Revisiting In-Context Learning with Long Context Language Models

Anonymous ACL submission

Abstract

In-Context Learning (ICL) is a technique by which language models make predictions based on examples provided in their input context. Previously, their context window size imposed a limit on the number of examples that can be shown, making example selection techniques crucial for identifying the maximally effective set of examples. However, the recent advent of Long Context Language Models (LCLMs) has significantly increased the number of examples that can be included in context, raising an important question of whether ICL performance in a many-shot regime is still sensitive to the method of sample selection. To answer this, we revisit these approaches in the context of LCLMs through extensive experiments on 18 datasets spanning 4 tasks. Surprisingly, we observe that sophisticated example selection techniques do not yield significant improvements over a simple random sample selection method. Instead, we discover that the advent of LCLMs has fundamentally shifted the challenge of ICL from that of selecting the most effective examples to that of collecting sufficient examples to fill the context window. Specifically, in certain datasets, including all available examples does not fully utilize the context window; however, by augmenting the examples in context with a simple data augmentation approach, we substantially improve ICL performance by 5%.

1 Introduction

In-Context Learning (ICL) has emerged as a powerful paradigm in natural language processing that enables Language Models (LMs) to learn, adapt, and generalize from examples provided within their input context, eliminating the need for extensive training and parameter updates (Brown et al., 2020; Min et al., 2022; von Oswald et al., 2023). However, due to the limited context lengths of earlier LMs (which accommodate only a few thousand tokens), much of previous ICL work has focused on optimizing sample selection strategies (Liu et al., 2021; Rubin et al., 2022; Sorensen et al., 2022; An et al., 2023; Mavromatis et al., 2023; Liu et al., 2024). With the advent of Long Context Language Models (LCLMs), which are capable of processing over a million tokens in a single context window, these constraints are significantly relaxed as it enables including a large number of examples to be used in ICL, known as many-shot ICL (Agarwal et al., 2024; Bertsch et al., 2024).

This expansion of context length raises an important question: do previous sample selection strategies, designed for shorter context windows in earlier LMs, generalize to the many-shot ICL regime? To answer this, we systematically revisit existing sample selection strategies by conducting extensive experiments across 18 datasets spanning diverse tasks (namely, classification, translation, summarization, and reasoning) with multiple LCLMs. Our experiments include multiple types of sample selection methods: relevance, diversity, and difficulty-based sample selection, as outlined in Dong et al. (2023). From these experiments, we uncover novel and surprising findings: contrary to prevailing expectations that carefully selected ICL demonstrations would yield performance improvements, they are similarly effective with a simple random selection approach, offering no statistically meaningful improvements in almost all cases (Figure 1). An additional reason to prefer the naive sample selection approach is that it enables greater efficiency through key-value caching of in-context examples (as the same examples can be reused across multiple queries), unlike sophisticated sample selection methods where the examples vary for each sample.

While the expanded context length in LCLMs allows us to focus less on selecting optimal subsets of examples, it introduces a new challenge: effectively utilizing this expanded capacity when the number of examples is limited. Specifically, in scenarios where available data is sparse (such as

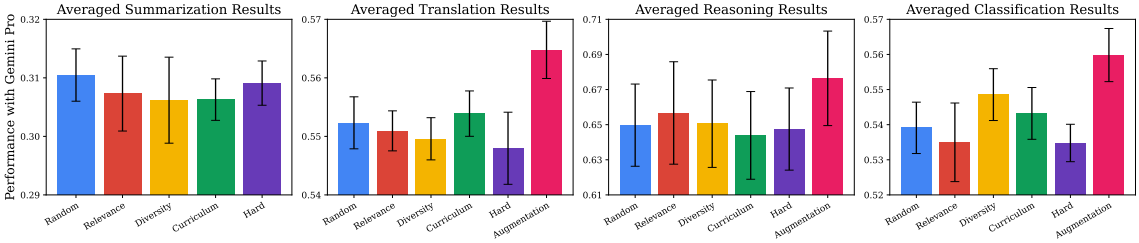

Figure 1: Results of various sample selection approaches in 64-shot ICL with LCLMs. Approaches include Retrieval that selects examples similar to the target query, Diversity that aims for maximizing example variety, Curriculum that arranges examples in order from easiest to hardest, and Hard that uses only challenging examples, alongside Random that selects examples without any constraints. Results indicate that sample selection methods provide no significant improvement over the naive (random) approach and sometimes perform worse. Meanwhile, Augmentation refers to the approach that generates additional demonstrations and uses them along with original samples for ICL, particularly for low-resource tasks (such as translation, reasoning, and classification) that do not contain enough samples to utilize the full capacity of LCLMs, showing substantial performance gains.

low-resource translation or reasoning tasks where annotated data samples are difficult or costly to obtain), the examples available only utilize a small fraction of the full context window. This mismatch between context capacity and example availability introduces a new direction in ICL research, shifting the focus from optimizing sample selection to maximally utilizing the long context window. To address this, we propose a simple yet effective data augmentation approach to increase the number of in-context examples, which consists of two steps: (1) generating synthetic examples and (2) filtering out low-quality examples through LCLM prompting contextualized with real examples. Then, by adding these augmented data samples to the context, we significantly improve ICL performance.

Moreover, we explore other key factors unique to LCLM-enabled ICL. Specifically, we investigate the capacity of LCLMs to comprehend extremely long context (where a large number of examples up to the context length are present), as well as how they handle scenarios in which some of these examples introduce noise. Through comprehensive analyses, we find that while performance generally improves as the number of in-context examples increases, it eventually plateaus and begins to decline as the context length approaches the limit. This diminishing return highlights the need to carefully balance context length and example quantity. Also, we observe that LCLMs exhibit robustness to noisy examples in relatively simple tasks, but become vulnerable to noise in more complex scenarios to which they might be less exposed during training, such as extremely low-resource translation tasks.

2 Examining Sample Selection Methods for In-Context Learning with LCLMs

2.1 Background

We begin with formally introducing LCLMs, followed by describing the setup of ICL with LCLMs.

Long-Context Language Models A language model (LM), which takes an input sequence of tokens $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ and generates an output sequence of tokens $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]$, can be denoted as follows: $\boldsymbol{y} = \texttt{LM}_\theta(\boldsymbol{x})$, where θ is the set of model parameters. A long-context LM (LCLM) is an advanced LM (Reid et al., 2024) that is designed to accommodate sequences with a large number of tokens (e.g., n can exceed 1 million), typically far surpassing the context sizes of earlier LMs.

