Effective Demonstration Annotation for In-Context Learning via Language Model-Based Determinantal Point Process

Peng Wang[♠], **Xiaobin Wang**[♡], **Chao Lou**[♢], **Shengyu Mao**[♠], **Pengjun Xie**[♡], **Yong Jiang**[♡*]
[♠]Zhejiang University, [♡]Alibaba Group, [♢]ShanghaiTech University
peng2001@zju.edu.cn, yongjiang.jy@alibaba-inc.com

Abstract

In-context learning (ICL) is a few-shot learning paradigm that involves learning mappings through input-output pairs and appropriately applying them to new instances. Despite the remarkable ICL capabilities demonstrated by Large Language Models (LLMs), existing works are highly dependent on large-scale labeled support sets, not always feasible in practical scenarios. To refine this approach, we focus primarily on an innovative selective annotation mechanism, which precedes the standard demonstration retrieval. We introduce the **L**anguage **M**odel-based **D**eterminant **P**oint **P**rocess (LM-DPP) that simultaneously considers the uncertainty and diversity of unlabeled instances for optimal selection. Consequently, this yields a subset for annotation that strikes a trade-off between the two factors. We apply LM-DPP to various language models, including GPT-J, LlaMA, and GPT-3. Experimental results on 9 NLU and 2 Generation datasets demonstrate that LM-DPP can effectively select canonical examples. Further analysis reveals that LLMs benefit most significantly from subsets that are both low uncertainty and high diversity.

1 Introduction

As large pre-trained language models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022a; Tay et al., 2023; Touvron et al., 2023; Workshop, 2023) grow in scale, they not only exhibit enhanced linguistic capabilities and expanded world knowledge but also demonstrate a novel ability for in-context learning. Specifically, LLMs have shown proficiency in learning from a limited set of input-output examples (known as demonstrations (Brown et al., 2020)), and effectively applying these learned mappings to new, unseen instances. This novel few-shot learning paradigm, which avoids parameter updates, has become a popular and efficient method for utilizing LLMs (Liu et al., 2021b; Dong et al., 2023; Liu et al., 2021a).

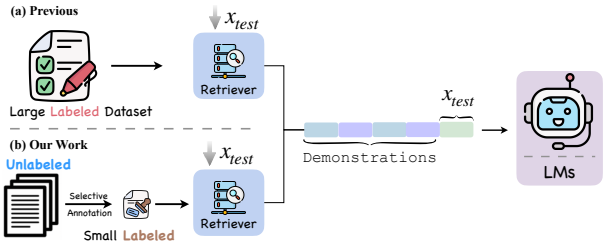

Figure 1: Left (**Step 1**): Without assuming access to a large amount of labeled data, we employ active data collection, selectively annotating demonstration examples. Right (**Step 2**): Prompt construction and model inference.

Previous studies have investigated which instances can serve as effective prompts for ICL (Liu et al., 2021a; Zhang et al., 2022b; Li and Qiu, 2023). They have demonstrated that retrieving specific similar contexts for individual test queries can significantly improve performance (`instance level`) and ground truth matters for support examples. To assign appropriate demonstrations to all test queries, support sets necessitate diversity and broad coverage, usually achieved through large labeled data, following the principle that Monte Carlo estimation accuracy improves with larger samples. Nonetheless, these extensive datasets are often impractical to obtain.

We investigate the selection of demonstrations from the perspective of Active Learning (AL) (Cohn et al., 1996; Settles, 2009). Based on the core principle that not all data points are of equal value, AL aims to identify the most effective instances in an unlabeled data pool for annotation. Margatina et al. (2023) elucidates that high semantic similarity, low uncertainty, and high diversity comprise an effective and efficient annotation strategy. Similarly, Gonen et al. (2022) demonstrates

[*] Corresponding Author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1266–1280
November 12-16, 2024

that lower prompt perplexity is closely associated with better performance. While Su et al. (2022)'s Vote-k framework adopts a data-centric perspective (i.e., selecting examples that balance diversity and representativeness), it neglects the assessment of uncertainty and the inter-relationship among context examples. In this paper, we pursue a more universally applicable yet straightforward solution, incorporating confidence signals of LLMs to select annotation instances that are maximally diverse and exhibit low uncertainty.

To address this need, we introduce a generic approach, LM-DPP, which jointly models uncertainty and diversity within the support set through a conditional Determinantal Point Process. Specifically, we employ LLMs' perplexity to score each candidate instance in the support set, which serves as a measure of the LLMs' uncertainty. Then a Gram matrix is constructed to balance the uncertainty and diversity of candidate instances and polynomial-time maximum a posteriori (MAP) inference (Chen et al., 2018) is applied to identify the most useful subset of instances to be annotated. From the perspective of selective annotation, we consider extremely low-resource ICL scenarios as those in which the available annotated examples are limited to a few dozen instances. Our focus centers on identifying which specific set of demonstrations can most effectively harness the capabilities of LLMs within this challenging context.

We validate our method through extensive experiments on 9 NLU and 2 Generation datasets. We also demonstrate the versatility of LM-DPP by adapting it to the large language model GPT-3 (175B). The experimental results illustrate that our approach can effectively balance two critical factors, uncertainty and diversity. In summary, our contributions are as follows.

- We revisit the setup of ICL from the perspective of selective annotation. We introduce a novel approach, **LM-DPP**, to select instances that balance uncertainty and diversity for annotation, aiming to reduce the human engineering workload.
- The experimental results indicate that the proposed method outperforms the previous best-performing selection methods by a large relative improvement and exhibits commendable generalizability across model size (§4.2) and annotation budget (§4.3) scaling.
- Comprehensive analysis confirms that LLMs can benefit from a demonstration set that exhibits both low uncertainty and diversity (§4.1) and gold annotation matters for ICL performance (§5.2).

2 Methodology

In this section, we introduce technical details of LM-DPP for selecting annotation instances exhibiting both high diversity and low uncertainty. Formally, given a set of unlabeled samples $\mathcal{X} = \{x_i\}_{i=1}^{N}$, LM-DPP aims to select a subset $\mathcal{L} \subset \mathcal{X}$ for annotation, where $|\mathcal{L}| = M$ is the annotation budget, such that the Language Models (LLMs) maintains high ICL performance on the test set $\mathcal{D}_{\text{test}}$. As shown in Figure 2, given a Pre-trained Language Model (PLM) G, we first score candidate instances x_i using the perplexity of the LLMs (§2.1). We then compute vector representations for the candidate instances, utilizing a conditional kernel matrix to balance diversity and low uncertainty (§2.2). Subsequently, we perform a greedy MAP inference algorithm to filter the candidate annotation set (§2.3).

2.1 Uncertainty

As off-the-shelf LLMs do not contain a classification head fine-tuned for specific tasks, calculating entropy, a common measure of uncertainty used in AL, across all possible outputs is challenging, if not unfeasible. Alternatively, we adopt the SPELL method proposed by (Gonen et al., 2022), using the perplexity of the LLMs, to score candidate examples $\widetilde{x}$. The scoring function $r(\widetilde{x})$ is defined as:

$$r(\widetilde{x}) = \frac{1}{PPL(\widetilde{x})} = \exp\left(\frac{1}{t}\sum_{i=1}^{t}\log G_\theta(\widetilde{x}_i|\widetilde{x}_{<i})\right) \quad (1)$$

Recent research also delineates that LLMs are essentially a form of lossless data compression (Delétang et al., 2023), and perplexity, serving as a proxy for the occurrence of the prompt in some form in the training data, inherently indicates the model's expectancy of the prompt. Therefore, perplexity-based demonstration selection can, to some extent, avoid LLM sampling from low-frequency distributions. We also conduct pilot experiments (Appendix B) that select instances of high uncertainty, observing a substantial decrease in performance.

