Selecting Demonstrations for Many-Shot In-Context Learning via Gradient Matching

Jianfei Zhang[1], Bei Li[2], Jun Bai[4], Rumei Li[1,2], Yanmeng Wang[3], Chenghua Lin[5], Wenge Rong[1]
[1]School of Computer Science and Engineering, Beihang University, China
[2]Meituan, Inc., China [3]Ping An Technology, China
[4]Beijing Institute for General Artificial Intelligence, China
[5]Department of Computer Science, The University of Manchester, UK
{zhangjf, lirumei3232, w.rong}@buaa.edu.cn, libei17@meituan.com
baijun@bigai.ai, wangyanmeng219@pingan.com.cn, chenghua.lin@manchester.ac.uk

Abstract

In-Context Learning (ICL) empowers Large Language Models (LLMs) for rapid task adaptation without Fine-Tuning (FT), but its reliance on demonstration selection remains a critical challenge. While many-shot ICL shows promising performance through scaled demonstrations, the selection method for many-shot demonstrations remains limited to random selection in existing work. Since the conventional instance-level retrieval is not suitable for many-shot scenarios, we hypothesize that the data requirements for in-context learning and fine-tuning are analogous. To this end, we introduce a novel gradient matching approach that selects demonstrations by aligning fine-tuning gradients between the entire training set of the target task and the selected examples, so as to approach the learning effect on the entire training set within the selected examples. Through gradient matching on relatively small models, e.g., Qwen2.5-3B or Llama3-8B, our method consistently outperforms random selection on larger LLMs from 4-shot to 128-shot scenarios across 9 diverse datasets. For instance, it surpasses random selection by 4% on Qwen2.5-72B and Llama3-70B, and by around 2% on 5 closed-source LLMs. This work unlocks more reliable and effective many-shot ICL, paving the way for its broader application.

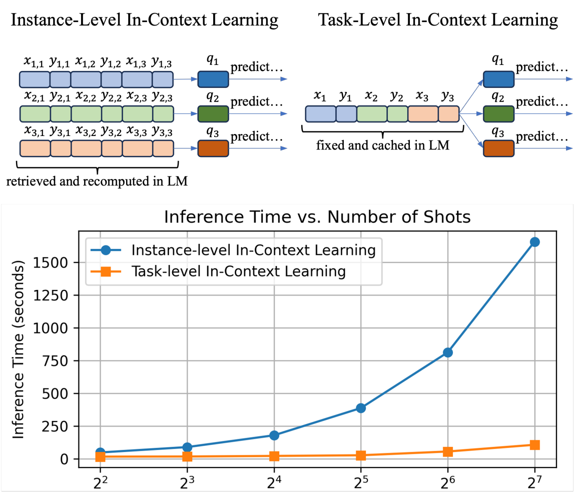

Figure 1: Instance-level retrieved demonstrations are not suitable for Many-Shot ICL, since the variable context cannot utilize prefix-caching and can lead to theoretically $O(n^2)$ inference complexity (Vaswani et al., 2017) when using n-shot demonstrations. Here we illustrate the runtime of Llama3-70B (Dubey et al., 2024) on CMSQA (Talmor et al., 2019) using vLLM with 4 A100 GPUs in tensor-parallel (Kwon et al., 2023).

1 Introduction

In-Context Learning (ICL) enables pre-trained Large Language Models (LLMs) to perform tasks by learning from input-output examples (or "*demonstrations*") provided during inference (Brown et al., 2020). This allows LLMs to adapt to new tasks through forward propagation, without weight updates via back-propagation, offering a flexible alternative to traditional fine-tuning.

Since the performance of ICL is often sensitive to the choice of demonstrations (Liu et al., 2022), significant efforts have been made to improve demonstration selection. Most studies concentrate on instance-level retrieval, aiming to identify suitable demonstrations for each test query independently (Luo et al., 2023; Rubin et al., 2022). This mainly involves considering similarity (Luo et al., 2023; Rubin et al., 2022) between demonstrations and the query, as well as auxiliary factors such as complexity (Fu et al., 2023), perplexity (Gonen et al., 2023), difficulty (Drozdov et al., 2023), and diversity (Li and Qiu, 2023). Another line of work focuses on task-level selection, which seeks to find a fixed set of demonstrations that achieves the best average performance across all test queries from a target task (Zhang et al., 2022; Wang et al., 2023). Current approaches primarily explore reinforcement learning on the selection policy (Zhang et al., 2022) and Bayesian inference to explain the demonstration effectiveness (Wang et al., 2023) as

Findings of the Association for Computational Linguistics: ACL 2025, pages 11686–11704
July 27 - August 1, 2025

potential solutions.

However, for the recently emerged Many-Shot in-context learning (Agarwal et al., 2024) paradigm, demonstrations are selected simply by random in existing work (Agarwal et al., 2024; Bertsch et al., 2024; Song et al., 2025; Jiang et al., 2024), which may lead to suboptimal performance. Nevertheless, existing methods for few-shot demonstration selection are not well-suited to many-shot scenarios. On one hand, applying instance-level demonstrations conflicts with caching and reusing the hidden states of the same long context in language models (Pope et al., 2023), leading to theoretically $O(n^2)$ inference complexity (Vaswani et al., 2017) for n-shot demonstrations, as illustrated in Fig. 1. On the other hand, existing task-level selection methods are designed for very few-shot scenarios (e.g., 4-shot in their implementations), facing challenges from exploration complexity (Zhang et al., 2022) and performance early saturation (Wang et al., 2023) for more shots. To further release the potential of many-shot ICL, it now requires a selection method with scalable effectiveness to many-shot scenarios.

To address such research gap, we revisit ICL demonstration selection from a "learning" perspective—we hypothesize that the *data requirements for in-context learning and fine-tuning are analogous*. Building on this premise, we conduct latent concept learning (Wang et al., 2023) on a small Language Model (LM) to learn the optimal in-context task embeddings. Then we compute the latent gradient that each example provides to the learning process. Finally, we select n-shot demonstrations from the whole fine-tuning set with the minimized L_2 distance between their average latent gradients, so as to approach the learning effect on all examples within the selected n-shot demonstrations.

We validate our proposed method on 9 datasets from 5 distinct NLP tasks, each covering scenarios from 4-shot to 128-shot. Our method selects demonstrations through gradient matching on relatively small models, e.g., Qwen2.5-3B or Llama3-8B, and surpasses the widely-adopted random selection by an average of 4% on Llama3-70B and Qwen2.5-72B. Furthermore, the selected demonstration set exhibits transferability to closed-source LLMs, consistently outperforming random selection by around 2% on Qwen-turbo, GLM-4-flash, Doubao-pro-32k, GPT-4o-mini, and DeepSeek-V3. Our source code is available at `https://github.com/zhangjf-nlp/ManyShotICL-CLG.git`.

2 Related Work

2.1 Many-Shot In-Context Learning

In-Context Learning (ICL) refers to the capability of LLMs to learn from data during inference through forward propagation, without the need for backward propagation or weight updates (Li et al., 2023b). Prior work is limited to few-shot scenarios (Jiang et al., 2024) by the context window size, e.g., 2048 tokens in GPT-3 (Brown et al., 2020).

Recently, advancements in expanding the context window size of LLMs (Ding et al., 2024; Chen et al., 2023), e.g., up to 128k tokens in GPT-4o (OpenAI, 2023) and Qwen2.5 (Yang et al., 2024), have enabled the exploration of many-shot ICL (Agarwal et al., 2024). These studies have observed that Many-Shot ICL, which includes up to hundreds or even thousands of demonstrations, can make substantial improvements compared to few-shot ICL on various tasks (Agarwal et al., 2024; Bertsch et al., 2024). In this paradigm, the benefits of instance-level retrieval over using a fixed random set of demonstrations tend to diminish (Bertsch et al., 2024), while a fixed demonstration set can largely reduce the inference cost through prefix-caching (Agarwal et al., 2024). Therefore, researchers tend to randomly select a demonstration set and reuse it across all queries in many-shot ICL (Agarwal et al., 2024; Bertsch et al., 2024; Jiang et al., 2024; Song et al., 2025).

While random selection is widely adopted for many-shot ICL, it is underexplored whether this strategy is truly optimal. This motivates our work to explore and develop a better demonstration selection strategy for the many-shot paradigm.

2.2 Demonstration Selection

The performance of ICL is sensitive to the choice of demonstrations (Liu et al., 2022; Perez et al., 2021), prompting efforts in demonstration selection. Existing approaches fall into two categories:

Instance-level retrieval selects demonstrations for each query using semantic similarity, such as cosine similarity (Rubin et al., 2022) and BM25 (Li et al., 2023a). Some work further emphasizes additional factors including diversity (Li and Qiu, 2023), complexity (Fu et al., 2023), difficulty (Drozdov et al., 2023), and perplexity (Gonen et al., 2023). While providing relevant and useful information, instance-level retrieval prevents the use of prefix-caching during inference and leads to the-