In-Context Learning with LCLMs Given a set of k input-output pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^k$ as well as an input query $\boldsymbol{x}'$, the goal of ICL is to produce an output $\boldsymbol{y} = \texttt{LCLM}(\boldsymbol{x}'|\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^k)$, where the model (LCLM) uses the contextual examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^k$ to make predictions for $\boldsymbol{x}'$. In prior research before the advent of LCLMs, the value of k was often limited by the relatively short context lengths of earlier models, which constrained the number of examples that could be utilized for ICL. Subsequently, significant work has focused on developing sample selection techniques to optimize performance within these restricted contexts (Liu et al., 2021; Rubin et al., 2022; Sorensen et al., 2022; An et al., 2023; Mavromatis et al., 2023; Liu et al., 2024). In the meantime, the expanded context capacity of LCLMs enables a larger k, facilitating many-shot learning with a far greater number of examples.

2.2 Experimental Setup

We now discuss the detailed experimental design.

Tasks and Datasets We experiment with 18 different datasets across four tasks to evaluate the effectiveness and robustness of various approaches.

- **Translation:** This task evaluates the ability of models to translate text from one language to another. We include translations from English to low-resource languages (namely, Bemba, Northern Kurdish, and Ewe) and high-resource lan-

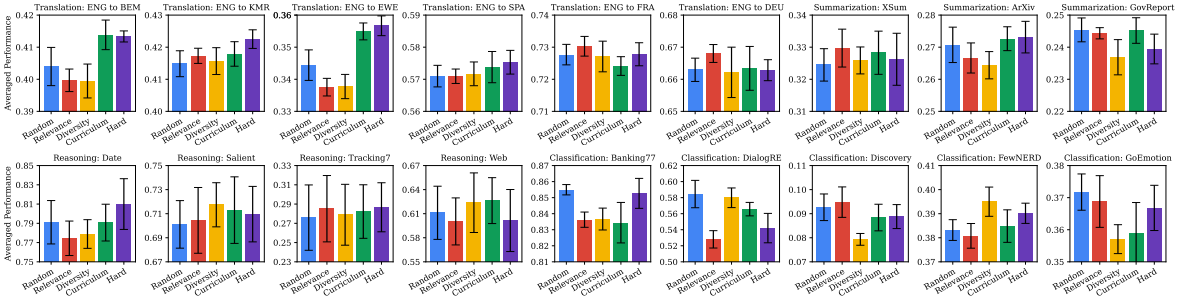

Figure 2: Results of various sample selection approaches on ICL of 64 examples with LCLMs, where we average the performance over all models: Gemini Pro, Gemini Flash, and Llama 3.1, across four different tasks with 18 datasets. Each bar represents the averaged performance, with the upper and lower limits indicating standard deviation. See Figure 9 for results on each model.

guages (Spanish, French, and German) from the FLORES-200 benchmark (NLLB et al., 2022), with chrF scores (Popovic, 2015) as the metric.

- **Summarization:** This task assesses the capability of models to generate concise and coherent summaries from articles. We include one widely-used XSum dataset (Narayan et al., 2018) and two long-context summarization datasets: ArXiv and GovReport (Cohan et al., 2018; Huang et al., 2021). ROUGE-L score is used for evaluation.
- **Reasoning:** This task evaluates the ability of models on complex reasoning. We use four challenging datasets from Big Bench Hard (Suzgun et al., 2022) following the experimental setting of Long-Context Frontiers (LOFT) benchmark (Lee et al., 2024a), where each data sample follows a multiple-choice question answering format.
- **Classification:** This task includes challenging benchmark datasets for ICL from Li et al. (2024), particularly designed for classification problems with diverse classes and long inputs.

ICL Sample Selection Strategies To ensure comprehensive coverage of previously explored sample selection strategies, we follow the category of three core dimensions from Dong et al. (2023) (that extensively summarizes around ICL 200 papers). This includes selecting samples based on their diversity, difficulty, and relevance to the query, with the baseline of random sample selection.

- **Naive:** This method randomly selects examples from a dataset and uses this initial set of selected examples as ICL demonstrations for all queries.
- **Relevance:** This method selects examples that are most similar to the input query to maximize the alignment of ICL demonstrations with the query. To compute semantic similarity between the query and each example, we use the state-of-the-art embedding model (Lee et al., 2024b).
- **Diversity:** This method selects examples that are maximally distinct from each other to capture a broad coverage of features within the task space. We embed each example in a shared embedding space with Lee et al. (2024b) and utilize k-means clustering (where k corresponds to the number of desired ICL examples) to group the examples into subcategories. We then select the example closest to each cluster center as the representative to capture a diverse subset of the task features.
- **Difficulty:** This method selects examples based on their difficulty. We examine two approaches: the first method (called **Curriculum**) follows a curriculum learning paradigm where examples are ordered from easiest to hardest; the second one (called **Hard**) includes only difficult examples, as simpler examples may already be well-understood by models. To assess example difficulty, we use model-based evaluation (Liu et al., 2023) with the state-of-the-art LCLM (Reid et al., 2024), which prompts a model 30 times and averages difficulty scores weighted by probabilities.

LCLM Configurations for ICL We consider LCLMs that support extensive token capacities to evaluate performance in long-context, many-shot ICL scenarios, such as those with context window lengths on the order of millions: Gemini 1.5 Flash (1M tokens) and Gemini 1.5 Pro (2M tokens) (Reid et al., 2024). Also, we consider the Llama 3.1 70B model (Dubey et al., 2024), which, while supporting the comparatively smaller context size of 128K tokens, is still considered an LCLM. To provide a comprehensive view of performance under different shots, we vary the number of ICL examples, starting from one and sequentially doubling to 2, 4, 8, 16, 32, and so forth, until reaching either the context size limit or the maximum number of dataset samples, whichever is exhausted first. Furthermore, to ensure the reliability of our results, we conduct multiple runs for each setup: 3 runs for translation and summarization tasks and 10 runs for reasoning and classification tasks. The prompts used to elicit responses from ICL are provided in Appendix A.

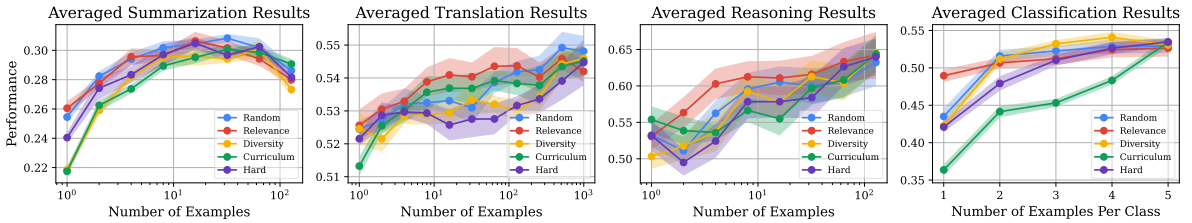

Figure 3: Results with varying the number of examples for ICL with Gemini Pro, where we average the results for each task.