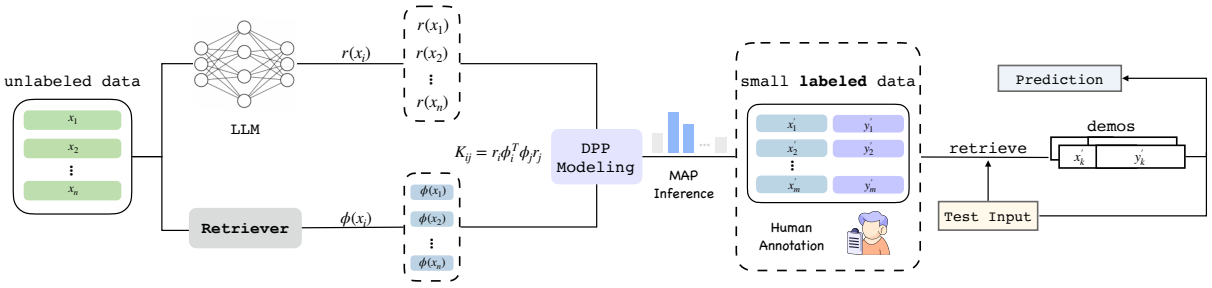

Figure 2: An illustration of our proposed approach. There are three steps in LM-DPP: (1) Estimate the perplexity for each unlabeled data point, with the reciprocal denoted as $r(x_i)$. (2) Employ conditional DPP to jointly model uncertainty and diversity, selecting a small set of examples for annotation before test time. (3) At test time, the context is constructed by retrieving relevant examples from the small annotated pool.

2.2 DPP Modeling

We consider similarity as the primary qualitative feature of the DPP diversification process. In this section, we present the decomposition of DPP that more directly elucidates the tension between diversity and the uncertainty measure for each candidate instance. Since the DPP kernel, L is typically written as a Gram matrix, $L = B^T B$, where the columns of B represent vectors from the candidate set $\mathcal{X}$. We define B_i as the product of the LLMs uncertainty term $r_i \in \mathbb{R}^+$ and the normalized diversity feature vector $\phi_i \in \mathbb{R}^D$, with $|\phi_i| = 1$. The new DPP kernel matrix can now be written as $K_{ij} = r_i \phi_i^T \phi_j r_j = r_i r_j \langle \phi_i^T \phi_j \rangle$ (Ye et al., 2023). r_i can be regarded as the intrinsic evaluation of the LLMs for the candidate instance and $\langle \phi_i^T \phi_j \rangle$ as the measure of similarity between instances x_i and x_j. Therefore, we arrive at $L = \text{Diag}(r) \cdot \phi \cdot \text{Diag}(r)$, and the unnormalized log probability for the subset S is $\log\det(L_S) = \sum_{i \in S} \log(r_i^2) + \log\det(\phi_S)$. To adjust the trade-off between uncertainty and diversity, we introduce a balancing parameter λ, thus modifying the log probability of L_S to:

$$\log\det(L_S)' = \lambda \cdot \sum_{i \in S} r_i + (1 - \lambda) \cdot \log\det(L_S) \tag{2}$$

This corresponds to a DPP with kernel $L' = \text{Diag}(\exp(\alpha r)) \cdot \phi \cdot \text{Diag}(\exp(\alpha r))$, where $\alpha = \lambda/(2(1-\lambda))$. In Equ. (2), the first term corresponds to the low perplexity of the selected instances, while the second term increases with the diversity of the selected instances. Without the diversity model, we would choose examples of low uncertainty, but the DPP would tend to repeatedly select similar examples. Without the low uncertainty model, although we could obtain a highly diverse set, we might fail to include in S those examples most favorable to the LLMs. By combining them, we can achieve a more balanced outcome.

2.3 Inference

The solution to the MAP for DPP, which is to find the set of examples with the highest probability, is a complex process and an NP-hard problem. (Chen et al., 2018) have proposed an improved greedy algorithm that can quickly solve it approximately. In specific, this algorithm greedily selects the demonstration from the candidate set that maximizes the marginal gain to be added to the final result subset, until the stopping condition is satisfied. That is, each time an example j is chosen to be added to the candidate set S_{map}, which is initialized as an empty set. The formalization is as follows:

$$j = \arg \max_{j \in \mathcal{X} \setminus S_{map}} \log\det(L_{S_{map} \cup \{j\}}) - \log\det(L_{S_{map}}) \tag{3}$$

By performing a Cholesky decomposition on $L_{S_{\text{map}}}$, and incrementally updating the Cholesky factor, the complexity of solving $\det(L_{S_{\text{map}}})$ can be reduced from $O(K^3)$ to $O(K)$. Therefore, the complexity of each iteration is $O(NK)$. This implies that it is possible to return K annotation examples within $O(NK^2)$ time. Once we have selected and annotated a subset of examples$\mathcal{L}$ from the **unlabeled** support set, following recent work (Liu et al., 2021a), we retrieve examples from $\mathcal{L}$ that are semantically similar to the test query samples. We use `Sentence-BERT` (Reimers and Gurevych, 2019) representations for $\mathcal{L}$ and $\mathcal{D}_{\text{test}}$ again and employ cosine similarity as the metric. The underlying principle is that demonstrations most similar to the test example will best assist the model in answering the query. For the order of demonstrations, we adhere to the configuration established by Su et al. (2022), where the order of the retrieved