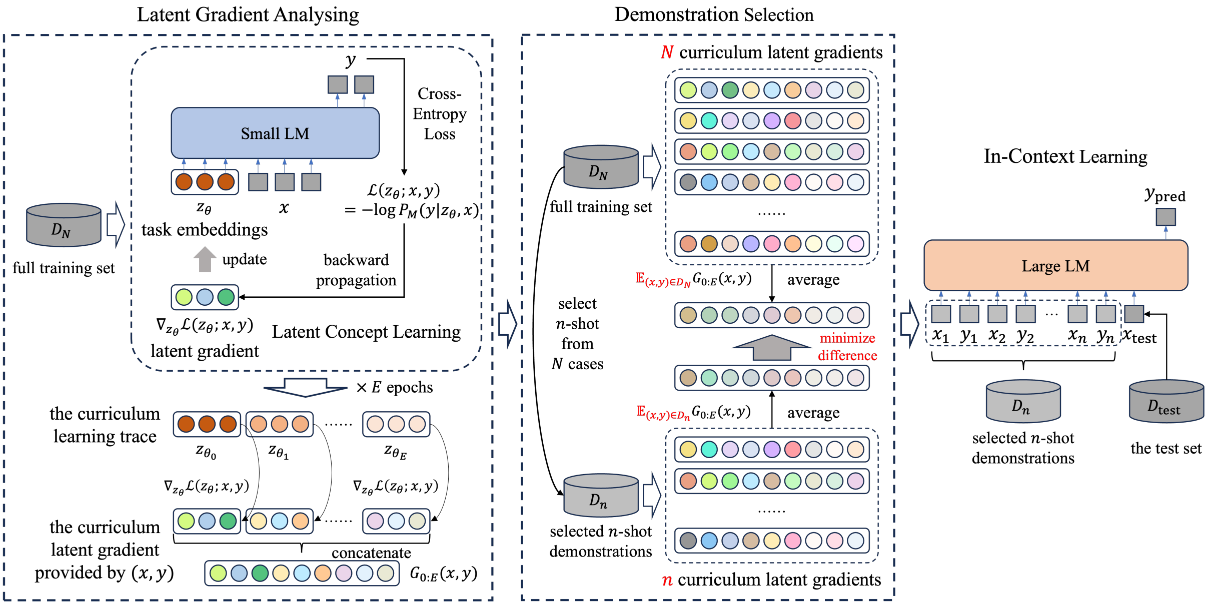

Figure 2: The overview of our proposed method for many-shot demonstration selection. We perform latent concept learning (Wang et al., 2023) on a small LM and compute the Curriculum Latent Gradient (CLG) provided by each example throughout the E-epoch training. Then we find an n-shot demonstration set to mimic the average gradient on the entire training set. We show that the selected n-shot demonstrations can effectively improve the many-shot In-Context Learning performances on both same-series larger LMs and various closed-source LLMs.

oretically $O(n^2)$ inference complexity (Vaswani et al., 2017) for n-shot demonstrations. In contrast, the runtime of many-shot ICL under a fixed demonstration set increases only linearly with a large number of shots (Agarwal et al., 2024). This inefficiency makes instance-level retrieval less practical as the number of demonstration shots increases.

Task-level selection aims to identify a fixed set of demonstrations that achieve optimal average performance across all queries from a target task. For example, Zhang et al. (2022) use Reinforcement Learning (RL) to gradually explore the optimal demonstration set. Wang et al. (2023) view LLMs as latent variable models and select demonstrations that can best infer the optimal latent task concept. However, these methods are both limited to few-shot scenarios (e.g., 4-shot in their implementations), facing challenges from complexity in exploring potential many-shot demonstration sets (Zhang et al., 2022) and performance early saturation beyond 4-shot (Wang et al., 2023) respectively.

In summary, current demonstration selection methods are fundamentally designed for and evaluated within few-shot scenarios, facing challenges in generalizing to many-shot scenarios. To the best of our knowledge, our work presents the first attempt to enhance demonstration selection for many-shot ICL, beyond random selection in existing work.

3 Methodology

Previous research has identified several key attributes that contribute to the effectiveness of in-context demonstrations, such as similarity (Luo et al., 2023), diversity (Li and Qiu, 2023), and coverage (Li and Qiu, 2023). While intuitively beneficial, these attributes lack a direct connection to the *learning dynamics* induced by demonstrations within the language model. We posit that the ideal demonstration set for ICL should not only present static attributes of high-quality fine-tuning data but, more fundamentally, should actively facilitate and guide the language model's learning process on the target task. Drawing inspiration from the principles of effective training data selection in supervised learning, particularly from the field of *dataset condensation* (Zhao et al., 2021), we introduce Curriculum Latent Gradient (CLG), a novel approach for task-level in-context demonstration selection. CLG leverages the concept of *latent task embeddings* and analyses their *learning trajectory* to identify demonstrations that can effectively guide the model's learning process.

We begin by reviewing Latent Concept Learning (Sec. 3.1), highlighting its strengths and limitations as a foundation. We then articulate the core principles behind Curriculum Latent Gradient (Sec. 3.2), detailing how we capture and leverage the learning dynamics of latent task embeddings to guide demonstration selection. Finally, we describe the demonstration selection process based on these learning dynamics (Sec. 3.3). An overview of our proposed methodology is depicted in Fig. 2.

3.1 Preliminary: Latent Concept Learning

Latent Concept Learning Wang et al. (2023) aims to learn an optimal latent task concept z_θ as a proxy of in-context demonstrations. Specifically, prediction from a language model P_M through in-context learning over demonstrations $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ can be expressed in Eq. 1,

$$P(y \mid x; D_n) = P_M(y \mid x_1, y_1, \ldots, x_n, y_n, x) \tag{1}$$

where D_n contains the n-shot demonstrations, and faces optimization challenges for its discrete nature. To address this, they replace the demonstration set with continuous and optimizable embeddings of the *latent task concept* z_θ, as expressed in Eq. 2,

$$P(y \mid x; z_\theta) = P_M(y \mid z_\theta, x) \tag{2}$$

where z_θ can be optimized through maximum likelihood estimation on the entire training set D_N, as formulated in Eq. 3.

$$\tilde{z}_\theta = \arg\max_{z_\theta} \sum_{(x,y)\in D_N} \log P(y \mid x; z_\theta) \tag{3}$$

Latent-Bayesian subsequently utilizes the optimized $\tilde{z}_\theta$ to select task-level demonstrations. Under a series of simplifying assumptions, it finally selects the top-n demonstrations with the highest posterior probabilities of $\tilde{z}_\theta$ in language model P_M, as expressed in Eq. 4, where the top-n operator selects the n samples with highest function values.

$$D_n = \underset{(x,y)\in D_N}{\text{top-}n} \; P_M(z_\theta \mid x, y) \tag{4}$$

While Latent-Bayesian offers an interpretable approach under specific assumptions, its reliance on static posterior probabilities and independence assumptions limits its effectiveness, particularly for many-shot ICL. Crucially, it overlooks the *dynamic interactions* between demonstrations – such as redundancy or synergistic effects – which become increasingly vital as the number of demonstrations grows. This motivates our departure from static, result-oriented approaches towards a method that explicitly considers the *learning process* itself.

3.2 Curriculum Latent Gradient

Our method borrows the idea of latent concept learning, i.e., learning task concept embeddings that play the role of context. Notably, we focus on the learning process instead of the learnt result. We quantify the learning dynamics under different demonstrations through the optimizing gradients on the task concept embeddings provided by each demonstration, throughout the training process.

Specifically, given a small language model P_M and a target task, we construct k prefix token embeddings $z_\theta \in R^{k\times h}$ to represent the latent task concept, where h denotes the embedding size. These embeddings are then trained on the target task with the negative log-likelihood loss in Eq. 5.

$$\mathcal{L}(z_\theta; x, y) = -\log P_M(y|z_\theta, x) \tag{5}$$

We randomly initialize the latent concept tokens $z_\theta = z_{\theta_0}$, and employ Stochastic Gradient Descent (Bottou, 2010) to optimize z_θ over the full training set D_N for E epochs. Crucially, we save the learnt latent concept at the end of each epoch, denoted as $z_{\theta_1}, z_{\theta_2}, \ldots, z_{\theta_E}$. For each training example $(x, y) \in D_N$, we calculate its curriculum latent gradient, $G_{0:E}(x, y) \in R^{k\times h\times(E+1)}$, which is a concatenation of the gradients of the loss function with respect to z_θ, evaluated both at the initial point $z_\theta = z_{\theta_0}$ and at the end of each epoch $z_\theta = z_{\theta_i}$ for $1 \le i \le E$. This is formulated in Eq. 6, where $[;]$ denotes the concatenation operation.

$$\begin{aligned} G_{0:E}(x, y) = \big[&\nabla_{z_\theta}\mathcal{L}(z_\theta; x, y)\big|_{z_\theta=z_{\theta_0}};\\ &\nabla_{z_\theta}\mathcal{L}(z_\theta; x, y)\big|_{z_\theta=z_{\theta_1}};\\ &\ldots;\\ &\nabla_{z_\theta}\mathcal{L}(z_\theta; x, y)\big|_{z_\theta=z_{\theta_E}}\big] \end{aligned} \tag{6}$$

3.3 Demonstration Selection via Gradient Matching

To select an effective n-shot demonstration set for ICL, we aim to identify an optimal n-shot subset $D_n^* \subseteq D_N$ that induces similar learning dynamics in the LM as the full training set D_N. Specifically, we propose to achieve this by minimizing the L_2 distance between the average curriculum latent gradient over D_N and that over the n-shot subset D_n, as formalized in Eq. 7, ensuring that the model's learning behaviour on the subset D_n, as summarized in $G_{0:E}(D_n)$, closely aligns with that on the

Algorithm 1 Gradient Matching through Greedy Search and Local Optimization

Input: the entire training set D_N, the curriculum latent gradient $G_{0:E}(x,y)$ for each example in the training set $(x,y) \in D_N$, the number of shots n, the maximum iteration steps l for local optimization.
Output: the n-shot demonstrations D_n^* that approximately minimize the gradient matching objective in Eq. 7.

```
Compute the average curriculum latent gradient on D_N:
    G_{0:E}(D_N) = E_{(x,y)∈D_N} [G_{0:E}(x,y)].
Initialize D_0 = ∅.
for i = 1 to n do
    Find the next example (x_j, y_j) ∈ (D_N \ D_{i-1}) that
    minimizes the distance between the average curriculum
    latent gradient on D_N and that on D_{i-1} ∪ {(x_j, y_j)}.
    Set D_i = D_{i-1} ∪ {(x_j, y_j)}.
end for
Set d_old = ||G_{0:E}(D_N) - E_{(x,y)∈D_n} G_{0:E}(x,y)||_2.
for i = 1 to l do
    Find (x_j, y_j) ∈ D_N \ D_n and (x_k, y_k) ∈ D_n to
    minimize the distance between the average curriculum
    latent gradient on D_N and that on D_n ∪ {(x_j, y_j)} \
    {(x_k, y_k)}.
    Set D_new = D_n ∪ {(x_j, y_j)} \ {(x_k, y_k)}.
    Set d_new = ||G_{0:E}(D_N) - E_{(x,y)∈D_new} G_{0:E}(x,y)||_2.
    if d_new ≥ d_old then
        break
    else
        Set D_n = D_new and d_old = d_new.
    end if
end for
return D_n as the optimized n-shot demonstrations D_n^*.
```

complete set D_N, as summarized in $G_{0:E}(D_N)$.