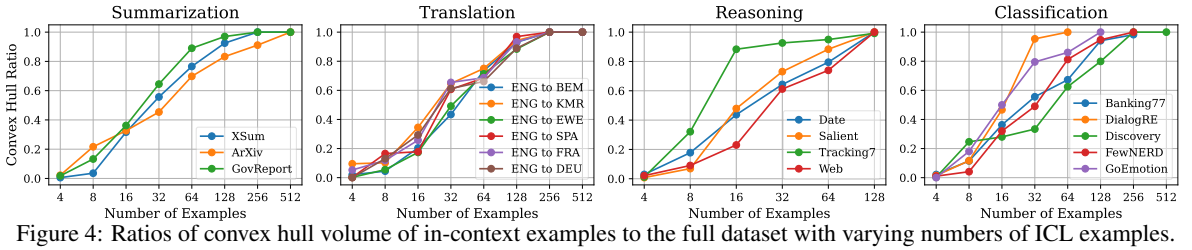

Figure 4: Ratios of convex hull volume of in-context examples to the full dataset with varying numbers of ICL examples.

Table 1: Counting the statistical significance of sophisticated selection approaches over random selection on each experiment instance, by conducting the t-test with 95% confidence threshold. Tran., Summ., Reas, Clas, denote translation, summarization, reasoning, and classification tasks, respectively.

LCLMs	Methods	Tran.	Summ.	Reas.	Clas.	Total
Gemini Pro	Relevance	0 / 6	0 / 3	0 / 4	0 / 5	0 / 18
	Diversity	0 / 6	0 / 3	1 / 4	2 / 5	3 / 18
	Curriculum	1 / 6	0 / 3	0 / 4	1 / 5	2 / 18
	Hard	0 / 6	0 / 3	1 / 4	0 / 5	1 / 18
Gemini Flash	Relevance	0 / 6	0 / 3	0 / 4	2 / 5	2 / 18
	Diversity	0 / 6	0 / 3	0 / 4	2 / 5	2 / 18
	Curriculum	0 / 6	0 / 3	0 / 4	0 / 5	0 / 18
	Hard	0 / 6	0 / 3	0 / 4	0 / 5	0 / 18
Llama 3.1	Relevance	1 / 6	0 / 3	1 / 4	1 / 5	3 / 18
	Diversity	0 / 6	0 / 3	0 / 4	2 / 5	2 / 18
	Curriculum	0 / 6	0 / 3	0 / 4	1 / 5	1 / 18
	Hard	0 / 6	0 / 3	0 / 4	2 / 5	2 / 18
Total	Relevance	1 / 18	0 / 9	1 / 12	3 / 15	5 / 54
	Diversity	0 / 18	0 / 9	1 / 12	6 / 15	7 / 54
	Curriculum	1 / 18	0 / 9	0 / 12	2 / 15	3 / 54
	Hard	0 / 18	0 / 9	1 / 12	2 / 15	3 / 54

Table 2: Results with varying the order of ICL samples, where Ascending and Descending represent cases where examples closer to the query appear earlier and later in the LCLM context, respectively. In contrast, random denotes the case where examples are arranged randomly without a specific order.

Methods	Summarization	Translation	Reasoning	Classification
Random	0.310 ± 0.004	0.553 ± 0.004	0.650 ± 0.023	0.539 ± 0.007
Ascending	0.307 ± 0.006	0.557 ± 0.004	0.641 ± 0.027	0.534 ± 0.010
Descending	0.309 ± 0.003	0.552 ± 0.007	0.648 ± 0.021	0.539 ± 0.005

2.3 Experimental Results

Results on Sample Selection Strategies We report the detailed results of various sample selection approaches in many-shot ICL scenarios in Figure 2. To rigorously evaluate each sample selection approach and their statistically significant gains, we conduct a t-test with a 95% confidence threshold and report the results in Table 1. From these results, we observe that previously effective sample selection methods, designed for shorter context LMs, yield little to no performance gains over the random selection approach when applied to LCLMs. Aggregated results across three different LCLMs indicate statistical significance in fewer than 15% of instances, indicating that they are not reliable.

Analysis on Number of ICL Examples To see the performance of ICL with respect to the number of examples, we visualize results in Figure 3. Overall, for any sampling method, we observe that performance increases as the number of examples increases. Also, when the number of examples is relatively small, the relevance-based sample selection approach performs particularly well, as focusing on highly relevant examples maximizes learning effectiveness with the limited number on examples. However, as the number of examples increases, the performance gap between various sample selection methods diminishes, indicating that performance is less dependent on selection strategies in many-shot scenarios. Lastly, in the summarization task (where samples tend to be longer than those in other tasks), we observe an initial increase in performance, followed by a decline once the context becomes heavily populated with a large number of examples. We argue this decline likely reflects the challenges LCLMs face in processing extremely long contexts, discussed in Section 4.2.

Analysis on Converge of ICL Examples To further investigate why the performance gap between different approaches diminishes as the number of examples increases, we analyze the representational coverage of examples in-context relative to the full examples. Specifically, we measure the convex hull volume spanned by the embeddings of ICL examples (where we vary their numbers) and compare it to that of the entire dataset, which can serve as a proxy for how well the samples in-context capture the distribution of the full data. Our results, visualized in Figure 4, show that, when the number of ICL examples is moderate (e.g., 64), they already span over 80% of the convex hull volume of the full dataset in almost all tasks and datasets. This suggests that, beyond a certain threshold, adding more

Table 3: Results of LCLM-enabled ICL on four different tasks, where Random indicates the naive sample selection approach without selection criteria, Best Selection indicates the model that achieves the best performance among sophisticated sample selection methods for each experiment unit, and Augmentation indicates the proposed approach that generates demonstrations and uses them alongside original samples with random selection. We emphasize statistically significant results over Random in bold. We exclude Llama from the augmentation scenario as its context capacity is approximately ten times smaller than that of Gemini, allowing it to fully utilize its available context with the original examples alone, making augmentation unnecessary.