Model	Budget	Method	Natural Language Inference				Classification			Multi-Choice		Avg
			RTE	MNLI	MRPC	QNLI	SST-5	DBpedia	TREC	Hellaswag	COPA	
GPT-J 6B	$\|\mathcal{L}\| = 16$	Random	$48.24_{3.1}$	$40.92_{3.0}$	$\mathbf{64.75}_{5.0}$	$51.86_{3.5}$	$46.49_{3.6}$	$82.72_{7.7}$	$56.94_{16.1}$	$67.77_{1.5}$	$83.11_{2.0}$	$60.31_{6.6}$
		Kmeans	$46.58_{2.6}$	$39.84_{1.0}$	$59.48_{8.6}$	$51.47_{2.1}$	$41.80_{4.7}$	$88.77_{0.8}$	$68.46_{3.5}$	$66.90_{2.2}$	$83.40_{1.3}$	$60.74_{3.8}$
		Vote-k	$47.86_{0.9}$	$40.04_{2.9}$	$59.96_{7.3}$	$51.37_{3.9}$	$40.24_{3.7}$	$89.26_{3.5}$	$72.07_{7.9}$	$68.56_{2.9}$	$83.40_{1.6}$	$61.42_{4.4}$
		Fast Vote-k	$48.34_{0.7}$	$39.26_{3.9}$	$58.89_{5.0}$	$50.39_{1.7}$	$\mathbf{50.80}_{5.8}$	$\mathbf{89.65}_{3.4}$	$75.10_{5.5}$	$67.38_{3.8}$	$83.10_{0.8}$	$62.54_{3.8}$
		LM-DPP (ours)	$\mathbf{49.81}_{1.5}$	$40.92_{1.7}$	$64.36_{1.4}$	$\mathbf{52.96}_{2.0}$	$47.66_{5.0}$	$89.06_{3.0}$	$\mathbf{75.20}_{2.6}$	$\mathbf{69.44}_{2.6}$	$\mathbf{83.60}_{2.1}$	$\mathbf{63.67}_{2.6}$
	$\|\mathcal{L}\| = 100$	Random	$47.64_{2.2}$	$39.41_{2.8}$	$63.59_{3.1}$	$51.11_{3.5}$	$47.43_{0.9}$	$90.30_{1.5}$	$76.36_{1.3}$	$67.88_{0.8}$	$\mathbf{84.03}_{1.7}$	$63.08_{2.2}$
		Kmeans	$48.22_{0.5}$	$41.74_{3.8}$	$64.40_{5.0}$	$51.52_{3.1}$	$46.18_{1.6}$	$90.55_{1.7}$	$77.09_{5.6}$	$67.63_{0.5}$	$83.30_{1.8}$	$63.40_{3.1}$
		Vote-k	$49.12_{1.3}$	$40.26_{2.9}$	$61.24_{4.1}$	$50.62_{3.1}$	$47.85_{1.2}$	$86.92_{2.0}$	$\mathbf{82.18}_{2.5}$	$67.79_{1.8}$	$82.12_{2.8}$	$63.12_{2.6}$
		Fast Vote-k	$51.93_{4.1}$	$39.53_{4.2}$	$65.73_{1.2}$	$50.41_{2.6}$	$49.39_{0.9}$	$\mathbf{91.60}_{2.1}$	$81.45_{5.4}$	$68.23_{1.0}$	$83.84_{3.9}$	$64.68_{3.2}$
		LM-DPP (ours)	$\mathbf{54.44}_{2.6}$	$\mathbf{42.31}_{2.4}$	$\mathbf{67.10}_{1.3}$	$\mathbf{53.26}_{1.5}$	$\mathbf{49.62}_{1.0}$	$91.03_{2.2}$	$82.01_{3.2}$	$\mathbf{68.92}_{1.5}$	$83.80_{1.7}$	$\mathbf{65.83}_{2.0}$
LLAMA-2 7B	$\|\mathcal{L}\| = 16$	Random	$54.70_{1.4}$	$38.81_{1.4}$	$60.42_{1.9}$	$53.03_{2.1}$	$54.10_{4.1}$	$86.82_{6.0}$	$67.48_{14.4}$	$77.25_{2.1}$	$88.58_{2.5}$	$64.57_{5.6}$
		Kmeans	$54.88_{1.3}$	$36.62_{4.9}$	$60.94_{8.0}$	$52.54_{1.8}$	$53.32_{2.7}$	$90.04_{1.8}$	$76.95_{8.4}$	$77.25_{2.1}$	$\mathbf{89.06}_{1.4}$	$65.73_{4.5}$
		Vote-k	$52.83_{0.5}$	$\mathbf{41.21}_{4.8}$	$62.89_{1.3}$	$\mathbf{55.57}_{0.4}$	$53.42_{2.6}$	$87.79_{1.6}$	$\mathbf{79.10}_{2.5}$	$77.24_{2.4}$	$87.70_{1.3}$	$66.42_{2.3}$
		Fast Vote-k	$52.25_{1.2}$	$38.28_{4.0}$	$59.67_{4.4}$	$53.13_{1.7}$	$53.32_{4.3}$	$88.28_{1.8}$	$75.46_{4.7}$	$77.15_{2.9}$	$88.48_{1.9}$	$65.11_{3.3}$
		LM-DPP (ours)	$\mathbf{58.99}_{3.5}$	$38.28_{5.6}$	$\mathbf{63.09}_{4.5}$	$53.81_{2.6}$	$\mathbf{55.37}_{3.3}$	$\mathbf{93.65}_{1.5}$	$76.28_{4.5}$	$77.25_{1.2}$	$88.67_{1.1}$	$\mathbf{67.26}_{3.5}$
	$\|\mathcal{L}\| = 100$	Random	$58.01_{1.2}$	$39.85_{5.1}$	$60.48_{4.0}$	$51.66_{1.9}$	$54.50_{1.6}$	$92.87_{1.2}$	$83.69_{2.6}$	$76.76_{3.1}$	$87.91_{1.2}$	$67.30_{2.8}$
		Kmeans	$56.54_{1.3}$	$\mathbf{42.29}_{2.9}$	$64.85_{2.2}$	$53.32_{2.1}$	$54.78_{1.9}$	$93.75_{2.0}$	$84.96_{2.9}$	$78.03_{2.3}$	$87.70_{1.5}$	$68.47_{2.2}$
		Vote-k	$58.40_{0.7}$	$42.19_{3.2}$	$65.33_{4.0}$	$53.71_{1.4}$	$57.13_{2.3}$	$90.82_{1.5}$	$84.38_{2.7}$	$78.42_{3.3}$	$86.14_{1.6}$	$68.50_{2.5}$
		Fast Vote-k	$61.72_{0.3}$	$39.55_{1.5}$	$63.18_{1.4}$	$51.95_{1.0}$	$56.15_{2.1}$	$93.46_{0.7}$	$\mathbf{85.74}_{1.9}$	$77.83_{3.0}$	$88.18_{1.5}$	$68.64_{1.7}$
		LM-DPP (ours)	$\mathbf{58.99}_{2.7}$	$41.31_{5.3}$	$\mathbf{66.80}_{2.3}$	$\mathbf{56.15}_{0.9}$	$\mathbf{57.62}_{3.0}$	$\mathbf{94.82}_{0.4}$	$83.50_{2.2}$	$\mathbf{78.91}_{2.1}$	$\mathbf{89.36}_{1.8}$	$\mathbf{69.72}_{2.6}$

Table 1: Results with GPT-J and LlaMA-2-7B on NLU task. We compare various selective annotation methods with $\{100, 16\}$ annotated examples. Bold numbers indicate the highest accuracy among all methods, while those underlined indicate the second-best. The subscript denotes the standard deviation.

examples is such that $s(q_i, x) \leq s(q_j, x)$ whenever $i < j$. $s(q_i, x)$ denotes the similarity between the retrieved example q_i and the test example x. This setup potentially leverages the recency bias inherent in LLMs (Zhao et al., 2021).

3 Experiments

3.1 Experimental Settings

Datasets We conduct experiments on 9 NLU and 2 Generation tasks involving different task formulations, including **Sentiment Classification**: SST-5 (Socher et al., 2013); **Natural Language Inference**: RTE (Bentivogli et al., 2009), MNLI (Williams et al., 2017), MRPC (Dolan et al., 2004), QNLI (Wang et al., 2018); **Topic Classification**: TREC (Hovy et al., 2001), DBpedia (Lehmann et al., 2015); **Multiple-choice Question Answering**: Hellaswag (Zellers et al., 2019), COPA (Roemmele et al., 2011); **Abstractive Summarization**: XSUM (Narayan et al., 2018) and **Open Domain QA**: NQ (Kwiatkowski et al., 2019). In the main experiment, the budget of annotation is set as $(\{16, 100\})$. For datasets with publicly available test data, we use the test data for evaluation. For others, we follow previous work (Lan et al., 2019; Su et al., 2022) and use the dev set for evaluation.

Baselines We compare LM-DPP with four strong selective annotation methods. And in our study, we primarily utilize **GPT-J-6B** (Wang and Komatsuzaki, 2021) and **LlaMA-2-7B** (Touvron et al., 2023) as scoring and inference language models, More details about baselines and implementation can be found in Appendix A.3, A.2 respectively.