$$
\begin{aligned}
G_{0:E}(D_N) &= \mathbb{E}_{(x,y)\in D_N}\left[G_{0:E}(x,y)\right] \\
G_{0:E}(D_n) &= \mathbb{E}_{(x,y)\in D_n}\left[G_{0:E}(x,y)\right] \\
D_n^* &= \underset{D_n \subset D_N}{\arg\min} \left\|G_{0:E}(D_N) - G_{0:E}(D_n)\right\|_2
\end{aligned} \tag{7}
$$

Since the optimization problem in Eq. 7 is indeed an NP-Complete Subset Sum Problem (SSP) (Lagarias and Odlyzko, 1985), we resort to heuristic approximation to find a near-optimal solution for D_n^*, involving two key phases: *Greedy Search* and *Local Optimization*. In the greedy phase, we incrementally construct the n-shot subset D_n by iteratively adding the example that minimizes the target L_2 distance. Then the local optimization phase refines the n-shot subset through up to $l = 32$ iterations to further reduce the target L_2 distance, by replacing a selected example with an unselected one. The detailed steps of the selection algorithm are presented in Algorithm 1, which requires only a few minutes to execute in practice.

	designed for task-level	extended from instance-level
learning-free	Random Best-of-N	BM25-Major BGE-KMeans
learning-based	Latent-Bayesian CLG (ours)	EPR-KMeans

Table 1: Types of Methods

4 Experiments

We conduct experiments on 9 datasets from 5 distinct tasks, and show that demonstrations selected by our proposed method can effectively improve the many-shot ICL performances over random selection as well as some straightforward methods.

4.1 Baselines

In addition to random selection, the commonly-adopted method in existing work of many-shot ICL, we examine various straightforward methods for demonstration selection, including those designed for task-level selection and those extended from instance-level retrieval. We classify these methods in Table 1 and introduce them in details below.

Random: The most basic demonstration selection method, which is widely-adopted in existing work for many-shot ICL (Agarwal et al., 2024; Bertsch et al., 2024; Song et al., 2025; Jiang et al., 2024). We implement this across 5 random seeds and report the mean and standard deviation.

BM25-Major: BM25 (Robertson and Zaragoza, 2009) is a popular term-based scoring method for instance-level retrieval (Li et al., 2023a). We extend it to task-level demonstration selection through Majority Voting, i.e., selecting n-shot demonstrations with the n highest average scores to be retrieved by the other examples on the training set.

BGE-KMeans: BGE-M3 (Chen et al., 2024) provides off-the-shelf sentence embeddings with state-of-the-art performances on multiple retrieval tasks. We extend it to task-level selection through KMeans clustering, as implemented in the Scikit-Learn library, over its text embeddings.

EPR-KMeans: EPR (Rubin et al., 2022) learns to retrieve demonstrations through contrastive learning, supervised by the ICL likelihoods on a relatively small LM. We train EPR retrievers and apply them to task-level selection through KMeans clustering over their learnt dense embeddings.

Type	Dataset	Task	#Train	#Validation	Avg. Tokens	Metric
Classification	SST-5 (Socher et al., 2013)	Sentiment Analysis	8,544	1,101	26.55	Acc
	MNLI (Williams et al., 2018)	Natural Language Inference	50,000	10,000	43.21	Acc
	CMSQA (Talmor et al., 2019)	Commonsense Reasoning	9,741	1,221	45.70	Acc
	HellaSwag (Zellers et al., 2019)	Commonsense Reasoning	50,000	10,000	79.42	Acc
Open-ended	GeoQuery (Shaw et al., 2021)	Code Generation	600	280	22.66	EM
	NL2Bash (Lin et al., 2018)	Code Generation	8,090	609	32.56	BLEU
	Break (Wolfson et al., 2020)	Semantic Parsing	44,321	7,760	61.44	LF-EM
	MTOP (Li et al., 2021)	Semantic Parsing	15,667	2,235	34.12	EM
	SMCalFlow (Andreas et al., 2020)	Semantic Parsing	50,000	10,000	53.78	EM

Table 2: Datasets used in our experiments. We use at most 50,000 training instances and 10,000 validation instances. Following previous work (Ye et al., 2023), we use Accuracy (Acc) on classification tasks, and use Exact Match (EM), Logical Form EM (LF-EM) (Hasson and Berant, 2021), and character-level BLEU on open-ended tasks.

Best-of-N: Best-of-N is a commonly used baseline for Reinforcement Learning (RL), and demonstrates competitive performance to RL-based task-level demonstration selection (Zhang et al., 2022). We implement this through randomly selecting N=5 demonstration sets, evaluating them on a relatively small LM over all training instances, and selecting the best-performing one.

Latent-Bayesian: The task-level demonstration selection method based on Bayesian inference (Wang et al., 2023), as we introduce in Sec 3.1.

4.2 Datasets and Evaluation

We list all the datasets in Table 2, paired with their tasks, sizes, average tokens, and evaluation metrics. We illustrate cases on each dataset in Appendix A. We select demonstrations from the training set and report their ICL performances on the validation set, since the test set is private for some datasets.

4.3 Implementation Details

For LM-based approaches, we utilize Llama3-8B to select demonstrations for Llama3 series LLMs (Dubey et al., 2024), and utilize Qwen2.5-3B to select demonstrations for Qwen2.5 series LLMs (Yang et al., 2024) and closed-source LLMs. We train EPR retrievers from BERT-base-uncased (Devlin et al., 2019) under the supervision of LMs, using a learning rate of $1e-5$ and a batch size of 8, for at most 120 epochs until regression. We conduct latent concept learning on LMs for Latent-Bayesian and CLG, with a learning rate of $1e-3$ and a batch size of 64 for $E = 10$ epochs.

For ICL performance evaluation, we utilize the vLLM framework (Kwon et al., 2023) to perform greedy search, involving 4-shot, 8-shot, 16-shot, 32-shot, 64-shot, and 128-shot settings.[1] On classification tasks, we only allow the model to generate tokens contained in the options.

4.4 Main Results

We illustrate the results of 128-shot ICL on Llama3-70B and Qwen2.5-72B in Table 3. It can be observed that, demonstrations selected by random are in fact generally suboptimal for task-level in-context learning. Among all methods, our proposed CLG performs the best on average, surpassing that of random selection by 4% on both models.

We further evaluate the selected demonstrations by Random, Best-of-N, and CLG on 5 closed-source models: Qwen-turbo, GLM-4-flash, Doubao-pro, GPT-4o-mini, and DeepSeek-V3.[2] We include at most 300 instances from each dataset and constrain the context length to within 7000 tokens due to budget limitations. The average performances across all datasets are illustrated in Table 4, where CLG consistently outperforms Random and Best-of-N on all closed-source models. This demonstrates the transferability of demonstrations selected by our CLG to different series of LLMs, as well as the effectiveness of CLG on closed-source LLMs.

4.5 Overall Scaling Performance

In Fig. 3, we examine the scaling trends of ICL performance w.r.t. the model size and the shot number. For simplicity, we present the average performance scores over 9 datasets. The results reveal a clear

[1]Since the context window sizes of Llama3-8B and Llama3-70B are limited to 8192, some of the 128-shot demonstrations may be truncated to fewer shots, ensuring at most 7000 prompt tokens (the demonstrations and the test query).

[2] https://www.alibabacloud.com/help/en/model-studio/getting-started/models, https://bigmodel.cn/dev/howuse/model, https://www.volcengine.com/product/doubao, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence, and https://api-docs.deepseek.com/news/news1226.

Method	SST-5	MNLI	CMSQA	HeSwag.	GeoQ.	NL2Bash	Break	MTOP	SMCal.	Average
Llama3-70B										
Random	43.71	61.83	**84.73**	76.11	73.36	29.72	35.55	44.66	36.68	54.04
BM25-Major	43.69	51.36	82.23	72.01	43.21	27.92	15.05	8.41	14.38	39.81
BGE-KMeans	46.41	65.09	83.78	73.90	**84.64**	28.39	35.13	44.79	38.73	55.65
Best-of-N	43.69	70.92	84.19	74.57	80.71	**29.76**	35.86	43.40	37.66	55.64
EPR-KMeans	43.05	49.78	83.62	71.32	78.21	22.92	**38.39**	**50.43**	38.10	52.87
Latent-Bayesian	46.59	69.84	84.60	70.09	62.14	29.20	32.59	41.21	23.13	51.04
CLG (ours)	**48.32**	**76.37**	84.52	**80.77**	**84.64**	29.59	37.33	47.74	**40.35**	**58.85**
Qwen2.5-72B										
Random	36.00	58.20	87.47	86.58	57.29	33.28	36.67	45.32	38.23	53.23
BM25-Major	37.78	45.94	86.65	85.61	52.50	38.11	22.81	16.24	11.57	44.13
BGE-KMeans	37.24	49.65	87.71	86.67	61.79	40.86	35.84	43.71	**41.29**	53.86
Best-of-N	36.60	60.57	87.55	86.83	61.07	37.13	35.73	46.89	39.22	54.62
EPR-KMeans	36.15	58.41	**87.96**	86.22	**62.86**	26.53	**39.12**	**47.53**	39.65	53.83
Latent-Bayesian	27.25	47.96	85.75	86.72	27.50	37.36	35.35	5.23	22.61	41.75
CLG (ours)	**38.33**	**77.29**	87.71	**87.68**	62.50	**45.19**	37.07	47.07	40.16	**58.11**

Table 3: Main results (in %) of Llama3-70B and Qwen2.5-72B under the 128-shot setting. The best scores on each model are in bold, and the second-best ones are underlined. Full results are illustrated in Appendix C. BM25-Major and Latent-Bayesian generally perform worse than Random selection; we analyse the reasons in Appendix B.