		Translation						Reasoning	
LCLMs	**Methods**	ENG to BEM	ENG to KMR	ENG to EWE	ENG to SPA	ENG to FRA	ENG to DEU	Date	Salient
Gemini Pro	Random	0.470 ± 0.003	0.439 ± 0.001	0.419 ± 0.004	0.580 ± 0.006	0.734 ± 0.002	0.676 ± 0.010	0.854 ± 0.009	0.776 ± 0.035
	Best Selection	0.470 ± 0.004	0.443 ± 0.004	0.418 ± 0.002	0.583 ± 0.004	**0.745** ± 0.005	0.676 ± 0.004	**0.896** ± 0.021	0.772 ± 0.017
	Augmentation	**0.487** ± 0.007	**0.469** ± 0.003	**0.437** ± 0.003	**0.595** ± 0.005	0.748 ± 0.007	0.694 ± 0.005	**0.927** ± 0.019	0.784 ± 0.018

		Reasoning		Classification					All
LCLMs	**Methods**	Tracking7	Web	Banking77	DialogRE	Discovery	FewNERD	GoEmotion	Average
Gemini Pro	Random	0.294 ± 0.029	0.675 ± 0.021	0.878 ± 0.002	0.661 ± 0.009	0.195 ± 0.007	0.568 ± 0.012	0.393 ± 0.007	0.574 ± 0.010
	Best Selection	0.311 ± 0.031	**0.700** ± 0.028	**0.886** ± 0.004	**0.709** ± 0.014	0.204 ± 0.011	0.569 ± 0.006	**0.413** ± 0.006	0.586 ± 0.011
	Augmentation	0.307 ± 0.031	**0.768** ± 0.040	**0.889** ± 0.004	**0.698** ± 0.010	**0.209** ± 0.009	0.574 ± 0.008	**0.428** ± 0.006	**0.601** ± 0.012

		Translation						Reasoning	
LCLMs	**Methods**	ENG to BEM	ENG to KMR	ENG to EWE	ENG to SPA	ENG to FRA	ENG to DEU	Date	Salient
Gemini Flash	Random	0.419 ± 0.006	0.427 ± 0.004	0.363 ± 0.002	0.573 ± 0.004	0.726 ± 0.004	0.666 ± 0.005	0.754 ± 0.022	0.682 ± 0.019
	Best Selection	0.421 ± 0.002	0.434 ± 0.002	0.360 ± 0.003	0.575 ± 0.002	0.732 ± 0.003	0.673 ± 0.001	0.777 ± 0.030	0.687 ± 0.015
	Augmentation	**0.436** ± 0.006	**0.460** ± 0.002	**0.378** ± 0.004	**0.594** ± 0.007	0.737 ± 0.010	0.676 ± 0.012	**0.804** ± 0.037	**0.714** ± 0.013

		Reasoning		Classification					All
LCLMs	**Methods**	Tracking7	Web	Banking77	DialogRE	Discovery	FewNERD	GoEmotion	Average
Gemini Flash	Random	0.256 ± 0.030	0.582 ± 0.033	0.868 ± 0.004	0.541 ± 0.008	0.065 ± 0.007	0.521 ± 0.006	0.362 ± 0.016	0.520 ± 0.011
	Best Selection	0.270 ± 0.031	0.566 ± 0.031	0.872 ± 0.006	0.547 ± 0.012	**0.083** ± 0.007	**0.532** ± 0.002	**0.385** ± 0.006	0.528 ± 0.010
	Augmentation	0.281 ± 0.035	0.609 ± 0.040	**0.880** ± 0.006	**0.578** ± 0.025	**0.090** ± 0.005	**0.537** ± 0.009	**0.392** ± 0.015	**0.544** ± 0.015

examples does not significantly improve coverage, as the selected examples, regardless of selection methods, can approximate the full data distribution.

Analysis on Example Order Previous work has shown that earlier LMs are sensitive to the order of examples when doing few-shot ICL. For example, LMs tend to follow the answer in the last example (Zhao et al., 2021; Lu et al., 2022). To investigate whether similar issues arise in many-shot ICL with LCLMs, we experiment by comparing performance when ordering ICL examples randomly, by increasing similarity, and by decreasing similarity. The results in Table 2 suggest that the order of examples does not affect performance of LCLMs.

Analysis on Computational Complexity In addition to performance, computational complexity is a critical factor to consider when assessing the practicality of many-shot ICL with LCLMs, as they often handle million-token contexts. We note that for approaches that adjust ICL examples based on the given query (such as relevance-based selection), the complexity scales quadratically, $\mathcal{O}(n^2)$, where n represents the number of tokens used for ICL demonstrations. In contrast, the simpler naive selection approach, which uses the same set of randomly selected examples for all queries, offers a significantly more efficient complexity of $\mathcal{O}(kn)$, where k is the number of tokens only within the target query ($n \gg k$). This is because the selected examples do not change based on the query; thus, the same set of examples can be key-value cached. As a result, random selection is a practical choice due to its equivalent performance with other selection methods and the added advantage of efficiency.

3 Augmenting ICL Demonstrations to Increase Context Capacity of LCLMs

3.1 ICL Example Augmentation Approach

Recall that recent advances in LCLMs offer unprecedented context capacity, potentially amplifying ICL performance by including more examples. However, the available examples sometimes fall short of filling this expanded capacity, and this under-utilization of the context may result in suboptimal performance. To address this, we introduce a simple yet effective ICL sample augmentation approach designed to increase the context capacity of LCLMs, while being scalable for many-shot scenarios. This method consists of synthetic example generation and low-quality example filtering.

Generation of Synthetic Examples Formally, let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{k}$ be a set of available ICL examples for a target task. The objective is to generate a set of synthetic examples $\mathcal{D}' = \{(\boldsymbol{x}'_j, \boldsymbol{y}'_j)\}_{j=1}^{m}$ (to supplement the original dataset $\mathcal{D}$), such that the augmented set of examples $\mathcal{D}_{\text{AUG}} = \mathcal{D} \cup \mathcal{D}'$ can increase the utilization of the available context capacity of LCLMs. To operationalize this, we generate each synthetic example $(\boldsymbol{x}'_j, \boldsymbol{y}'_j)$ by prompting an LM with randomly selected real examples from $\mathcal{D}$ as context, to ensure the generated data retains meaningful characteristics relevant to the task.

Filtering Out Low-Quality Examples Once the synthetic examples are generated, we filter out low-quality instances that may introduce noise or irrelevant information. To do this, we design a function f that assigns a quality score to each synthetic example $(\boldsymbol{x}'_j, \boldsymbol{y}'_j)$ based on its contextual relevance and alignment with real examples as well as over-

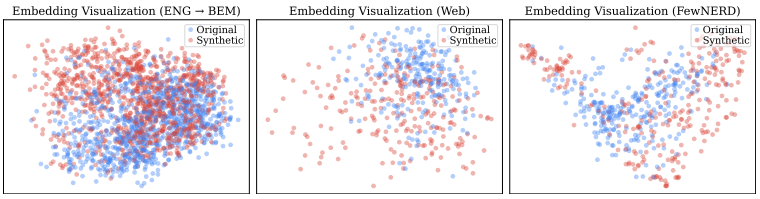

Figure 5: Visualization of embedding-space with original and synthetic examples.