Methods		Random	Kmeans	Vote-k	Fast Vote-k	LM-DPP
$\mathcal{L} = 16$						
NQ	*ACC.*	$21.74_{4.39}$	$22.78_{3.63}$	$22.79_{3.37}$	$22.01_{3.75}$	$\mathbf{23.83}_{3.10}$
XSUM	*R-L*	$24.57_{0.03}$	$23.65_{0.29}$	$24.88_{1.03}$	$24.74_{1.20}$	$\mathbf{26.34}_{1.07}$
	FactCC	$35.07_{4.26}$	$\mathbf{36.72}_{2.41}$	$32.49_{1.44}$	$34.68_{2.86}$	$33.53_{3.70}$
$\mathcal{L} = 100$						
NQ	*ACC.*	$23.57_{3.54}$	$22.92_{3.13}$	$24.48_{4.01}$	$23.70_{3.51}$	$\mathbf{24.61}_{3.74}$
XSUM	*R-L*	$25.11_{0.41}$	$24.47_{0.46}$	$24.66_{0.84}$	$24.63_{1.37}$	$\mathbf{27.29}_{0.55}$
	FactCC	$35.64_{5.86}$	$34.86_{2.97}$	$36.12_{2.40}$	$\mathbf{36.53}_{3.84}$	$35.16_{2.01}$

Table 2: Results with LlaMA-2-7B on Generation Task.

Metrics We compare the predicted answers with the true outcomes and report the accuracy (Acc.) for all NLU tasks and exact matching scores (Rajpurkar et al., 2016) for NQ. For summarization tasks, we assess factual consistency using FactCC (Kryscinski et al., 2020) [1], a BERT-based (Devlin et al., 2019) metric for evaluating output faithfulness. Simultaneously, for quality assessment, we report the ROUGE-L F1 score (Lin, 2004) to evaluate the summary against the reference.

3.2 Main Results

NLU Task From Table 1, we can observe that LM-DPP consistently improves the **on-average** accuracy across a variety of NLU tasks under

[1] https://huggingface.co/manueldeprada/FactCC

different annotation budgets ($|\mathcal{L}| = 16$, $|\mathcal{L}| = 100$). Specifically, with a larger budget, LM-DPP achieves an average absolute gain of 1.15% on GPT-J and 1.08% on LlaMA, compared to the best-performing baseline. This demonstrates that balancing uncertainty and diversity ensures that the chosen demonstrations are more likely to contain complementary information that enhances performance. On GPT-J, LM-DPP exhibits the lowest average standard deviation (2.6, 2.0), and on LlaMA-2, it shows greater stability than the Random baseline, albeit marginally lower than Vote-k. This indicates that LM-DPP can maintain a relatively stable performance across different experimental setups, substantially increasing the reliability and robustness of contextual learning. Furthermore, we observe that as the annotation budget increases, performance fluctuations decrease across different selection methods.

Generation Task Experiments on LlaMA-2 (as shown in Table 2) reveal that LM-DPP achieves notable improvement on the NQ task across various annotation budgets, especially at $\mathcal{L} = 16$, where it surpasses the best baseline by 1.04%. In the XSUM task, applying LM-DPP consistently enhances Rouge scores, particularly achieving a 2.18% increase at $\mathcal{L} = 100$. This underscores the efficacy of the proposed method in improving the generality and reference similarity of generated text. However, this improvement comes **at the cost of some degree of factual consistency** with the reference, potentially due to the pursuit of diversity reducing the focus on task-specific relevance (see Appendix C.2 for a more detailed analysis). Overall, LM-DPP boosts the model's generalization and accuracy and highlights the potential for performance optimization with increased annotation budgets. Despite some variability in factual consistency, these insights pave the way for future research on efficiently allocating annotation resources in NLG tasks (Dong et al., 2022).

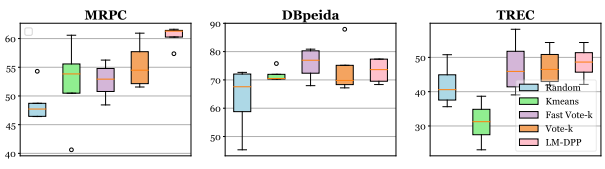

Figure 3: LlaMA-2-7B Results with $\mathcal{L} = 4$.

Smaller In-Context Examples We investigate the impact of the number of examples and labels on ICL performance. As shown in Figure 3, LM-DPP surpasses the other baselines in terms of accuracy and stability on MRPC and TREC but is slightly inferior to Vote-k on DBpedia. Further analysis suggests that a well-balanced demonstration set does not always result in improved performance or reduced variance (see Appendix C.3 for more details). In TREC, performance increases with more labels, whereas in MRPC, demonstrations with a single label (all being `equivalent`) lead to better performance than a balanced demonstration set, with less variance.

λ	MRPC	QNLI	TREC	DBpedia	Hellaswag
0.0	62.57	51.43	79.40	90.67	67.16
0.2	66.42	52.64	78.82	89.47	66.73
0.4	65.34	53.21	77.69	90.22	65.05
0.5	66.89	**53.38**	81.43	**91.52**	68.89
0.6	**67.10**	53.26	**82.01**	91.03	**68.92**
0.8	66.39	52.18	81.24	90.77	67.42
0.9	66.51	52.97	79.36	84.25	66.27
1.0	66.14	51.45	81.57	79.49	59.73

Table 3: The GPT-J performance of different trade-off factors λ. ($\lambda = \{0.0, 1.0\}$) correspond respectively to the vanilla DPP and the Perplexity baseline (§A.3).

4 Analysis

4.1 Impacts of the Trade-off Between Uncertainty and Diversity

We analyze to investigate how the trade-off between diversity and uncertainty impacts the performance of downstream tasks. With an annotation budget of 100, we test the performance under different (λ) values utilizing GPT-J as the inference model. As evident from Table 3, a complete inclination towards uncertainty ($\lambda = 1.0$) generally yields poorer outcomes across all tasks, likely due to selective annotation excessively concentrating on a small portion of data, thereby diminishing ICL's generalization capacity. Optimal effects are often observed at (λ) values of **0.5** or **0.6** (which approximate a balance between the two factors), suggesting that moderate uncertainty coupled with a degree of diversity is beneficial for the model's downstream task performance. Moreover, different tasks demonstrate varied sensitivities to the (λ) value. For instance, QNLI shows minor performance shifts ($\pm 1.95\%$), whereas DBpedia exhibits significant performance variations at certain (λ) values (exceeding $\pm 10.00\%$), indicating that the optimal selection of (λ) may relate to the tasks' characteristics and difficulty levels. Despite such

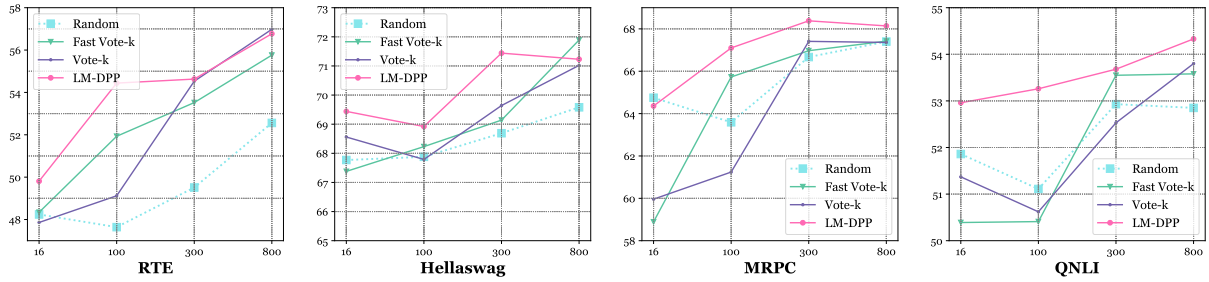

Figure 4: Comparisons of various selection methods with ({16, 100, 300, 800}) annotated examples on four representative tasks: RTE, MRPC paraphrase detection, QNLI, and Hellaswag commonsense answering for GPT-J.