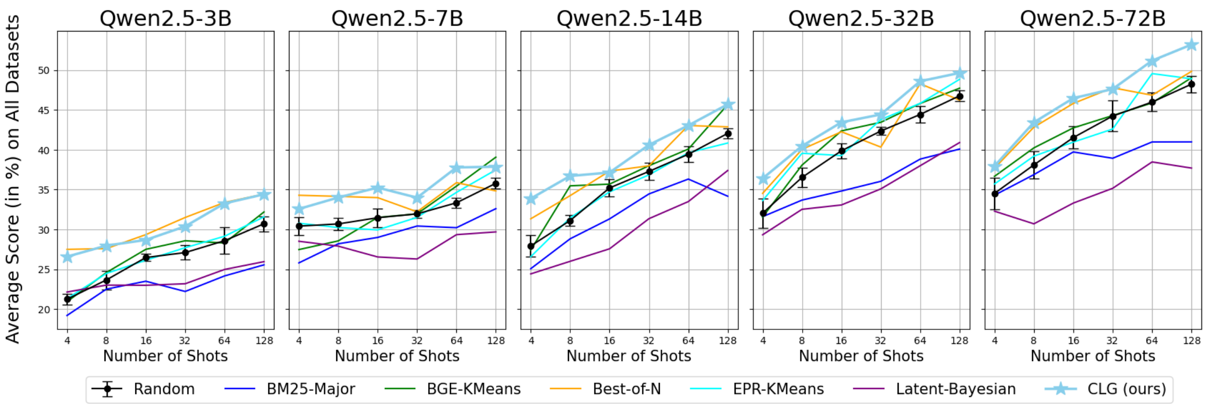

Figure 3: The scaling performances with respect to (w.r.t.) the number of shots and the model size. For random selection under 5 different random seeds, standard deviations are illustrated through error bars.

Method	Qwen.	GLM4	DouB.	GPT4o	DeepS.
Random	60.08	56.17	62.78	60.22	66.83
±std	±0.22	±2.24	±0.26	±0.45	±0.52
Best-of-N	59.63	57.65	61.69	60.38	66.79
CLG (ours)	**61.87**	**58.50**	**64.09**	**61.48**	**68.08**

Table 4: ICL performances of 128-shot demonstrations, average over 9 datasets, on Qwen-turbo, GLM-4-flash, Doubao-pro, GPT-4o-mini, and DeepSeek-V3. Best scores are in bold. More details are in Appendix D.

positive correlation between the overall ICL performance and the model size as well as the shot number, indicating that larger models and more shots generally lead to better performances.

Among the methods compared, CLG generally outperforms other methods across various conditions, achieving significant improvements (far more than the standard deviation) over random selection. In contrast, the most competitive Best-of-N baseline exhibits relatively lower scaling efficiency with the increased model size or the increased shot numbers. This further highlights the scalability of CLG with respect to the model size and the shot number.

4.6 Scaling to a Thousand Shots

To further compare CLG against Random selection in long-context settings with enough demonstrations, we select 1024-shot demonstrations through both methods, and evaluate their performances on MNLI, HellaSwag, Break, and SMCalFlow, the four datasets with the most examples.

As illustrated in Table 5, CLG consistently achieves better results than the average performance of Random selection on all datasets. Such performance gains on 1024-shot are generally smaller than those on 128-shot. This could be explained by the possibility that 1024-shot demonstrations provide sufficient task knowledge needed for ICL, even when selected randomly.

Besides, we found that Break and SMCalFlow

Method	MNLI	HeSwag.	Break	SMCal.
128-shot				
Random	61.57	86.58	36.67	38.23
±std	±1.12	±0.33	±0.60	±1.42
CLG (ours)	**77.29**	**87.68**	37.07	40.16
Δ	+15.72	+1.10	+0.40	+1.93
1024-shot				
Random	55.08	77.64	42.08	52.67
±std	±1.75	±1.83	±0.28	±0.35
CLG (ours)	58.94	79.12	**42.21**	**52.82**
Δ	+3.86	+1.48	+0.13	+0.15

Table 5: Comparison of 128-shot and 1024-shot ICL on Qwen2.5-72B (in %). Best scores are in bold.

show significant improvements when scaled to 1024-shot, while MNLI and HellaSwag exhibit some degree of performance degradation. To explain this phenomenon, we analyse the scaling trends of ICL on each dataset with increasing shot numbers in Appendix E. We find that open-ended tasks, such as Break and SMCalFlow, are more likely to benefit from a larger number of shots.

4.7 Ablation Study: Gradient Mismatching

We conduct an ablation study to assess the effectiveness of the gradient matching term in Eq. 7. Specifically, we reverse the optimization target: instead of minimizing the L_2 distance in Eq. 7, we maximize it to select demonstrations that misalign with the full training set in learning dynamics.

As illustrated in Fig. 4, this reversal results in significant degradation in the ICL performance, which can be much poorer than that of random selection. These results empirically validate that the quality of demonstrations is correlated with the degree of curriculum latent gradient matching.

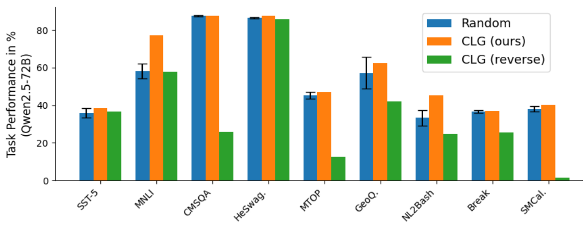

Figure 4: Ablation study on the impact of gradient mismatching on ICL performance, using 128-shot demonstrations with Qwen2.5-72B.

4.8 Improving Adaptability through Hybrid In-Context Learning

In addition to our main experiments on task-level many-shot in-context learning, we further explore the complementary use of task-level and instance-level demonstrations. Specifically, we combine the 128-shot task-level demonstrations selected by Random and CLG with the 4-shot instance-level demonstrations retrieved by BGE-M3.

As illustrated in Fig. 5, such hybrid strategy yields further improvements over using only task-level or instance-level demonstrations across most datasets. Meanwhile, CLG generally outperforms Random selection in such hybrid mode.

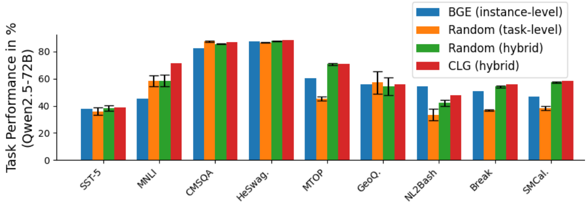

Figure 5: ICL with hybrid demonstrations: task-level many-shot followed by instance-level few-shot.

4.9 Correlation between FT and ICL

To validate our hypothesis that the data requirements for In-Context Learning (ICL) and Fine-Tuning (FT) are comparable, we fine-tune Qwen2.5-3B on the 128-shot demonstrations selected by each method. We employ prefix-tuning (Li and Liang, 2021) to avoid overfitting, and perform training with a batch size of 128 and a learning rate of $3e-4$ for 1000 epochs. We quantify the FT performances through the negative evaluation losses on the validation set. We analyse and illustrate the correlation between FT and ICL performance in Fig. 6, which shows a positive relationship between the two forms of machine learning, especially for open-ended tasks.

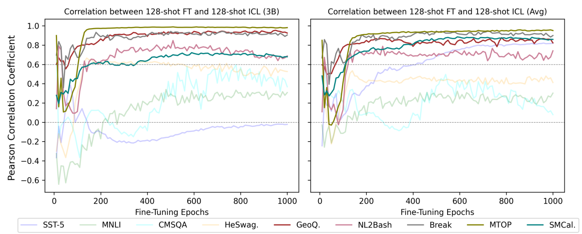

Figure 6: The correlation between FT and ICL. The left involves ICL performances on Qwen2.5-3B, and the right involves that averaged over all Qwen2.5 models. Datasets for open-ended tasks are lined in solid colors.

4.10 Diversity and Coverage Analysis

Since prior work has illustrated that diversity and coverage are important for effective in-context demonstrations (Li and Qiu, 2023), we examine whether our method inherently maintains diversity and coverage in selected demonstrations. Specifically, we analyze the label distributions of demonstration sets on classification datasets. We measure the KL divergence between the label distribution of selected demonstrations and that of the test set, with results shown in Table 6.

Method	SST-5	MNLI	CMSQA	HeSwag.
Random	0.020	0.016	0.016	0.012
$\pm$ std	$\pm$0.016	$\pm$0.009	$\pm$0.008	$\pm$0.007
Best-of-N	0.019	0.021	0.041	0.019
BGE-KMeans	0.017	0.036	0.010	0.026
Latent-Bayesian	4.050	0.115	0.134	4.296
CLG (ours)	**0.006**	**0.002**	**0.007**	**0.008**

Table 6: KL divergence between demonstration set label distribution and test set label distribution (lower is better). CLG achieves the best alignment with test set distributions, indicating it inherently maintains diversity and coverage of labels on classification datasets.

The results demonstrate that CLG naturally selects a demonstration set whose label distribution closely matches the test set distribution, significantly outperforming random selection. This suggests that gradient matching can inherently preserve data coverage in terms of label. In contrast, Latent-Bayesian shows particularly poor performance in maintaining label diversity, which may contribute to its suboptimal ICL performance.

4.11 Efficiency Analysis

To further verify the cost-effectiveness of our proposed CLG in practice, we analyse the computational cost of the complete demonstration selection pipeline. Table 7 details the time consumption for selecting 128-shot demonstrations from MNLI and SMCalFlow using Qwen2.5-3B, measured on 8 A100 GPUs with data-parallel implementation via `deepspeed` and `accelerate` libraries.

CLG Stage	# GPUs	MNLI	SMCal.
Prefix Tuning	8	80 min	87 min
Gradient Computation	8	145 min	210 min
Gradient Matching	1	6 min	10 min

Table 7: Time consumption for 128-shot demonstration selection from datasets with 50k training examples.

The results demonstrate that our approach maintains practical efficiency even at scale, with total computation time being approximately 3~4 times that of prefix-tuning, an already efficient fine-tuning method. This computational investment is well justified as our method enables significant ICL performance improvements on both open-source larger models and closed-source LLMs, for which fine-tuning may face challenge in computational resources and access to model weights. This validates the cost-effectiveness of our proposed CLG for many-shot ICL in practice.