Table 4: Results on Similarity (embedding-level similarity between original and synthetic examples) and Volume (relative expansion of the convex hull with augmented examples).

Tasks	Similarity	Volume
Translation	0.5715	1.6563
Reasoning	0.8099	3.2328
Classification	0.6252	2.7931

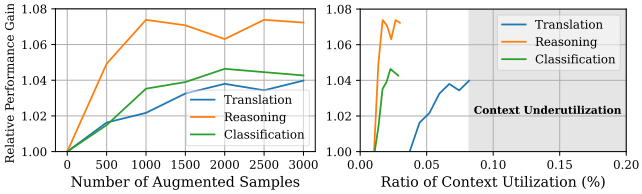

Figure 6: Results with augmented examples according to the size of synthetic samples (Left) and context utilization of Gemini Pro (Right).

Table 5: Results on ablation study, where w/o Filtering and w/o Original denote results based on augmented samples without filtering and without original samples, respectively. Only Original shows results without augmentation.

Methods	Translation	Reasoning	Classification
Augmentation	**0.571** ± 0.005	**0.696** ± 0.027	**0.560** ± 0.008
w/o Filtering	0.552 ± 0.005	0.666 ± 0.031	0.548 ± 0.009
w/o Original	0.544 ± 0.002	0.611 ± 0.025	0.531 ± 0.007
Only Original	0.553 ± 0.004	0.650 ± 0.023	0.539 ± 0.007

all quality. Specifically, each synthetic example is rated on a 5-point Likert scale by prompting the LM 30 times with the synthetic and 30 real examples. We then compute an aggregate score using a weighted average of scores with their corresponding probabilities from the LM. Only the synthetic examples that exceed the quality threshold, τ, are retained in the augmented example set, as follows:

$$\mathcal{D}_{\text{AUG}} = \mathcal{D} \cup \{(\boldsymbol{x}'_j, \boldsymbol{y}'_j) \mid f(\boldsymbol{x}'_j, \boldsymbol{y}'_j, \mathcal{D}) \geq \tau\}_{j=1}^{m}.$$

Notably, our data augmentation process is efficient, as it is performed offline and does not contribute to inference-time overhead. Also, it takes under 10 seconds per example, which can be done in parallel.

3.2 Experimental Setup

For synthetic data generation and filtering, we use Gemini Pro, one of the state-of-the-art LMs. We focus on tasks that underutilize the context capacity of LCLMs even when all available samples are provided, such as translation, reasoning, and classification. For each task, we generate 3,000 examples and retain only those with a quality score above the median among the generated samples. As a result, we use the original examples and 1,500 synthetic examples. The prompts used to elicit data generation and filtering are provided in Appendix A.

3.3 Experimental Results

Main Results As shown in Table 3, which compares the example augmentation approach (with random selection) to other sample selection strategies, the augmentation approach demonstrates substantial performance gains across various datasets, which can be attributed to the greater diversity and volume of ICL examples achieved through synthetic data generation, leading to the effective utilization of the context capacity of LCLMs. Also, like the random selection approach, our augmentation method allows the reuse of the same examples across all queries. Thus, due to key-value caching, the augmentation approach is as efficient as random selection while achieving superior performance.

Analysis on Augmented Data Beyond performance improvements, we analyze the characteristics of the augmented data to better understand its impact on ICL. First, as visualized in Figure 5, the embedding-space distribution of augmented examples closely follows that of real examples while expanding the overall data coverage, which suggests that the synthetic examples effectively capture task-relevant features without deviating substantially from the original data distribution. In addition, we further quantify this expansion through two metrics: the similarity between original and synthetic examples, and the relative expansion of the convex hull with augmented examples compared to that formed by original examples, and report results in Table 4. From this, we observe that while synthetic examples maintain a high degree of similarity to real examples (ensuring alignment with the task), they also significantly increase the volume of the data distribution. This balance between relevance and diversity highlights why our augmentation approach effectively enhances ICL performance.

Finally, we analyze the impact of the number of augmented examples on performance and their corresponding context utilization in LCLMs. As shown in Figure 6, while increasing the number of synthetic examples initially improves performance, it eventually plateaus, indicating diminishing returns. Also, despite augmentation improving context utilization, we find that even at peak performance, the augmented data occupies less than 3% of the full context capacity of LCLMs, which is significantly below the scale that LCLMs can handle (Figure 8). These suggest an interesting future work to develop more advanced augmentation strategies to increase the context utilization of LCLMs.

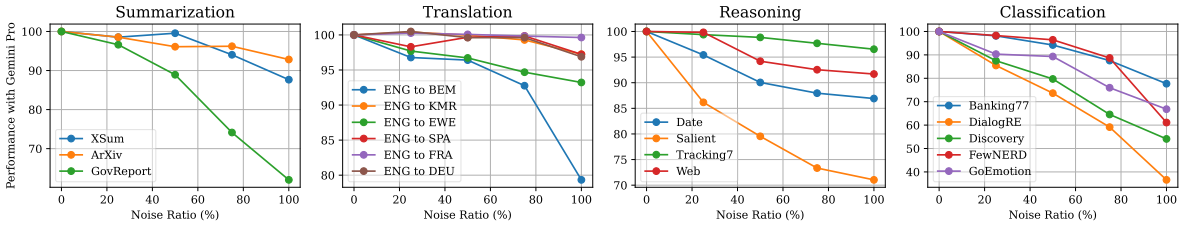

Figure 7: Results with varying the ratio of noisy examples within the context of LCLMs, where we report the relative performance over the ICL without noisy examples (i.e., the noise ratio of 0) and the results are averaged over multiple runs.