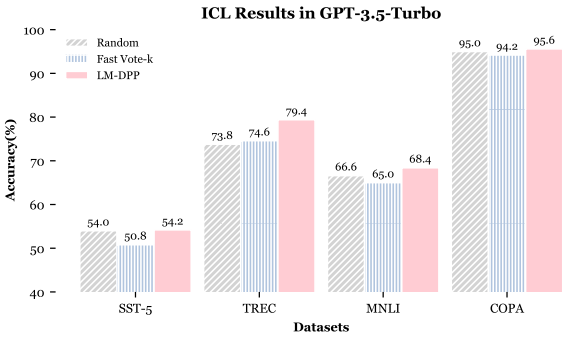

Figure 5: Results of GPT-3-Turbo (175B) with 100 annotated examples. LM-DPP consistently improves in-context learning on various datasets.

variability, we find that introducing this trade-off factor consistently surpasses the vanilla DPP and Perplexity baselines, which consider only diversity or uncertainty, thereby validating the effectiveness of LM-DPP.

4.2 Transferability across Different LMs

Small model for scoring Scoring every sample from the extensive unlabeled pool using a more resource-intensive LLM could be computationally demanding, particularly when the size of the unlabeled sample pool is substantial. Therefore, we attempt to use GPT2 (Radford et al., 2019) (117M, which possesses basic language modeling capabilities) as a surrogate for the source language model GPT-J, while maintaining GPT-J for the inference model. Across 9 NLU tasks (annotation size=100), the average accuracy was **64.76** (details in Appendix C.1). This indicates that LM-DPP exhibits strong transferability across different inference LMs, which means that the selected demonstrations can be reused.

Transfer to LLMs To gain some intuition on the effect of model size, we endeavor to transfer the proposed method to LLMs that are aligned with human expectations (`gpt-3.5-turbo-instruct`) (Ouyang et al., 2022).

In specific, we take the *logprobs* returned by the official API as a reference for measuring uncertainty, from which we calculate $r(x_i)$ and perform standard LM-DPP. As depicted in Figure 5, we report the experimental results of GPT-3.5-Turbo (175B) with LM-DPP on several datasets and compare them with the Random and Fast Vote-k baseline. In comparison to random selection, our results indicate that LM-DPP can significantly enhance the performance of GPT-3.5, as evidenced by the 5.6% improvement in TREC accuracy, 1.8% in MNLI, 0.2% in SST-5, and 0.6% in COPA. The proposed LM-DPP approach surpasses Fast Vote-k by an average of 3.25%, indicating that considering representativeness alone is not sufficient to extract a high-quality demonstration subset.

4.3 Varying budget of annotated examples

We further investigate how the size of the annotation set affects the performance of in-context learning. Under annotation sizes of ({16, 100, 300, 800}), we compare LM-DPP with Random selection, Fast Vote-k, and Vote-k, and report the results in Figure 4. It is observable that with increasing annotation budgets, most selective methods generally show a consistent overall improvement trend. This is in line with the expectation that more labeled data is more likely to retrieve relevant examples to assist LLMs in accurately answering, thereby improving the performance of in-context learning. The proposed approach, LM-DPP, outperforms other methods at an annotation size of 16 on RTE, Hellaswag, and QNLI, suggesting that even with extremely low annotation budgets, LM-DPP can ensure the effectiveness and diversity of context. Additionally, with a sufficient annotation budget ($\mathcal{L} = 800$), LM-DPP exhibits commendable performance, achieving the best results on two

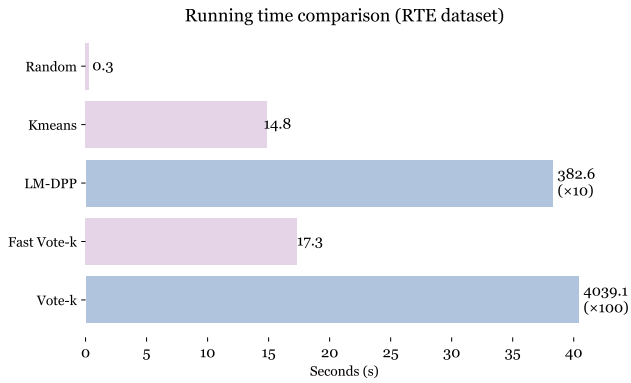

Figure 6: The time consumed to select 300 demonstrations from the RTE dataset (comprising 2491 instances).

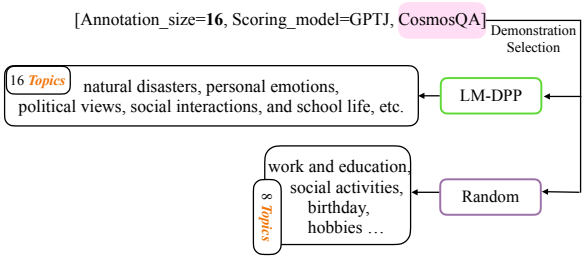

Figure 7: Case Study of selected demonstrations under the condition of annotation_size=16.

	Hellaswag	COPA	DBpedia	TREC	QNLI	MNLI
Random†	67.88	84.03	90.30	76.36	51.11	39.41
LM-DPP†	68.92	83.80	91.03	82.01	53.26	42.31
UN-LM-DPP	68.48 $_{\textbf{-0.64}}$	83.20 $_{\textbf{-0.72}}$	90.74 $_{\textbf{-0.32}}$	76.48 $_{\textbf{-6.74}}$	53.37 $_{\textbf{+0.21}}$	41.09 $_{\textbf{-2.88}}$

Table 4: The GPT-J performance on various datasets. †Resulting numbers are taken from Table 1. The annotation budget is 100. In UN-LM-DPP, the annotation set consists of two parts: $\mathcal{D}^i$ and $\mathcal{D}^u$, with standard ICL being implemented.

datasets, MRPC and QNLI. In contrast, the performance decline of Vote-k on QNLI may be attributed to the annotation of noisy data (high perplexity), with some case analyses provided in the appendix A.1. This reaffirms the necessity of balancing uncertainty and diversity.

4.4 Time Efficiency

We explore the execution efficiency of both the baseline methods and LM-DPP. As illustrated in Figure 6, the LM-Free approach significantly reduces the time required to select demonstrations compared to methods that require scoring by LM. Selecting 300 samples takes 4039.1s with Vote-k, 382.6s with LM-DPP, and only 0.3s with random selection. Since LM-DPP only requires a single forward pass per sample, we can optimize time efficiency in two ways: (1) preemptively compute perplexity for data samples in practical scenarios and devise methods to reset or update cached demonstration samples periodically. (2) using smaller-parameter scoring models (see §4.2) can achieve more than tenfold acceleration (24.4s).

5 Discussion

5.1 Case Study

We compare demonstrations selected via LM-DPP against Random in CosmosQA dataset (Huang et al., 2019). It reveals that demonstrations selected by the LM-DPP exhibit greater diversity in content, covering 16 distinct topics such as natural disasters, personal emotions, political views, social interactions, and school life, compared to only 8 topics covered by random selection (Figure 7). The selected demonstrations not only span a broad range of subjects but also offer a variety in style, including personal narratives, descriptive events, emotional expressions, and dialogues. This diversity enhances the model's ability to interpret and respond to questions.

5.2 Does annotation benefit from gold labels?

Min et al. (2022) observed that random substitution of labels in demonstrations minimally impacts the performance across a suite of tasks, while Yoo et al. (2022) highlighted that the integrity of input label mapping is a crucial factor. In this section, we explore whether Gold Labels (i.e., providing correct labels) are essential for achieving high performance in ICL.