5 Conclusion

Many-shot ICL has emerged as a promising way to utilize LLMs on downstream tasks, but the random selection applied in existing work may produce sub-optimal demonstrations and learning results.

In this work, we hypothesize that data requirements for in-context learning and fine-tuning are analogous, and propose Curriculum Latent Gradient (CLG) to select a demonstration set that aligns with the entire training set in learning dynamics on LMs. We validate our method across various datasets and LLMs, showing its significant improvements compared to random selection, e.g., improving by 4% on open-source LLMs and by 2% on closed-source LLMs.

This work unlocks more reliable and effective many-shot ICL, paving the way for its broader adoption and application in real-world scenarios.

Limitations

This work focuses on improving many-shot in-context learning (ICL) by optimizing the demonstration set. However, ICL performance can also be sensitive to the order of demonstrations (Lu et al., 2022), which our method does not address. We believe that investigating the impact of demonstration order on in-context learning could lead to further improvements in the learning performance, for example, by examining the learning dynamics at different stages of fine-tuning.

Besides, our method selects demonstrations according to analysis on the training set, which may suffer from biases in the training data, such as imbalances over gender, race, or culture. This could lead to discriminatory content in practice. Furthermore, improved many-shot ICL may be susceptible to malicious exploitation, resulting in misuse such as jailbreaking LLMs (Anil et al., 2024).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62477001.

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Stephanie Chan, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal M. P. Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. *CoRR*, abs/2404.11018.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Andrew Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Trans. Assoc. Comput. Linguistics*, 8:556–571.

Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel J Ford, et al. 2024. Many-shot jailbreaking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *CoRR*, abs/2405.00200.

Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *19th International Conference on Computational Statistics, COMPSTAT 2010, Paris, France, August 22-27, 2010 - Keynote, Invited and Contributed Papers*, pages 177–186. Physica-Verlag.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *CoRR*, abs/2402.03216.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *CoRR*, abs/2306.15595.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. Longrope: Extending LLM context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, Mohit Iyyer, Andrew McCallum, Donald Metzler, and Kai Hui. 2023. Parade: Passage ranking using demonstrations with large language models. *CoRR*, abs/2310.14408.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10136–10148. Association for Computational Linguistics.

Matan Hasson and Jonathan Berant. 2021. Question decomposition with dependency graphs. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H. Chen, and Andrew Y. Ng. 2024. Many-shot in-context learning in multimodal foundation models. *CoRR*, abs/2405.09798.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

J. C. Lagarias and Andrew M. Odlyzko. 1985. Solving low-density subset sum problems. *J. ACM*, 32(1):229–246.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2950–2962. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023a. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4644–4668. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding supporting examples for in-context learning. *CoRR*, abs/2302.13539.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023b. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore.

Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D. Ernst. 2018. Nl2bash: A corpus and semantic parser for natural language interface to the linux operating system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Seyed Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y. Zhao. 2023. Dr.icl: Demonstration-retrieved in-context learning. *CoRR*, abs/2305.14128.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. 2023. Efficiently scaling transformer inference. In *Proceedings of the Sixth Conference on Machine Learning and Systems, MLSys 2023, Miami, FL, USA, June 4-8, 2023*. mlsys.org.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 922–938. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. 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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Mingyang Song, Mao Zheng, and Xuan Luo. 2025. Can many-shot in-context learning help llms as evaluators? A preliminary empirical study. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 8232–8241. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Tomer Wolfson, Mor Geva, Ankit Gupta, Yoav Goldberg, Matt Gardner, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Trans. Assoc. Comput. Linguistics*, 8:183–198.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9134–9148. Association for Computational Linguistics.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2021. Dataset condensation with gradient matching. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

A Dataset Examples

We illustrate examples on all dataset in Table 8. We illustrate a query prompt with random 4-shot demonstrations on each dataset, where the demonstration answers are highlighted in red.

B Case Study: Poor Performance of BM25-Major and Latent-Bayesian

Since BM25-Major and Latent-Bayesian perform poorly, even underperforming Random selection, we examine the demonstrations selected by these two methods. We found they tend to be trapped in specific patterns and lose diversity in selection. We illustrate their demonstration cases and discuss these phenomena in Tables 9 and 10 respectively.

C Full Results

We illustrate all results on Qwen2.5 series models and Llama3 series models for 128-shot scenarios in Table 11, and results averaged over 4-shot to 128-shot scenarios in Table 12.

D Implementation Details and Full Results on Closed-Source LLMs

We evaluate demonstrations selected by Random, Best-of-N, and CLG over 5 closed-source LLMs: Qwen-turbo, GLM-4-flash, Doubao-pro-32k, GPT-4o-mini, and DeepSeek-V3. These models are primarily designed for chat-based interactions with users rather than text completion. To adapt them for our evaluation, we begin by presenting them with a system message that outlines the task requirements:

```
You are an In-Context Learner to learn from task
     demonstrations provided in context.
User will present you a prompt, composed of several
     demonstrations (query + response) and a test-
     query (only query).
You need to under stand the task concept and
     response format through the demonstrations, and
      return your response to the test query.

Example of a valid JSON response:
‘‘‘json
{
    "response": "...",
}‘‘‘
```

Next, we construct a user message containing the demonstrations (limited to 7000 tokens) and the test query, formatted as a JSON object. An example with 4-shot demonstrations is shown below:

```
‘‘‘json
{
    "demonstrations": [
        {
            "query": "as giddy and whimsical and
                 relevant today as it was 270 years
                 ago . It is ",
            "response": "great"
        },
        {
            "query": "after an uncertain start ,
                 murder hits and generally sustains
                 a higher plateau with bullock 's
                 memorable first interrogation of
                 gosling . It is ",
            "response": "good"
        },
        {
            "query": "an unbelievably stupid film ,
                 though occasionally fun enough to
                 make you forget its absurdity . It
                 is ",
            "response": "bad"
        },
        {
            "query": "even though it is infused with
                  the sensibility of a video
                 director , it does n't make for
                 completely empty entertainment It
                 is ",
            "response": "OK"
        }
    ],
    "test-query": {
        "query": "in his first stab at the form ,
             jacquot takes a slightly anarchic
             approach that works only sporadically .
              It is "
    }
}
‘‘‘
```

After chatting with closed-source models by these messages, we extract and evaluate responses from their JSON-formatted outputs. The full evaluation results are illustrated in Table 13.

E Scaling Trends of Random Selection on All Datasets

In Figure 7, we present the ICL performances of all models across various demonstration sizes, ranging from 4-shot to 128-shot, using randomly selected demonstrations. The results indicate that datasets for open-ended tasks tend to consistently benefit from more in-context demonstrations. This trend is likely attributable to the fact that open-ended tasks have significantly longer answers (see Appendix A), which provide richer information to learn from context and more space for improvements in prediction.

Qwen2.5-7B exhibits unusual performance on MNLI, significantly outperforming other models in the 4-shot setting but dropping to the lowest performance in the 128-shot setting. This anomaly is validated through comparisons across models and demonstration selection methods. As shown in Fig. 8, this behavior appears unique to Qwen2.5-7B and irrelevant to demonstration selection, suggesting Qwen2.5-7B may be a distinct case on MNLI.