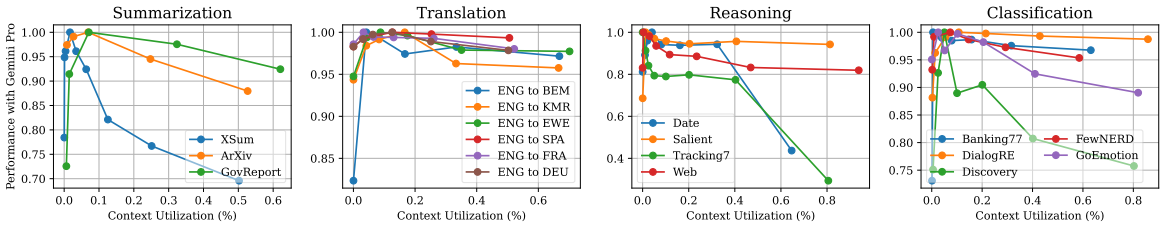

Figure 8: Results across different percentages of context size utilized in LCLMs, where the x-axis represents the percentage of the full LCLM context used (according to the number of tokens over the full token length), and the y-axis shows the relative performance compared to the highest performance achieved for each dataset. Results are averaged over multiple runs.

Ablation Study on Augmentation To see how each component in the augmentation approach contributes to performance gains, we conduct an ablation study. As shown in 5, we observe that the full augmentation method (called Augmentation), which uses both original and filtered synthetic examples, achieves the best performance. In contrast, when the filtering step is omitted, performance decreases, indicating that filtering contributes positively by removing lower-quality examples. Also, a large performance drop occurs when original samples are excluded from the augmented set. This suggests that although filtering helps maintain quality, the synthetic samples generated still do not match the quality of the original examples. Thus, while our augmentation approach is effective, further research could improve data generation techniques to improve the quality of the synthetic examples.

4 Behaviors of LCLM-Enabled ICL

4.1 LCLM-Based ICL with Noisy Examples

LCLMs can accommodate a large number of diverse ICL examples, which raises the question of the impact and risk of including noisy examples in the context. We investigate how the performance of LCLM-enabled ICL is impacted when some or all of the ICL examples are noisy. To simulate noisy examples, we modify the outputs of a subset of in-context demonstrations by replacing their outputs with outputs from other randomly selected demonstrations. As shown in Figure 7, LCLM-enabled ICL is largely robust to noise when the proportion of noisy examples is relatively low (i.e., below 25%). This observation highlights why augmented examples, even if slightly lower quality, can still enhance performance as it increases the utilization of the context window. In contrast, when the amount of noise exceeds this threshold, LCLMs become vulnerable to the negative effects of noise and the performance notably declines. This adverse effect is more pronounced for challenging tasks, such as low-resource translation (e.g., English to Bemba or Ewe). This is likely because LCLMs are less familiar with those tasks, and therefore rely more on learning from in-context examples.

4.2 LCLM-Based ICL with Long Context

As the context length capacity of LCLMs continues to grow, it becomes increasingly important to assess whether LCLMs can reliably utilize a large number of ICL examples. To investigate this, we conduct an experiment analyzing the performance as a function of the context utilization. Specifically, we gradually increase the number of examples by powers of two, and if the entire set of examples within the dataset is used, we further extend the context utilization by repeating these examples. The hypothesis being tested is that if LCLMs can effectively understand and utilize extremely long context, performance should remain consistent even with repeated examples, as the presence of duplicates should not impact contextual understanding. However, as shown in Figure 8, a substantial performance decline occurs when LCLMs are pushed to use extremely large contexts. Specifically, this decline generally begins when more than 25% of the available context capacity is utilized. Also, the performance drop is pronounced in tasks such as xsum, which requires generating abstractive summaries (unlike other summarization datasets like arXiv or GovReport) and in tasks demanding complex reasoning such as date understanding (Date) and

object tracking (Tracking7). These findings suggest that while LCLMs can handle moderately long contexts, they encounter limitations with exceedingly large contexts, particularly in tasks requiring fine-grained reasoning or abstractive generation. This may be due to challenges in distinguishing and integrating relevant information across numerous examples, especially when tasks require high levels of nuanced abstraction and precise reasoning.

5 Related Work

LCLMs The field of language modeling has witnessed remarkable advancements with Language Models (LMs) (Brown et al., 2020; OpenAI, 2023; Reid et al., 2024; Dubey et al., 2024). However, earlier LMs were constrained by relatively short context windows, typically handling only a few thousand tokens at a time, which limits their applicability in advanced tasks requiring broader context comprehension, such as document-level summarization or complex reasoning (Koh et al., 2023; Suzgun et al., 2022). To address this, recent efforts have led to the development of LCLMs, designed to process much larger contexts, sometimes accommodating over a million tokens within a single prompt (Reid et al., 2024). To mention a few, models like Longformer and BigBird (Beltagy et al., 2020; Zaheer et al., 2020) incorporate sparse attention mechanisms to efficiently handle extended contexts without compromising on computational feasibility. Also, LongRoPE extends the context window of LMs to 2M tokens by interpolating their specific positional embeddings (Ding et al., 2024).

In-Context Learning ICL is a recent paradigm that enables LMs to learn from examples in-context and perform given tasks (Brown et al., 2020; Min et al., 2022; von Oswald et al., 2023). Since its introduction, previous studies have concentrated on developing the strategies to optimize the quality and arrangement of in-context examples to maximize performance, especially given the limitations of early LMs on context length. For example, these approaches include selecting examples that maximize relevance to the target query (Liu et al., 2021; Rubin et al., 2022), ensuring diversity among examples to cover a range of possible cases (Sorensen et al., 2022; An et al., 2023), strategically ordering examples to improve model adaptation (Zhao et al., 2021; Lu et al., 2022), and prioritizing examples by their ease of learning based on their difficultly (Mavromatis et al., 2023; Liu et al., 2024). Yet, as the context capacity expands with LCLMs, these conventional selection strategies warrant re-evaluation, particularly in many-shot settings.

Many-Shot ICL Early approaches in many-shot ICL have primarily focused on the paradigm shift brought by the ability to incorporate a larger number of examples in-context (Agarwal et al., 2024; Bertsch et al., 2024), without giving much consideration to example selection strategies. Such many-shot ICL methods have demonstrated performance comparable to fine-tuning. Also, there is a very recent work that explores retrieval strategies in many-shot ICL (Bertsch et al., 2024); however, they use models with relatively limited context capacities (e.g., under 100k tokens with Llama 2), resulting in restrictions on the number of examples included and, consequently, making retrieval-based methods appear more advantageous. However, contrary to this finding, we uncover that this advantage diminishes as the context capacity increases, allowing random sampling to perform on par with more sophisticated selection methods when a large number of examples is used. Lastly, other recent efforts include establishing benchmarks for long-context ICL (Lee et al., 2024a; Li et al., 2024). Unlike prior studies, our work offers a novel perspective by systematically re-evaluating traditional selection strategies in the expanded context regime and highlighting the shift from selection optimization to effectively leveraging the extensive context space in many-shot ICL, with data augmentation.