Specifically, we divide the selective annotation process into several steps. Step 1: Annotate 50 instances to construct an in-domain dev set $\mathcal{D}^i$ (containing gold labels). Step 2: For the unannotated instances, we pair each input x_i with every possible label $y \in \mathcal{C}$ ($\mathcal{C}$ is the label set) to construct a train set $\mathcal{D}'$ carrying pseudo-labels. Step 3: Given the prompts $\mathcal{Z} \in \mathcal{D}'$, the ICL accuracy on the in-domain dev set $\mathcal{D}^i$ is denoted as Acc($\mathcal{Z}$). We select the Top-50 $\mathcal{Z}$, represented as $\mathcal{D}^u$. Therefore, the final annotation set ($|\mathcal{L}| = 100$) comprises two parts: $\mathcal{D}^i$ with gold labels, and $\mathcal{D}^u$ selected post-hoc. This process is referred to as **UN-LM-DPP**, followed by conducting standard ICL experiments.

As shown in Table 4, we observe that UN-LM-DPP, compared to LM-DPP with gold annotations, exhibits a certain performance decline in most

tasks but still surpasses Random selection in some datasets. The performance fluctuation varies significantly across different tasks, depending on the specific characteristics of the datasets, as evidenced by a decrease of -6.74% in TREC, yet only -2.88% in MNLI.

Dataset	Hellaswag	COPA	DBpedia	TREC	QNLI	MNLI
Gold-Labeled	47.63%	38.86%	25.11%	11.52%	52.30%	37.43%

Table 5: The proportion of golden-labeled examples identified within an unlabeled setting in UN-LM-DPP.

This suggests that, to a certain extent, ICL generally benefits from gold demonstrations. In addition, we report the proportion of gold demonstrations within the constructed $\mathcal{D}^u$ during Step 2, with the results presented in Table 5. In QNLI, there is a 52.30% gold label ratio, and surprisingly, we observe a slight performance improvement compared to LM-DPP. It is evident that within similar tasks, a higher ratio of gold-standard examples correlates with a smaller decline in ICL performance. However, this is not a generalized finding across the board, and we consider annotation-free ICL as a direction for future work.

6 Related Work and Background

Determinantal Point Process The Determinantal Point Process (DPP) is an elegant probabilistic model that captures negative correlations and allows for efficient algorithms in sampling, marginalization, and conditioning (Kulesza, 2012). Formally, a point process $\mathcal{P}$ is a probability measure on the power set of $\mathcal{V}$, that is, the set of all discrete items $2^{\mathcal{V}}$. If Y is a random subset drawn according to $\mathcal{P}$, then for every $S \subseteq Y$:

$$P(S \subseteq Y) = \det(L_S) \tag{4}$$

for some kernel matrix $L \in \mathbb{R}^{n \times n}$ that is symmetric, real and positive semidefinite. L_S denotes the submatrix of L obtained by restricting to the rows and columns indexed by S. The operator $\det(\cdot)$ represents the determinant of a matrix. Typically, the DPP kernel L can be written as a Gram matrix, $L_{ij} = K(a_i, a_j)$, where $K(\cdot, \cdot)$ is the kernel associated with the determinantal point process, often expressed as $\phi(a_i)^T \phi(a_j)$, ϕ is the feature map of a reproducing kernel (Ye et al., 2023).

Under distribution $\mathcal{P}$, our objective is maximum a posteriori (MAP) inference, which is to find the subset of items with the highest probability, corresponding to the most diverse subset of items.

$$S_{map} = \arg\max_{S \in Y} \det(L_S) \tag{5}$$

Although finding the mode of a DPP is NP-hard, pioneering works (Kulesza, 2012; Lee et al., 2009; Chen et al., 2018; Gillenwater et al., 2012) have largely relied on greedy algorithms or sampling methods, and have succeeded in performing greedy MAP inference within polynomial time.

In-context Learning The capacity for in-context learning has been observed in large-scale Pre-trained Language Models (PLMs) such as GPT-3, representing a few-shot learning paradigm that does not require any parameter updates. It involves pre-pending a small number of demonstrations as prompts before the test input, allowing LLMs to discern patterns and "learn" to predict.

Formally, let $\hat{x}$ be the test query to be addressed, and $s(\cdot, \cdot)$ be the cosine similarity. Standard ICL prompts the language model G with a set of example input-output pairs $\{(x_1, y_1) \ldots (x_m, y_m)\}$ and predicts the answer $\hat{y}$ for the query. Typically, the pairs (x_i, y_i) are retrieved from a train set $\mathcal{D}$ within the same domain through similarity.

$$\hat{y} = \arg\max_{y} G_\theta(y \mid \hat{x}, \mathcal{C}), \tag{6}$$
$$\mathcal{C} = \underset{(x_i, y_i) \in \mathcal{D}}{\text{TopK}} \; (s(\hat{x}, x_i)).$$

Recent works have aimed to enhance ICL by selecting valuable demonstrations (Liu et al., 2021a; Rubin et al., 2022), optimizing the order of demonstrations (Lu et al., 2022), etc. Su et al. (2022) utilize selective annotation to significantly reduce annotation costs while ensuring high ICL performance. Yang et al. (2023) explore the corpus-level in-context learning via DPP and mention the need to use gold labels to score candidate samples. CEIL (Ye et al., 2023) train the demonstration retriever with a learnable conditional DPP. However, these existing works are highly dependent on large annotated support sets.

7 Conclusion and Future Work

In this work, we focus primarily on an innovative selective annotation mechanism and introduce an efficient annotation practice, LM-DPP. It selects both diverse and low-uncertainty examples for annotation and demonstrates promising results in various LMs. Moreover, empirical results validate the

generalizability of LM-DPP across model size and annotation budget scaling. In the future, we plan to apply LM-DPP to more NLP tasks and explore annotation-free selection methods.

Limitations

The proposed work still has some limitations.

Selection Method. Previous studies have elucidated that low uncertainty ensures familiarity of the LLMs with the demonstrations (Gonen et al., 2022), while diversity ensures that the selected demonstrations may encompass a broad range of information, thereby enhancing the overall effectiveness of ICL (Margatina et al., 2023). However, we still lack pilot experiments tailored to these factors to examine their impact on ICL performance thoroughly.

Retrieval Method. We have implemented prompt retrieval based on similarity (TopK). However, it is currently unclear whether the proposed method applies to other prompt retrieval methods, such as Random Retrieval, Coverage-based Retrieval (Gupta et al., 2023), and Retrieval based on Mutual Information (Sorensen et al., 2022). We plan to extend our work to cover more scenarios.

Retriever. Retriever is indeed one of the variables in our experiments. However, we have solely employed a retriever based on the SentenceBert architecture. Validating our experimental results on a more diverse array of retrievers constitutes future extension work.

Language. We also acknowledge that all datasets considered in this work are in English, which does not ensure that our work can be broadly generalized to other languages.

Potential Risk

Previous works have shown Large language models contain rich biased data (Bender et al., 2021). Since we use LLMs like LlaMA, GPT-J, and GPT-3, the proposed LM-DPP approach may elicit some content with offensive language or discrimination.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7:8.

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.

Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. 1996. Active learning with statistical models.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. 2023. Language modeling is compression. *arXiv preprint arXiv:2309.10668*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William Dolan, Chris Quirk, Chris Brockett, and Bill Dolan. 2004. Unsupervised construction of large

paraphrase corpora: Exploiting massively parallel news sources.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. Near-optimal map inference for determinantal point processes. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the first international conference on Human language technology research*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Alex Kulesza. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jon Lee, Vahab Mirrokni, Viswanath Nagarjan, and Maxim Sviridenko. 2009. Non-monotone submodular maximization under matroid and knapsack constraints.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

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.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models.

