
> Answer: [Your answer text]
> Probability: [Write your probability between 0.0 and 1.0 here.]
>
> Post: {post}

B Detailed Analysis

B.1 Reasoning

This section provides the stage-by-stage details underlying the key findings summarized in section 5.

Statistical significance of uncertainty strategies. SC shows statistically significant improvements over random at both early and late acquisition stages, whereas PPL and Verbal confidence show no consistent pattern of significant gains across the acquisition budget.

Diversity strategy. LLM-Tags shows multiple statistically significant gains concentrated at mid-to-late acquisition stages, consistent with the interpretation that tag-based structural representations require a sufficiently large labeled pool before their advantage over random sampling becomes distinguishable. SE shows no stage at which it significantly exceeds random performance.

Rescue effect. The Diversity-First hybrid (SE + SC) closely follows pure SC during early acquisition stages (up to approximately 1,600 labeled samples) before establishing a small but consistent advantage from mid-to-late stages onward. This pattern suggests that the benefit of SE as a structural filter only manifests once the labeled pool is large enough for cluster-based partitioning to meaningfully separate reasoning subtypes.

Pipeline ordering. The advantage of Diversity-First ordering is most pronounced in the SC + SE configuration, where Uncertainty-First underperforms across most acquisition stages. A similar but weaker trend appears in lower-confidence proxy combinations such as Verbal + LLM-Tags, though with stage-dependent fluctuation at later phases—in some later-stage budgets, the ordering advantage narrows or briefly reverses, suggesting

that pipeline ordering effects are most stable when built on a strong uncertainty proxy (SC) and less predictable when built on weaker proxies (PPL, Verbal).

B.2 Reading Comprehension

This section provides the stage-by-stage details underlying the key findings summarized in section 5.

Variance overlap in the best hybrid. The SC + SE hybrid learning curve shows highly intersecting variances with pure SC across all labeled-sample budgets, confirming that the slightly higher AULC score of the hybrid does not correspond to a statistically distinguishable trajectory at any individual budget point.

Significant gains for Uncertainty-First ordering. For the SC + SE configuration, Uncertainty-First shows statistically significant gains over Diversity-First at 1,900, 2,200, and 4,300 labeled samples, while remaining marginally favorable at other budgets without reaching significance.

Volatility in weaker-proxy hybrids. Hybrids with weaker uncertainty indicators show substantial phase-dependent volatility in their ordering preference. PPL + SE favors Diversity-First during early acquisition stages before shifting to favor Uncertainty-First at later stages, while Verbal + SE shows a similarly unstable, inverted trajectory across the query budget. This contrasts with the stable Uncertainty-First preference observed for SC + SE, suggesting that ordering effects only stabilize once the underlying uncertainty signal itself is reliable; when the proxy is noisy, its interaction with cluster-based diversity filtering becomes unpredictable across acquisition stages.

B.3 Summarisation

This section provides the supporting detail underlying the key findings summarized in section 5.

Hybrid configurations Pipeline ordering All Uncertainty-First and Diversity-First hybrid permutations remain at or below the random baseline throughout the acquisition budget, with no configuration exhibiting the mid-to-late-stage divergence that characterizes the rescue effect observed for SE + SC on CommonSenseQA. This absence suggests that the rescue effect requires at least one component metric (e.g., SC) to carry a reliable signal for the hybrid to exploit.

The heatmap in fig. 4c shows near-uniformly small magnitude differences between Uncertainty-First and Diversity-First orderings across all combination strategies and acquisition stages, in sharp contrast to the pronounced, stage-dependent structure observed for CommonSenseQA and SQuAD_v2. This indicates that pipeline ordering becomes structurally inert once both uncertainty and diversity signals fail to capture task-relevant information.