6 Conclusion

We explored ICL in the context of LCLMs, investigating whether traditional sample selection strategies remain effective in many-shot scenarios and observing that they offer minimal to zero performance gains over simple random selection. We also highlighted the emerging challenge of underutilized context in low-resource tasks due to limited example availability, and proposed a data augmentation strategy, which substantially boosts performance by increasing context utilization of LCLMs. Lastly, we analyzed the behavior of LCLM-enabled ICL when operating with extremely long context and in the presence of noisy examples, and found that while performance improves with added examples, it plateaus and even declines when the context becomes too long, with increased vulnerability to noise in complex tasks. This suggests promising future directions in making LCLMs more robust to lengthy context and noise examples alongside the direction of extending their context length.

Limitations

While this work explores the new opportunity of ICL with LCLMs, a couple of limitations can be considered. First, the computational cost associated with LCLMs remains a significant challenge, particularly for researchers and practitioners in resource-constrained settings. Second, while the proposed data augmentation method enhances context utilization of LCLMs and improves ICL performance, the quality of synthetic examples often falls short of the quality of original data. Addressing them through cost-efficient strategies for leveraging LCLMs and developing improved data augmentation techniques would be an exciting area for future work.

Ethics Statement

We believe this work does not raise any direct ethical concerns, as it primarily focuses on advancing the understanding of ICL with LCLMs. However, as with any other application of LCLM-based ICL, careful consideration must be given to the quality of the examples used in the context. Specifically, the inclusion of biased, harmful, or otherwise problematic examples in the input context can propagate or amplify these issues in the model's outputs, and we advise practitioners to carefully evaluate and select ICL examples to avoid potential issues.

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. *ArXiv*, abs/2404.11018.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? 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 11027–11052. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

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. *ArXiv*, abs/2405.00200.

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*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, W. Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *North American Chapter of the 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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv*, abs/2301.00234.

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. *ArXiv*, abs/2407.21783.

Luyang Huang, Shuyang Cao, Nikolaus Nova Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1419–1436. Association for Computational Linguistics.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2023. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8):154:1–154:35.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, S'ebastien M. R. Arnold, Vincent Perot, Sid Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2024a. Can long-context language models subsume retrieval, rag, sql, and more? *ArXiv*, abs/2406.13121.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernández Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha R. Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024b. Gecko: Versatile text embeddings distilled from large language models. *ArXiv*, abs/2403.20327.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *ArXiv*, abs/2404.02060.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *ArXiv*, abs/2402.10738.

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.

Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. 2023. Which examples to annotate for in-context learning? towards effective and efficient selection. *ArXiv*, abs/2310.20046.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

NLLB, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

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

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub,

Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

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.

Taylor Sorensen, Joshua Robinson, Christopher Michael Rytting, Alexander Glenn Shaw, Kyle Jeffrey Rogers, Alexia Pauline Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. 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 819–862. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Huai hsin Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Annual Meeting of the Association for Computational Linguistics*.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Table 6: Results of LCLM-enabled ICL on reasoning datasets with and without Chain-of-Thought (CoT) (Wei et al., 2022).

Methods	Date	Salient	Tracking7	Web
Many-Shot ICL	0.927	0.784	0.307	0.768
Many-Shot ICL with CoT	0.918	0.810	0.334	0.771

A Prompts

We provide the prompts used for many-shot ICL on translation, summarization, and reasoning tasks in Table 7 and on classification tasks in Table 8. Also, we provide the prompts used for synthetic data augmentation and filtering in Table 9.

B Detailed Experimental Setup

Configuration For all experiments, we use the default hyperparameters for Gemini and Llama.

Ratio of Augmented Data We use original examples alongside 1,500 synthetic samples (filtered from an initial set of 3,000 examples according to their quality scores); therefore, the percentage of augmented samples varies depending on the size of the original examples in each dataset. Specifically, for the translation task where there are around 1,000 original examples, synthetic samples comprise around 60% of the total examples. For reasoning tasks (having around 100 to 150 examples), synthetic samples constitute 90-94% of the total examples. For the classification task (e.g., Banking77 dataset), with 385 original examples, synthetic samples account for around 80% of the total examples.

C Detailed Experimental Results

Results with CoT It is worth noting that while developing the approach to better utilize many examples within the expanded context windows of LCLMs with advanced prompting techniques, such as Chain-of-Thought (CoT) (Wei et al., 2022), represents an orthogonal but promising future research direction, as an initial foray into this area, we perform experiments with CoT on the reasoning task (as it may benefit from explicit step-by-step thinking procedures) and report results in Table 6. From this, we then observe that the CoT prompting strategy improves the performance on most datasets (except for Date whose performance is already high without CoT), demonstrating that there may be a potential to enhance the performance of LCLM-enabled many-shot ICL via advanced prompting.