Dataset	Query Prompt with 4-shot In-Context Demonstrations selected by Random
SST-5	more of the same from taiwanese auteur tsai ming-liang , which is good news to anyone who 's fallen under the sweet , melancholy spell of this unique director 's previous films . It is good it 's a great performance and a reminder of dickens ' grandeur . It is great a series of escapades demonstrating the adage that what is good for the goose is also good for the gander , some of which occasionally amuses but none of which amounts to much of a story . It is bad one just waits grimly for the next shock without developing much attachment to the characters . It is OK in his first stab at the form , jacquot takes a slightly anarchic approach that works only sporadically . It is
MNLI	Instead of ancient artefacts it shows the lifestyle and achievements of myriad Jewish communities around the globe through high-tec h audio-visual displays, hands-on exhibits, scale models (many of which are exquisite), and reconstructions. Can we say "The Jewish communities had many ancient artifacts on display."? No Theirs is now the dominant right-wing critique of integrationist programs. Can we say "Right wing politics is praised by the integrationist."? No Lorenzo the Magnificent and brother Giuliano lie in simple tombs beneath the sculptor's Madonna and Child, flanked by lesser artists' statues of the family patron saints Cosmas and Damian. Can we say "Lorenzo and Giuliano were related to one another."? Yes The difference, if any, between the reacquisition price and the net carrying value of the extinguished debt should be recognized as a loss or gain in accounting for interest on Treasury debt. Can we say "The difference is not recognized."? No The new rights are nice enough Can we say "Everyone really likes the newest benefits "?
CMSQA	Question: Bill sits down on a whoopee cushion, what sound does he make when he sits? (A) fall asleep (B) flatulence (C) sigh of relief (D) medium (E) comfort Answer: B Question: What is likely heard by those going to a party? (A) smoking pot (B) happiness (C) laughter (D) babies (E) meet new people Answer: C Question: A handsome prince is a stock character common to what? (A) england (B) fairy tale (C) castle (D) palace (E) court Answer: B Question: What covers the largest percentage of the pacific northwest? (A) united states (B) united states (C) washington (D) oregon (E) british columbia Answer: B Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? (A) bank (B) library (C) department store (D) mall (E) new york Answer:
HeSwag.	Choose an ending: The player is holding a bat walking in the field. The player... (A) in front stands across the screen and the ball dunks him to the ground. (B) makes a goal and the male scores. (C) runs down a purple net towards the ball. (D) begins to hit the ball. Answer: B Choose an ending: A woman sitting on the floor speaks to the camera. She... (A) begins blowing cover on the phone. (B) pulls a harmonica and ties the end of the violin. (C) begins to play a saxophone in the bedroom. (D) holds onto a cat in front of her. Answer: D Choose an ending: Someone jumps out the living room window. He... (A) dives for her pipe. (B) tuxedos her free eyes. (C) sees the girl chasing after him. (D) is as grim as he can. Answer: C Choose an ending: A group sits on horses as the stand and take a rest. The group... (A) continues jumping around a gym as the crowd claps. (B) slowly slowly make their way to the other side of the pool on the level. (C) rides on the horses along a trail. (D) pass between horses at the left. Answer: C Choose an ending: Students lower their eyes nervously. She... (A) pats her shoulder, then saunters toward someone. (B) turns with two students. (C) walks slowly towards someone. (D) wheels around as her dog thunders out. Answer:
GeoQ.	how many rivers are there in m0 answer: count(intersection(river,loc_2(m0))) through which states does the m0 flow answer: intersection(state,traverse_1(m0)) through which states does the m0 run answer: intersection(state,traverse_1(m0)) what is the most populous city in m0 answer: largest_one(population_1,intersection(city,loc_2(m0))) name all the rivers in m0 answer:
NL2Bash	Get the path of running Apache ps -ef | grep apache Gets domain name from dig reverse lookup. $dig -x 8.8.8.8 | grep PTR | grep -o google.* Find all *.xml files under current directory find -name *.xml Disables shell option 'nullglob'. shopt -u nullglob Add executable permission to "pretty-print"
Break	when did vincent von gogh die? 1#) return vincent von gogh 2#) return when did #1 die Does the Bronx have more non-Hispanic whites, or Hispanic whites? 1#) return the Bronx 2#) return non-Hispanic whites in #1 3#) return Hispanic whites in #1 4#) return number of #2 5#) return number of #3 6#) return which is highest of #4 , #5 What is the oldest log id and its corresponding problem id? 1#) return log ids 2#) return #1 that is the oldest 3#) return corresponding problem id of #2 4#) return #2 , #3 How many years apart were the British Saloon Car Championship season wins after 1970? 1#) return the British Saloon Car Championship 2#) return season wins of #1 3#) return years of #2 4#) return #3 that were after 1970 5#) return the difference of #4 what flights are available tomorrow from denver to philadelphia 1#)
MTOP	Ask Rob how bad the traffic is [IN:SEND_MESSAGE [SL:RECIPIENT Rob] [SL:CONTENT_EXACT how bad the traffic is]] Can I have the headlines please [IN:GET_STORIES_NEWS [SL:NEWS_TYPE headlines]] what are the headlines in ohio news? [IN:GET_STORIES_NEWS [SL:NEWS_TYPE headlines] [SL:NEWS_TOPIC ohio] [SL:NEWS_TYPE news]] What is the most recent news regarding local politics [IN:GET_STORIES_NEWS [SL:DATE_TIME the most recent] [SL:NEWS_TYPE news] [SL:NEWS_CATEGORY local politics]] call Nicholas and Natasha
SMCal.	What about with Kaitlin Taylor? (Yield (Execute (NewClobber (refer ((̂Dynamic) ActionIntensionConstraint)) ((̂Recipient) ConstraintTypeIntension) (intension (RecipientWithNameLike ((̂Recipient) EmptyStructConstraint) (PersonName.apply "Kaitlin Taylor")))))) Sorry I meant the weekend? (Yield (Execute (ReviseConstraint (refer ((̂Dynamic) roleConstraint (Path.apply "output"))) ((̂Event) ConstraintTypeIntension) (Event.start_? (DateTime.date_? (ThisWeekend)))))) CHECK FOR ALBERT'S BIRTHDAY EVENT (Yield (FindEventWrapperWithDefaults (Event.subject_? (? = "ALBERT'S BIRTHDAY")))) Add meetings same time everyday for this week. (FenceRecurring) Change the reservation for tonight to 6 people from 4.

Table 8: Dataset cases with random 4-shot in-context demonstrations.

Dataset	Query Prompt with 4-shot In-Context Demonstrations selected by Latent-Bayesian
SST-5	renner ? It is OK apart from its own considerable achievement , metropolis confirms tezuka 's status as both the primary visual influence on the animé tradition and its defining philosophical conscience . It is OK ... salaciously simplistic . It is OK ... would be a total loss if not for two supporting performances taking place at the movie 's edges . It is OK in his first stab at the form , jacquot takes a slightly anarchic approach that works only sporadically . It is
HeSwag.	Choose an ending: Richard Parker frantically swipes at the small fish as someone shields his face. The fluttering silver creatures... (A) spoon in around him. (B) inserting its wands into their palm. (C) pelt his exposed torso. (D) fly over him, catching and grabbing the sheet of flame. Answer: C Choose an ending: The sheer athleticism of his movement keeps him ahead of the choppers as he leads them on. Back at the window, someone... (A) ducks down, gathers his cube and hunches at one. (B) groggily wakes with a crowbar. (C) gazes into a narrow gap between a tall arch. (D) leans her head against the rotting frame. Answer: D Choose an ending: The door closes after him, and unable to stop, the keys herd into it. Huge broken stone statues... (A) lie on an maize field. (B) are about to emerge, which painted walls. (C) lie about the floor. (D) are drawn around the arena empty gleaming guns. Answer: C Choose an ending: A strong wind flings leaves high into the air and the hedges advance. As the hedges close in behind them, they both... (A) climb the rock - wakeboard on the side of the structure, to avoid it. (B) sip of the wand. (C) move, still throwing his legs to each other. (D) pose, once aftermath from all angles. Answer: D Choose an ending: Students lower their eyes nervously. She... (A) pats her shoulder, then saunters toward someone. (B) turns with two students. (C) walks slowly towards someone. (D) wheels around as her dog thunders out. Answer:

Table 9: Demonstration cases selected by Latent-Bayesian. On SST-5, Latent-Bayesian tends to exclusively select demonstrations with the neutral label `OK`. On HellaSwag, it predominantly selects demonstrations with answers `C` or `D`. Specifically, for 128-shot demonstrations on HellaSwag, Latent-Bayesian selected answers `A`, `B`, `C`, and `D` in proportions of 8, 0, 45, and 75 respectively. This is in contrast to the training set of $50,000$ examples where the four answers are distributed equally ($12,500 \pm 68$ each).

Dataset	Query Prompt with 4-shot In-Context Demonstrations selected by BM25-Major
Break	What shape is the object is a different shape and size than the other objects? 1#) return objects 2#) return shapes of #1 3#) return sizes of #1 4#) return number of #1 for each #2 5#) return number of #1 for each #3 6#) return #2 where #4 is one 7#) return #3 where #5 is one 8#) return #1 where #2 is #6 9#) return #1 where #3 is #7 10#) return #1 in both #8 and #9 11#) return the shape of #10 Are all the spheres the same material? 1#) return spheres 2#) return the materials of #1 3#) return number of #1 for each #2 4#) return #2 where #3 is the highest 5#) return #1 where #2 is #4 6#) return number of #5 7#) return number of #1 8#) return if #6 and #7 are equal What is the count and code of the job with the most employee? 1#) return employees 2#) return jobs of #1 3#) return number of #1 for each #2 4#) return the highest of #3 5#) return #2 where #3 is #4 6#) return the code of #5 7#) return #4 , #6 What color is the object that is a different material than the others? 1#) return objects 2#) return materials of #1 3#) return the number of #1 for each #2 4#) return #2 where #3 is one 5#) return #1 where #2 is #4 6#) return the color of #5 what flights are available tomorrow from denver to philadelphia 1#)
MTOP	remind me to call my father for fathers day [IN:CREATE_REMINDER [SL:PERSON_REMINDED me] [SL:TODO [IN:GET_TODO [SL:TODO [IN:CREATE_CALL [SL:CONTACT [IN:GET_CONTACT [SL:CONTACT_RELATED my] [SL:TYPE_RELATION father]]]]] [SL:TODO for fathers day]]]] Set a reminder to call my husband when I get home [IN:CREATE_REMINDER [SL:TODO [IN:GET_TODO [SL:TODO [IN:CREATE_CALL [SL:CONTACT [IN:GET_CONTACT [SL:CONTACT_RELATED my] [SL:TYPE_RELATION husband]]]]] [SL:LOCATION [IN:GET_LOCATION]]]]] Remind me to call my brother after work.[IN:CREATE_REMINDER [SL:PERSON_REMINDED me] [SL:TODO [IN:GET_TODO [SL:TODO [IN:CREATE_CALL [SL:CONTACT [IN:GET_CONTACT [SL:CONTACT_RELATED my] [SL:TYPE_RELATION brother]]]]] [SL:TODO after work]]]] Remind me to call my father today. [IN:CREATE_REMINDER [SL:PERSON_REMINDED me] [SL:TODO [IN:CREATE_CALL [SL:CONTACT [IN:GET_CONTACT [SL:CONTACT_RELATED my] [SL:TYPE_RELATION father]]]]] [SL:DATE_TIME today]] call Nicholas and Natasha
SMCal.	What is on my calendar tomorrow? I want to have lunch with Sara, Barack and Monica. (do (Yield (FindEventWrapperWithDefaults (EventOnDate (Tomorrow) (ˆ(Event) EmptyStructConstraint)))) (Yield (CreateCommitEventWrapper (CreatePreflightEventWrapper (& (& (Event.subject_? (?= "lunch")) (Event.start_? (DateTime.date_? (?= (Tomorrow))))) (Event.attendees_? (& (& (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Sara")))))) (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Barack"))))))) (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Monica"))))))))))))) I would like Tuesday. What time is my team sync with Abby, Elli and Kim? (do (Yield (Execute (ChooseCreateEventFromConstraint (Event.start_? (DateTime.date_? (Date.dayOfWeek_? (?= (Tuesday))))) (refer (ˆ(Dynamic) ActionIntensionConstraint))))) (Yield (Event.start (singleton (QueryEventResponse.results (FindEventWrapperWithDefaults (& (Event.subject_? (? = "team sync")) (Event.attendees_? (& (& (AttendeeListHasRecipientConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Abby"))) (AttendeeListHasRecipientConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Elli")))) (AttendeeListHasRecipientConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Kim")))))))))))) Change the lunch on Friday to being at 1pm. Also invite Sally, Alexa, and Sandra to the lunch. (Yield (UpdateCommitEventWrapper (UpdatePreflightEventWrapper (Event.id (singleton (QueryEventResponse.results (FindEventWrapperWithDefaults (& (Event.subject_? (? = "lunch")) (Event.start_? (DateTime.date_? (Date.dayOfWeek_? (?= (Friday)))))))))) (& (Event.start_? (DateTime.time_? (?= (NumberPM 1L)))) (Event.attendees_? (& (& (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Sally")))))) (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Alexa"))))))) (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Sandra")))))))))))) Put a lunch meeting with Abby, Kim, and Jesse at Taco Bell tomorrow (Yield (CreateCommitEventWrapper (CreatePreflightEventWrapper (& (& (& (Event.subject_? (?= "lunch meeting")) (Event.start_? (DateTime.date_? (?= (Tomorrow))))) (Event.location_? (?= (LocationKeyphrase.apply "Taco Bell")))) (Event.attendees_? (& (& (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Abby")))))) (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Kim"))))))) (AttendeeListHasRecipient (Execute (refer (extensionConstraint (RecipientWithNameLike (ˆ(Recipient) EmptyStructConstraint) (PersonName.apply "Jesse")))))))))))) Change the reservation for tonight to 6 people from 4.