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?

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

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

Burr Settles. 2009. Active learning literature survey.

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

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. *arXiv preprint arXiv:2203.11364*.

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

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, Singapore. Association for Computational Linguistics.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. Opt: Open pre-trained transformer language models.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

A Appendix

A.1 Details with perplexity estimation

	QNLI	
	$\|\mathcal{L}\| = 16$	$\|\mathcal{L}\| = 100$
$\text{Perplexity}_{\text{avg}}$	75.16	95.43
$\text{Perplexity}_{\text{max}}$	143.48	278.62

Table 6: Annotation Set (selected by Vote-k) Perplexity Statistics.

We report the perplexity of annotated instances when ($|\mathcal{L}| = \{16, 100\}$) (as shown in Table 6). It's observed that as the annotation cost increases to 100, there is a corresponding significant rise in perplexity. For instance, in COPA, the $\text{Perplexity}_{\text{avg}}$ increases by 4.01, and $\text{Perplexity}_{\text{max}}$ rises by 125.70. A similar phenomenon is also observed in DBpedia. This indicates to some extent that introducing demonstrations with high perplexity can lead to a decrease in ICL performance.

A.2 Implementation Details

The inference method we employed is direct (a regular inference used in (Brown et al., 2020)), which involves presenting demonstrations and candidate answers to the LLMs to select the candidate with the highest likelihood. For each test dataset, a specific prompt template (Table 12) is used for scoring and inference. For each test instance, we include as many retrieved samples as possible in the preceding prompt, up until the maximum token length was reached (e.g., 2048 for GPTJ, 4096 for LlaMA-2-7B). Sentence-BERT (Reimers and Gurevych, 2019) is used as the demonstration retriever. Following (Rubin et al., 2022), we adopt the `paraphrase-mpnet-base-v2` to encode the test input x_{test} and the inputs of the train set. All experiments are conducted on a single Tesla V100 GPU with 32GB of memory. Empirically, obtaining embeddings for unlabeled examples using `Sentence BERT` as described in Section 2.1 varies between 0.2 to 2 hours, contingent upon the dataset size. In Section 2.2, our approach requires approximately 6 seconds to generate the annotation set on a single CPU. Notably, ICL obviates the need for model training.

Dataset	Task Type	Split
SST-5	Sentiment Classification	8544/1101/2210
RTE	Natural Language Inference	2491/277/3000
MNLI	Natural Language Inference	392702/19647/19643
MRPC	Natural Language Inference	3668/408/1725
QNLI	Natural Language Inference	104743/5463/5463
TREC	Topic Classification	5452/0/500
DBpedia	Topic Classification	560000/0/70000
Hellaswag	Multiple-choice Question Answering	39905/10042/10003
COPA	Multiple-choice Question Answering	1000/0/500
CosmosQA	Multiple-choice Question Answering	9471/1221/1140
XSUM	Abstractive Summarization	204045/11332/11334
NQ	Open Domain QA	307373/7830/0

Table 7: Dataset Statistics in the Experiments.

We also acknowledge that acquiring unlabelled samples in practice is a process marked by significant variance(Su et al., 2022). To simulate this realistic scenario, **we randomly sample 3K instances** from the training set multiple times to serve as the pool of samples awaiting annotation. In all the experimental setups described in this paper, we utilize four distinct seeds **(0, 1, 42, 123)**, and the values presented in the tables (figures) reflect the average across four runs. Additionally, we provide the corresponding standard deviations for these values.

A.3 Baselines

Random A randomly selected annotation baseline is necessary, as it directly picks unlabeled training instances at random. Ideally, data points selected by any heuristic algorithm should yield better performance compared to it.

Perplexity (Gonen et al., 2022) reported that lower perplexity correlates with better performance. We rank candidate instances by their perplexity and select the top $|\mathcal{L}|$ instances with the lowest perplexity as our annotation set.

K-means As a representative selection method in the series of diversity approaches, we employ clustering techniques. Following (Yu et al., 2022), we first encode all data points using an *Encoder*, then perform k-means clustering with $|\mathcal{L}|$ clusters and select instances accordingly.

Vote-k (Su et al., 2022) selects $|\mathcal{L}|/10$ samples through a graph-based voting mechanism, after which the $|\mathcal{L}|/10$ labeled samples are used as context for the LLMs, to calculate confidence scores for the other unlabeled candidate instances. Finally, the instances are grouped according to percentile

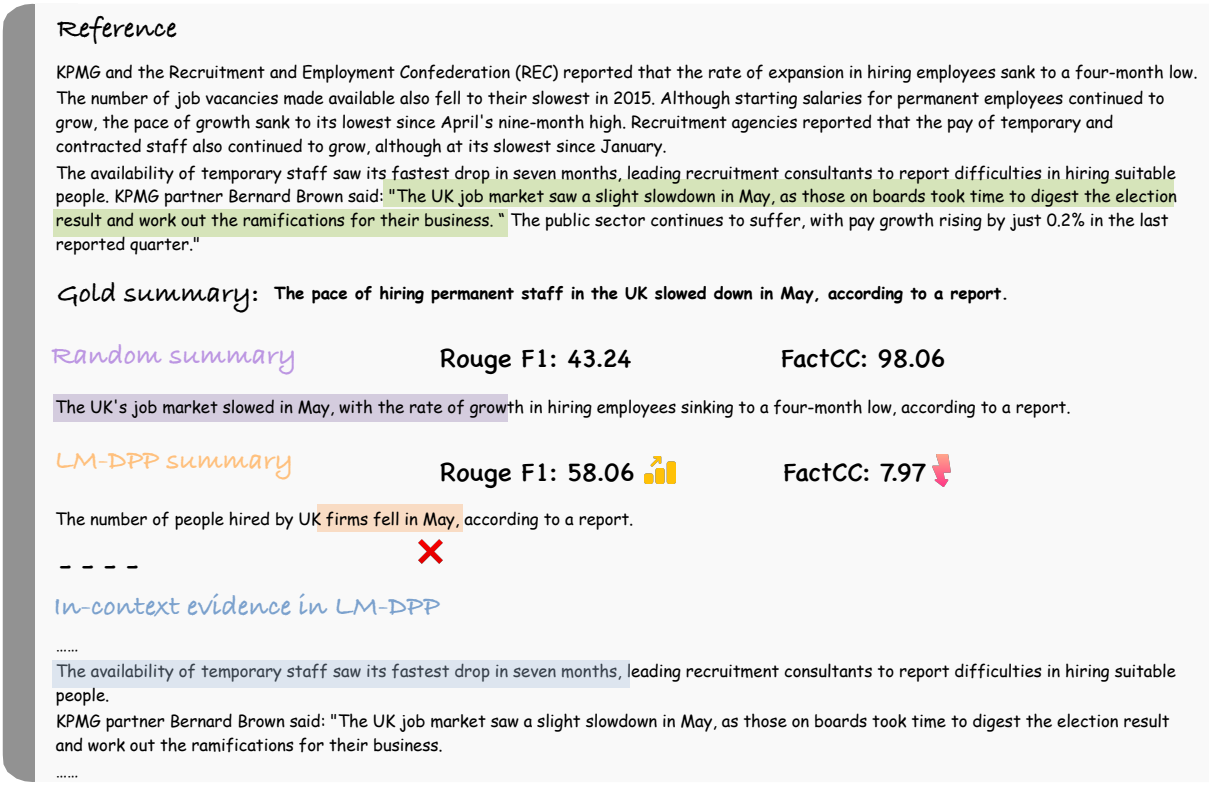

Figure 8: Case analysis in XSUM, we compare the performance of Random and LM-DPP on generation quality and fact consistency

ranks of confidence scores, and selection is made through voting within each group.