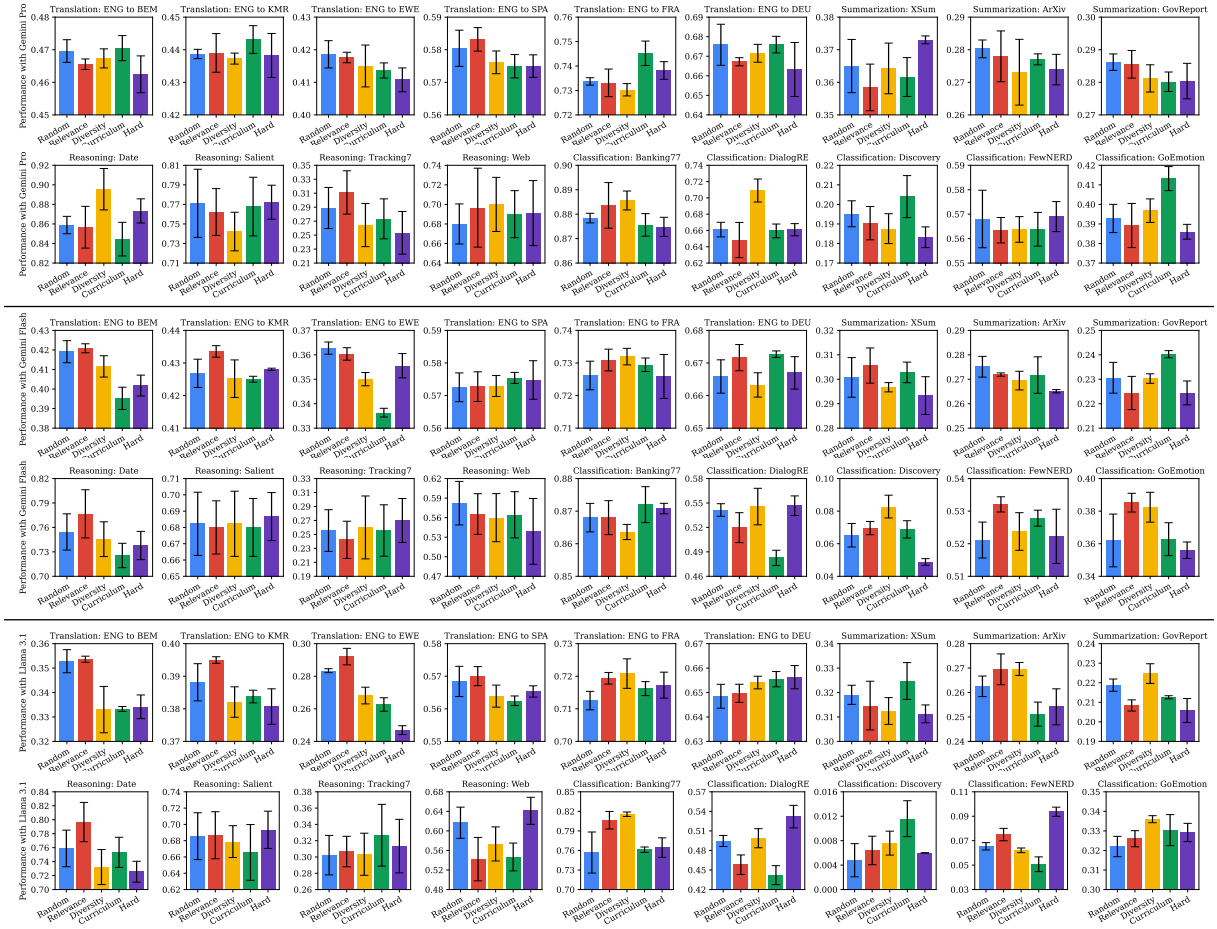

Figure 9: Detailed results of various sample selection approaches on ICL with LCLMs, such as Gemini Pro (Top), Gemini Flash (Middle), and Llama 3.1 (Bottom), across four different tasks (translation, summarization, reasoning, and extreme classification) with 18 datasets. Each bar represents the averaged performance, with the upper and lower limits indicating standard deviation.

Table 7: A list of prompts that we use for many-shot ICL on translation, summarization, and reasoning tasks.

Types	Prompts
Translation	You are an expert translator. I am going to give you one or more example pairs of text snippets where the first is in {SOURCE_LANGUAGE} and the second is a translation of the first snippet into {TARGET_LANGUAGE}. The sentences will be written as the following format: {SOURCE_LANGUAGE}: <first sentence> {TARGET_LANGUAGE}: <translated first sentence> After the example pairs, I am going to provide another sentence in {SOURCE_LANGUAGE} and I want you to translate it into {TARGET_LANGUAGE}. Give only the translation, and no extra commentary, formatting, or chattiness. Translate the text from {SOURCE_LANGUAGE} to {TARGET_LANGUAGE}. {EXAMPLES} {TARGET_QUERY}
Summarization	You are an expert in article summarization. I am going to give you one or more example pairs of article and its summary in fluent English. The pairs will be written as the following format: Article: <article> Summary: <summary> After the example pairs, I am going to provide another article and I want you to summarize it. Give only the summary, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}
Reasoning	You are an expert in multiple-choice question answering tasks. I am going to give you one or more example pairs of question and its answer in a multiple-choice question answering format. The pairs will be written as the following format: Question: <question> Answer: <answer> After the example pairs, I am going to provide another question and I want you to predict its answer. Give only the answer that follows a consistent format as in the provided examples, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}

Table 8: A list of prompts that we use for many-shot ICL on five different extreme classification tasks.

Types	Prompts
BANKING77	I am going to give you one or more example pairs of customer service query and its intent. The pairs will be written as the following format: service query: <query> intent category: <category> After the example pairs, I am going to provide another customer service query and I want you to classify the label of it that must be one among the intent categories provided in the examples. Give only the category, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}
DialogRE	I am going to give you one or more examples of the dialogue, the list of entity pairs within it, and their corresponding relation types. The examples will be written as the following format: Dialogue: <dialogue> The list of k entity pairs are (<entity 1>, <entity 2>), ... The k respective relations between each entity pair are: <relation>, ... After the examples, I am going to provide another dialogue along with its associated entity pairs, and I want you to classify their corresponding relation types that must be one among the relation types provided in the examples. Give only the relations, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}
Discovery	I am going to give you one or more example pairs of two sentences and the conjunction word between them. The pairs will be written as the following format: <sentence 1> ( ) <sentence 2> the most suitable conjunction word in the previous ( ) is <conjunction word> After the example pairs, I am going to provide another two sentences and I want you to classify the conjunction word between them that must be one among the conjunction words provided in the examples. Give only the conjunction word, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}
FewNERD	I am going to give you one or more examples of the sentence, the named entities within it, and their corresponding entity types. The examples will be written as the following format: Sentence: <sentence> <named entity>: <entity type> After the example pairs, I am going to provide another comment and I want you to classify the label of it that must be one among the emotion categories provided in the examples. Give only the category, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}
GoEmotion	I am going to give you one or more example pairs of comment and its emotion category. The pairs will be written as the following format: comment: <comment> emotion category: <category> After the example pairs, I am going to provide another sentence, and I want you to classify the named entities within it and their corresponding entity types that must be one among the entity types provided in the examples. Give only the named entities and their corresponding entity types, and no extra commentary, formatting, or chattiness. {EXAMPLES} {TARGET_QUERY}

Table 9: A list of prompts that we use for generating synthetic demonstrations and filtering them of low-quality.

Types	**Prompts**
Generation	You are an expert in data augmentation. You will be provided with a series of demonstrations that show how a task is performed. Your objective is to generate a new example that closely follows the pattern, structure, and style of the demonstrations. Carefully analyze the key steps, transitions, and output style in the provided demonstrations. Then, create a new sample that maintains consistency in format and correctness while introducing variety in content. Here are the demonstrations: {EXAMPLES} Now, as an expert, generate a new sample that aligns with the original demonstrations:
Filtering	You are an expert in assessing data quality. Given the original set of samples, your task is to carefully evaluate the provided sample in comparison to the original samples. Based on your expertise, determine whether the provided sample is of high quality, meeting or exceeding the standards set by the original set. Here are the original samples: {EXAMPLES} Now, as an expert, evaluate the provided sample: {GENERATED_SAMPLE} Please provide only a single numerical rating (1, 2, 3, 4, or 5) based on the quality of the sample, without any additional commentary, formatting, or chattiness.