Table 10: Demonstration cases selected by BM25-Major. BM25-Major tends to select significantly longer demonstrations. This can be attributed to the fact that longer examples tend to have higher BM25 scores.

Method	SST-5	MNLI	CMSQA	HeSwag.	GeoQ.	NL2Bash	Break	MTOP	SMCal.	Average
Llama3-8B										
Random	37.75	54.32	72.99	41.51	49.14	20.14	27.19	34.33	29.01	40.71
BM25-Major	38.78	40.62	70.84	39.62	35.71	28.66	2.16	5.37	6.56	29.81
BGE-KMeans	37.15	61.85	73.55	36.02	**58.21**	19.90	23.98	33.20	28.69	41.39
Best-of-N	39.33	51.92	73.38	40.58	53.93	26.78	25.88	35.48	**30.55**	41.98
EPR-KMeans	37.60	53.98	73.87	39.99	49.64	17.65	**30.09**	36.51	29.63	41.00
Latent-Bayesian	37.15	44.04	73.22	33.81	41.79	24.72	25.66	32.30	18.13	36.76
CLG (ours)	**39.96**	**61.98**	**73.96**	**46.01**	**58.21**	**33.83**	29.06	**37.18**	27.39	**45.29**
Llama3-70B										
Random	43.71	61.83	**84.73**	76.11	73.36	29.72	35.55	44.66	36.68	54.04
BM25-Major	43.69	51.36	82.23	72.01	43.21	27.92	15.05	8.41	14.38	39.81
BGE-KMeans	46.41	65.09	83.78	73.90	**84.64**	28.39	35.13	44.79	38.73	55.65
Best-of-N	43.69	70.92	84.19	74.57	80.71	**29.76**	35.86	43.40	37.66	55.64
EPR-KMeans	43.05	49.78	83.62	71.32	78.21	22.92	**38.39**	**50.43**	38.10	52.87
Latent-Bayesian	46.59	69.84	84.60	70.09	62.14	29.20	32.59	41.21	23.13	51.04
CLG (ours)	**48.32**	**76.37**	84.52	**80.77**	**84.64**	29.59	37.33	47.74	**40.35**	**58.85**
Qwen2.5-3B										
Random	19.05	51.15	68.76	35.30	47.93	14.08	23.29	25.94	19.07	33.84
BM25-Major	**26.70**	45.33	68.47	33.79	36.07	**28.53**	3.56	5.01	0.12	27.51
BGE-KMeans	26.43	40.70	**73.79**	32.72	51.43	19.74	19.83	26.26	**22.20**	34.79
Best-of-N	21.25	**56.02**	71.50	41.10	**55.00**	21.36	23.45	26.31	21.15	**37.46**
EPR-KMeans	19.35	54.53	70.93	33.61	49.29	11.79	23.80	**30.07**	17.66	34.56
Latent-Bayesian	21.07	44.66	69.62	25.68	28.93	11.26	20.91	23.80	11.56	28.61
CLG (ours)	24.07	54.42	72.56	**44.77**	52.50	17.62	**24.90**	26.49	19.40	37.41
Qwen2.5-7B										
Random	26.48	31.88	84.88	71.38	35.43	16.93	28.22	32.72	27.99	39.55
BM25-Major	**29.34**	31.88	83.78	71.50	41.43	**32.27**	8.65	7.79	11.98	35.40
BGE-KMeans	26.25	31.88	85.26	**74.10**	**56.79**	19.46	25.28	33.20	**30.73**	**42.55**
Best-of-N	24.80	31.88	84.77	72.69	26.07	15.72	27.80	32.75	28.71	38.35
EPR-KMeans	24.70	**31.94**	85.26	71.33	51.79	15.44	**31.17**	**35.30**	25.43	41.37
Latent-Bayesian	12.62	31.88	84.93	60.58	21.43	13.31	26.21	26.13	17.66	32.75
CLG (ours)	24.89	**31.94**	**85.34**	71.76	50.00	19.54	28.27	32.98	28.03	41.42
Qwen2.5-14B										
Random	29.55	56.76	82.29	55.53	61.71	28.00	31.41	38.74	33.73	46.41
BM25-Major	19.53	55.98	79.28	46.52	51.79	41.82	14.85	10.02	12.82	36.96
BGE-KMeans	**34.79**	**57.31**	82.88	60.61	**74.64**	31.99	30.13	40.89	31.93	49.46
Best-of-N	28.97	55.17	82.56	52.92	59.64	36.30	30.44	42.24	32.95	46.80
EPR-KMeans	18.07	56.24	82.06	51.67	58.93	24.39	**32.41**	**43.18**	33.95	44.54
Latent-Bayesian	22.16	56.20	81.98	63.08	47.50	24.77	30.39	28.95	16.71	41.30
CLG (ours)	32.79	56.30	**83.05**	**69.35**	60.36	**42.43**	32.20	39.51	**34.95**	**50.10**
Qwen2.5-32B										
Random	13.93	61.58	85.54	84.62	72.29	28.51	36.24	43.19	37.43	51.48
BM25-Major	10.72	62.23	85.09	80.69	57.86	**41.68**	21.38	11.23	21.06	43.55
BGE-KMeans	12.62	60.96	85.26	83.12	**80.00**	28.11	35.35	43.76	38.51	51.97
Best-of-N	**17.44**	62.24	57.25	85.31	73.57	37.11	34.91	46.94	35.36	50.01
EPR-KMeans	14.62	61.41	85.67	83.60	76.79	28.35	**38.32**	**49.13**	**40.98**	53.21
Latent-Bayesian	0.00	58.68	85.91	81.66	58.57	31.57	34.88	33.24	22.21	45.19
CLG (ours)	16.17	**73.33**	**86.40**	**86.53**	73.57	29.97	36.26	46.62	36.75	**53.96**
Qwen2.5-72B										
Random	36.00	58.20	87.47	86.58	57.29	33.28	36.67	45.32	38.23	53.23
BM25-Major	37.78	45.94	86.65	85.61	52.50	38.11	22.81	16.24	11.57	44.13
BGE-KMeans	37.24	49.65	87.71	86.67	61.79	40.86	35.84	43.71	**41.29**	53.86
Best-of-N	36.60	60.57	87.55	86.83	61.07	37.13	35.73	46.89	39.22	54.62
EPR-KMeans	36.15	58.41	**87.96**	86.22	**62.86**	26.53	**39.12**	**47.53**	39.65	53.83
Latent-Bayesian	27.25	47.96	85.75	86.72	27.50	37.36	35.35	5.23	22.61	41.75
CLG (ours)	**38.33**	**77.29**	87.71	**87.68**	62.50	**45.19**	37.07	47.07	40.16	**58.11**

Table 11: Main results, merely 128-shot, on all models. The best scores on each model are in bold, and the second-best ones are underlined.