Fast Vote-k A rapid and efficient alternative to Vote-k, it circumvents the use of LLMs to compute confidence scores. It directly selects the $|\mathcal{L}|$ samples with the highest voting scores.

A.4 Dataset Statistics

Table 7 presents the data statistics of the datasets employed in our experiments.

A.5 Prompt Template

The prompt templates utilized for each task are reported in Table 12.

B High Uncertainty

	RTE	MNLI	MRPC	QNLI	SST-5
LM-DPP$_{\text{high_uncertainty}}$	51.29	42.91	66.17	52.30	48.74
	DBpedia	TREC	HellaSwag	COPA	
	93.18	81.40	66.95	83.80	

Table 8: Results of selecting high-uncertainty instances (GPTJ + annotation_size=100+LM-DPP). Improvements in high uncertainty are underlined.

Apart from the MNLI and DBpedia datasets, selecting instances of high uncertainty led to a certain degree of performance degradation (Table 8). Therefore, we prioritize the selection of low-uncertainty instances in our experiments and hope to inspire further work in the area of perplexity estimation.

C Analysis and supplement

C.1 Small Model for scoring

	RTE	MNLI	MRPC	QNLI	SST-5
LM-DPP$_{\text{gpt2_scoring}}$	51.96	41.79	66.81	51.43	47.32
	DBpedia	TREC	HellaSwag	COPA	Avg
	90.67	81.85	67.94	83.09	64.76

Table 9: Results of using GPT2 as a surrogate.

Table 9 presents the results of using GPT2 as a surrogate.

C.2 Fact Consistency in XSUM

Upon closer analysis (as shown in Figure 8), we find that in pursuit of diversity and uncertainty in demonstrations, LM-DPP may retrieve content that is topically related but not completely factually

Examples	
LM-DPP:	`equivalent, equivalent, equivalent, equivalent`
Random:	`equivalent, not equivalent, not equivalent, not equivalent`

Table 10: In MRPC, the four demonstration label examples selected by Random and LM-DPP.

consistent. For example, while the source text emphasizes a "The UK job market saw a slight slowdown in May," the LM-DPP generated summary mentions "fell in May," shifting the focal point of the information and potentially misleading readers to interpret a deterioration in actual employment conditions rather than a deceleration in growth rate. This discrepancy is also reflected in the context evidence cited by LM-DPP, which notes "the availability of temporary staff saw its fastest drop in seven months," further reinforcing a negative portrayal of employment circumstances, despite not fully reflecting the source's focus or theme.

We further observe that balancing the Rouge scores with FactCC scores, ensuring factual consistency while maintaining high levels of abstractiveness and textual similarity, presents a significant challenge for LM-DPP. This observation suggests that future research might need to explore more nuanced demonstration selection strategies or introduce stronger fact-checking and correction mechanisms to mitigate the potential risks to factual consistency arising from the pursuit of diversity and uncertainty. This provides valuable insights on how to further optimize the method moving forward.

C.3 Impact of label coverage

At $\mathcal{L} = 4$, the Acc. of Random and LM-DPP on MRPC and TREC are respectively (47.30, 40.63) and (61.36, 49.64). Combined with Tables 10 and 11, it can be seen that as the label coverage increases, performance on MRPC decreases, while TREC shows an expected pattern. This may be related to the difficulty of the task; moreover, from the perspective of data, an imbalanced label distribution might more closely approximate the statistical characteristics of real-world data. In certain cases, imbalanced examples could reflect key signals of specific categories, aiding the model in learning effective decision boundaries more swiftly. We look forward to further research in this area.

Random
Input: *What are the factors leading to the high teen pregnancy rate in Spartanburg , South Carolina?* **Label**: *description and abstract concept*
Input: *Who invented Make-up ?* **Label**: ***human being***
Input: *Who is the current UN Secretary General ?* **Label**: ***human being***
Input: *What does God create in the first sentence of the Bible ?* **Label**: *entity*
LM-DPP
Input: *How much caffeine is in a 16 oz cup of coffee ?* **Label**: *numeric value*
Input: *What is the fastest growing state in the U.S.A. in 1998 ?* **Label**: *location*
Input: *What British female pop singing star of the 1960s and early 1970s was a child actress in the 1940s and '50s* **Label**: *human being*
Input: *Why was Muhammad Ali stripped of his title and barred from boxing in 1967 ?* **Label**: *description and abstract concept*

Table 11: In TREC, the four demonstration examples selected by Random and LM-DPP.

Dataset	Prompt Template	Example
SST-5	How do you feel about the following sentence? \n {Input} \n answer:{Output}	**Input**: this is a stunning film, a one-of-a-kind tour de force. **Output**: very positive
RTE	{Input1}. Based on that information, is the claim {Input2} "entailment", or "contradiction"? \n answer:{Output}	**Input1**: No Weapons of Mass Destruction Found in Iraq Yet. **Input2**: Weapons of Mass Destruction Found in Iraq. **Output**: contradiction
MNLI	{Input1}. Based on that information, is the claim {Input2} "True", "False", or "Inconclusive"? \n answer:{Output}	**Input1**: Good luck, my friends. **Input2**: I wish my friends luck. **Output**: True
MRPC	Are the following two sentences "equivalent" or "not equivalent"? \n {Input1}.\n {Input2}. \n answer:{Output}	**Input1**: Staff writer Dave Michaels contributed to this report. **Input2**: Staff writers Frank Trejo and Robert Ingrassia contributed to this report. **Output**: equivalent
BoolQ	{Input1}. Based on that information, is the claim {Input2} "True", or "False"? \n answer:{Output}	**Input1**: is there going to be another season of Britannia. **Input2**: In March 2018, is was announced that Sky Atlantic had renewed the show for a second season. **Output**: True
QNLI	{Input1}. Based on that information, is the claim {Input2} "entailment", or "contradiction"? \n answer:{Output}	**Input1**: About 40,000,000 tons were produced in 1984. **Input2**: How many tons of bitumen ere produced in 1984? **Output**: entailment
TREC	content: {Input} \n {Output}	**Input**: What films featured the character Popeye Doyle ? **Output**: entity
DBpedia	title: {Input1}; content: {Input2} \n {Output}	**Input1**: Panay Technological College **Input2**: Panay Technological College is a higher institution in Kalibo Aklan. **Output**: educational institution
Hellaswag	The topic is {Input1}. {Input2} \n {Output}	**Input1**: Hurling **Input2**: A group of lacrosse players are shown on a field. they **Output**: run around, trying to get the ball away from each other.
COPA	{Input2}. What was the {Input1} of this? \n {Output}	**Input1**: cause **Input2**: My body cast a shadow over the grass. **Output**: The sun was rising.
CosmosQA	{Input1}. {Input2} \n {Output}	**Input1**: El dropped me off at B. 's house. She welcomed El . and me into her home . **Input2**: Why did she welcome us into the house ? **Output**: She liked us and enjoys our company .
Subj	Input: {Input}. \n Type: {Output}	**Input**: katie is a young girl who loves to climb . **Output**: objective
XSUM	write a short summary:\n {Input}. \n TL;DR: {Output}	**Input**: A lone hiker salutes the aptly named Wet Sleddale Reservoir in Cumbria, as it overflows down a 21 metre high dam wall... **Output**: Photograph by Jeff Overs / BBC
NQ	Write an answer: {Input} \n {Output}	**Input**: who is credited with creating the gothic art movement **Output**: Abbot Suger

Table 12: Prompt templates and corresponding examples used in each dataset.