Method	SST-5	MNLI	CMSQA	HeSwag.	GeoQ.	NL2Bash	Break	MTOP	SMCal.	Average
Llama3-8B										
Random	36.64	45.95	72.71	39.84	28.26	19.53	19.80	16.87	14.97	32.73
BM25-Major	32.26	37.67	71.51	38.40	23.09	22.65	1.43	3.44	4.72	26.13
BGE-KMeans	37.16	45.90	72.99	35.80	32.20	20.23	14.21	18.31	16.72	32.61
Best-of-N	**39.15**	48.50	**73.63**	40.51	**35.77**	27.22	21.14	20.42	17.32	35.96
EPR-KMeans	35.12	43.26	73.53	38.75	30.60	18.87	21.03	19.02	16.09	32.92
Latent-Bayesian	36.69	37.88	72.66	34.15	18.15	16.15	18.96	14.21	9.79	28.74
CLG (ours)	38.72	**51.86**	73.52	**44.27**	34.11	**35.47**	**21.41**	**20.85**	**18.18**	**37.60**
Llama3-70B										
Random	42.39	50.82	83.59	72.82	46.21	31.42	29.39	22.76	18.72	44.24
BM25-Major	41.10	44.79	81.89	71.01	35.18	31.18	9.83	5.58	12.70	37.03
BGE-KMeans	42.72	53.00	82.90	60.84	50.06	32.34	27.11	22.67	19.89	43.50
Best-of-N	44.69	49.75	83.24	75.63	**51.79**	33.67	30.08	24.41	21.77	46.11
EPR-KMeans	41.34	45.13	83.25	71.58	49.16	25.84	**31.36**	26.56	21.40	43.96
Latent-Bayesian	44.26	50.75	83.30	55.33	27.44	29.23	25.36	18.84	11.53	38.45
CLG (ours)	**46.17**	**57.28**	**83.69**	**79.48**	51.61	**36.15**	30.47	**27.18**	**24.30**	**48.48**
Qwen2.5-3B										
Random	17.81	49.79	70.77	36.96	25.67	15.98	14.48	13.18	9.80	28.27
BM25-Major	21.15	50.18	68.40	36.65	20.83	18.32	1.01	2.26	0.16	24.33
BGE-KMeans	21.69	46.76	72.37	34.03	28.87	17.23	9.93	13.23	12.18	28.48
Best-of-N	**23.28**	**60.32**	72.73	**41.12**	**33.09**	19.18	15.28	14.87	**13.57**	**32.60**
EPR-KMeans	15.99	50.41	71.55	34.41	30.89	12.99	14.14	**16.41**	12.45	28.81
Latent-Bayesian	20.85	51.93	67.98	32.97	13.21	10.81	13.37	9.50	10.20	25.65
CLG (ours)	22.09	59.70	**73.59**	39.39	27.14	**27.20**	**16.61**	15.54	8.24	32.17
Qwen2.5-7B										
Random	23.84	45.36	84.88	75.65	14.69	17.38	21.92	16.09	14.81	34.96
BM25-Major	24.51	41.30	83.58	73.83	17.86	20.58	5.44	3.70	8.76	31.06
BGE-KMeans	24.95	32.15	84.98	75.48	20.24	18.97	17.00	17.41	16.95	34.24
Best-of-N	**28.82**	43.78	85.04	75.38	15.53	21.66	22.18	18.16	17.20	36.42
EPR-KMeans	19.94	44.04	**85.16**	75.38	18.81	15.30	**23.80**	**19.63**	15.31	35.26
Latent-Bayesian	15.41	40.92	84.07	66.86	8.33	16.61	17.54	10.48	13.70	30.44
CLG (ours)	25.10	**47.65**	84.86	**75.84**	**23.81**	**23.12**	21.66	18.50	**17.53**	**37.56**
Qwen2.5-14B										
Random	25.67	52.56	76.58	64.15	34.23	30.25	24.96	19.53	17.27	38.35
BM25-Major	16.09	48.36	78.31	61.59	36.61	37.25	8.92	6.42	8.16	33.52
BGE-KMeans	25.81	52.96	77.98	64.23	38.81	31.64	21.83	20.55	17.43	39.03
Best-of-N	21.03	51.27	76.58	64.44	**40.18**	40.16	25.57	22.59	18.81	40.07
EPR-KMeans	17.68	51.29	76.32	62.11	37.14	23.59	25.67	**23.32**	18.19	37.26
Latent-Bayesian	3.90	48.72	78.33	64.72	26.85	23.10	20.74	11.60	13.13	32.34
CLG (ours)	**29.32**	**53.83**	**79.06**	**70.19**	35.24	**41.38**	**25.79**	21.93	**20.44**	**41.91**
Qwen2.5-32B										
Random	12.85	62.43	85.56	85.48	41.83	32.31	28.97	22.12	19.01	43.40
BM25-Major	11.52	58.64	85.56	83.73	34.46	35.52	12.34	5.60	15.12	38.06
BGE-KMeans	**20.78**	57.03	**85.63**	84.82	**49.40**	31.00	26.03	21.89	19.29	43.99
Best-of-N	17.94	**66.03**	71.27	85.58	44.46	40.06	28.53	26.05	19.69	44.40
EPR-KMeans	11.16	62.86	85.52	85.01	48.63	31.08	**29.82**	**26.16**	21.72	44.66
Latent-Bayesian	0.05	57.73	83.70	84.42	27.08	33.02	24.21	13.17	15.05	37.60
CLG (ours)	19.10	63.38	85.46	**86.67**	44.46	**41.55**	28.01	25.46	**22.66**	**46.31**
Qwen2.5-72B										
Random	34.13	57.59	87.12	87.73	32.49	35.95	29.40	23.37	19.67	45.27
BM25-Major	28.76	49.16	85.72	86.88	36.25	43.46	16.47	7.78	13.41	40.88
BGE-KMeans	**38.86**	43.75	87.24	87.36	40.18	40.63	27.29	25.22	21.57	45.79
Best-of-N	37.54	62.57	86.86	88.02	39.94	40.88	30.37	25.14	20.98	48.03
EPR-KMeans	29.31	56.36	**87.99**	87.44	**42.02**	30.30	**30.67**	**25.89**	21.85	45.76
Latent-Bayesian	24.81	52.97	85.03	88.14	15.42	34.75	25.17	3.31	15.17	38.31
CLG (ours)	38.30	**64.05**	87.03	**88.32**	41.75	**47.67**	29.09	25.63	**23.79**	**49.51**

Table 12: Main results, averaged over 4-shot to 128-shot settings, on all models. The best scores on each model are in bold, and the second-best ones are underlined.

Method	SST-5	MNLI	CMSQA	HeSwag.	GeoQ.	NL2Bash	Break	MTOP	SMCal.	Average
Qwen-turbo-latest										
Random	56.87	80.93	86.47	79.07	73.43	62.61	44.20	28.73	28.40	60.08
±std	±2.36	±0.80	±1.21	±0.44	±3.97	±1.15	±2.88	±2.22	±1.99	±0.22
Best-of-N	52.67	80.33	85.33	**80.00**	72.86	61.79	42.00	**34.00**	27.67	59.63
CLG (ours)	**59.67**	**83.33**	**87.67**	79.00	**77.14**	**64.39**	**46.33**	30.67	**28.67**	**61.87**
GLM-4-flash										
Random	56.13	74.54	**87.73**	69.60	64.07	61.03	41.20	28.20	23.07	56.17
±std	±1.99	±1.24	±0.74	±13.97	±1.82	±1.03	±3.04	±0.96	±2.14	±2.24
Best-of-N	53.00	**76.33**	87.00	76.67	**66.07**	61.44	44.67	**32.33**	21.33	57.65
CLG (ours)	**56.33**	72.67	87.33	**77.67**	63.93	**62.24**	**51.67**	30.67	**24.00**	**58.50**
Doubao-pro-32k-0528										
Random	54.53	81.00	84.33	**79.07**	75.64	**64.57**	59.27	38.33	28.27	62.78
±std	±1.94	±1.97	±0.92	±0.68	±2.16	±0.78	±1.53	±2.09	±1.56	±0.26
Best-of-N	53.67	78.67	**84.67**	77.67	74.64	62.86	58.67	37.67	26.67	61.69
CLG (ours)	**55.00**	**81.33**	83.67	77.67	**79.64**	62.82	**60.33**	**46.33**	**30.00**	**64.09**
GPT-4o-mini										
Random	55.93	76.40	84.33	76.53	70.36	65.05	53.73	32.33	**27.33**	60.22
±std	±1.81	±1.06	±0.79	±0.75	±0.96	±0.44	±4.20	±2.03	±1.07	±0.45
Best-of-N	53.67	75.67	84.33	**77.67**	**72.86**	63.87	55.67	36.33	23.33	60.38
CLG (ours)	**56.00**	**78.67**	**86.67**	76.67	68.93	**65.72**	**57.33**	**36.67**	26.67	**61.48**
DeepSeek-V3										
Random	**58.20**	85.20	87.80	80.93	83.43	**67.77**	59.73	41.60	36.80	66.83
±std	±2.57	±1.17	±1.00	±1.65	±2.48	±0.60	±2.09	±3.71	±2.90	±0.52
Best-of-N	57.67	**85.67**	86.33	81.33	83.93	65.84	59.00	**47.33**	34.00	66.79
CLG (ours)	56.33	85.33	**88.00**	**83.33**	**87.14**	67.28	**62.00**	45.33	**38.00**	**68.08**

Table 13: Full results of closed-source models under the 128-shot setting. We only evaluate at most 300 instances on each dataset for limited budget. Best scores on each model are in bold.

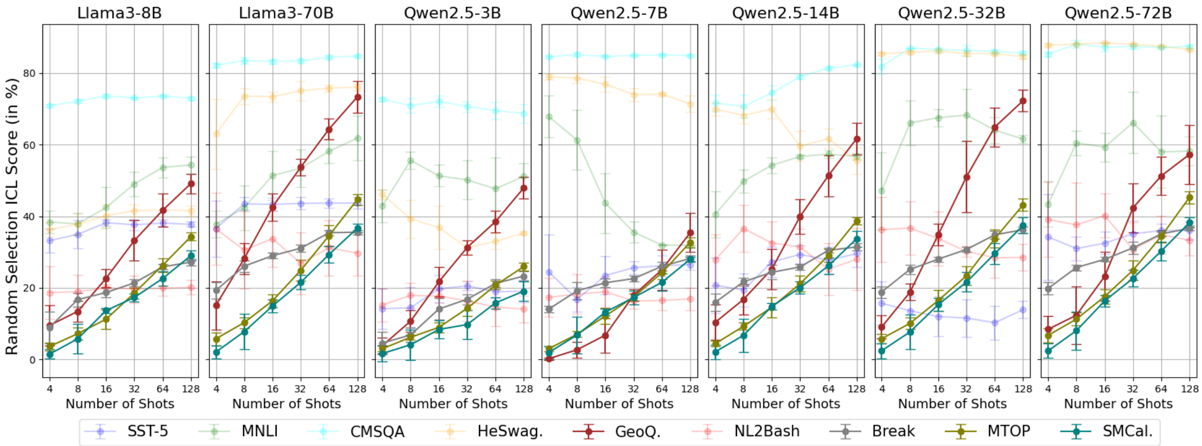

Figure 7: The scaling trends of random selection with respect to the number of shots on all models. We illustrate the standard deviations under 5 random seeds through error bars. The datasets that can steadily benefit from more shots are in solid colors, and the others are in semi-transparent colors.

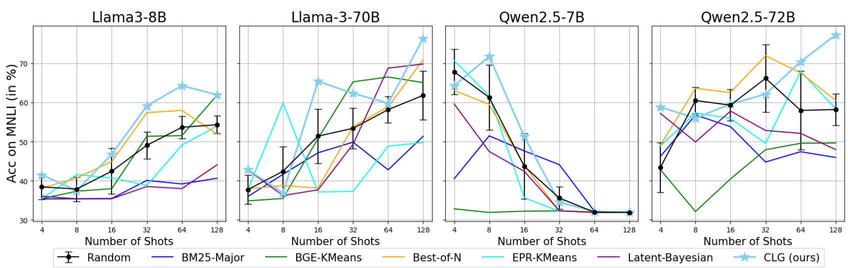

Figure 8: The performance of different models on MNLI.