Topic Coverage-based Demonstration Retrieval for In-Context Learning

Anonymous ACL submission

Abstract

The effectiveness of in-context learning relies heavily on selecting demonstrations that provide all the necessary information for a given test input. To achieve this, it is crucial to identify and cover fine-grained knowledge requirements. However, prior methods often retrieve demonstrations based solely on embedding similarity or generation probability, resulting in irrelevant or redundant examples. In this paper, we propose **TopicK**, a topic coverage-based retrieval framework that selects demonstrations to comprehensively cover topic-level knowledge relevant to both the test input and the model. Specifically, TopicK estimates the topics required by the input and assesses the model's knowledge on those topics. TopicK then iteratively selects demonstrations that introduce previously uncovered required topics, in which the model exhibits low topical knowledge. We validate the effectiveness of TopicK through extensive experiments across various datasets and both open- and closed-source LLMs. Our source code is available at `https://anonymous.4open.science/r/TopicK_ARRmay`

1 Introduction

Large language models (LLMs) (Grattafiori et al., 2024; Yang et al., 2024; Hurst et al., 2024) have demonstrated a remarkable capacity to internalize and utilize novel information solely from contextual input, without requiring any parameter updates. This ability, referred to as *in-context learning* (ICL) (Brown et al., 2020), enables LLMs to leverage a small set of input-output demonstrations to solve previously unseen tasks or adapt to new domains. However, prior studies (Liu et al., 2022; Peng et al., 2024) have shown that the effectiveness of ICL is highly sensitive to the choice of these demonstrations. Consequently, identifying the most informative demonstrations is critical to fully realizing the generalization potential of LLMs through ICL.

In early work, *similarity-based* approaches (Gao et al., 2021; Liu et al., 2022; Ye et al., 2023; Gupta et al., 2023) employed retrieval modules to select relevant demonstrations from a candidate pool, given a test input. They either utilize a BM25 retriever (Robertson et al., 2009) to select exemplars with high lexical overlap, or dense retrievers (Reimers and Gurevych, 2019; Liu et al., 2019) to identify K-nearest-neighbors in the embedding space. These approaches encode the test input and candidate demonstrations separately, enabling efficient retrieval with low latency. However, such off-the-shelf retrievers operate independently of the inference models (i.e., LLMs), thus failing to account for their parametric knowledge.

To address this limitation, recent *uncertainty-based* approaches (Iter et al., 2023; Wang et al., 2024a; Peng et al., 2024) propose selecting demonstrations that reduce the LLM's predictive uncertainty. They measure the generation probability of either the test input (Peng et al., 2024) or the model output (Iter et al., 2023), conditioned on each candidate. Demonstrations are then ranked based on these probabilities. While this aligns retrieval with LLMs, it requires a separate LLM inference for every test-candidate pair, incurring a substantial computational burden at test time. Moreover, as demonstrations are ranked by independently computed probabilities, these methods fail to ensure diversity among the selected demonstrations.

Figure 1 presents a motivating case study for a test input from SciQ (Welbl et al., 2017). The similarity-based approach (Ye et al., 2023) retrieves a demonstration solely based on embedding similarity, thereby failing to capture the specific topics required by the test input. Meanwhile, the uncertainty-based method (Peng et al., 2024) selects a redundant demonstration about "`Carnivore`", where the model exhibits the highest uncertainty, overlooking the diversity among demonstrations. These shortcomings highlight

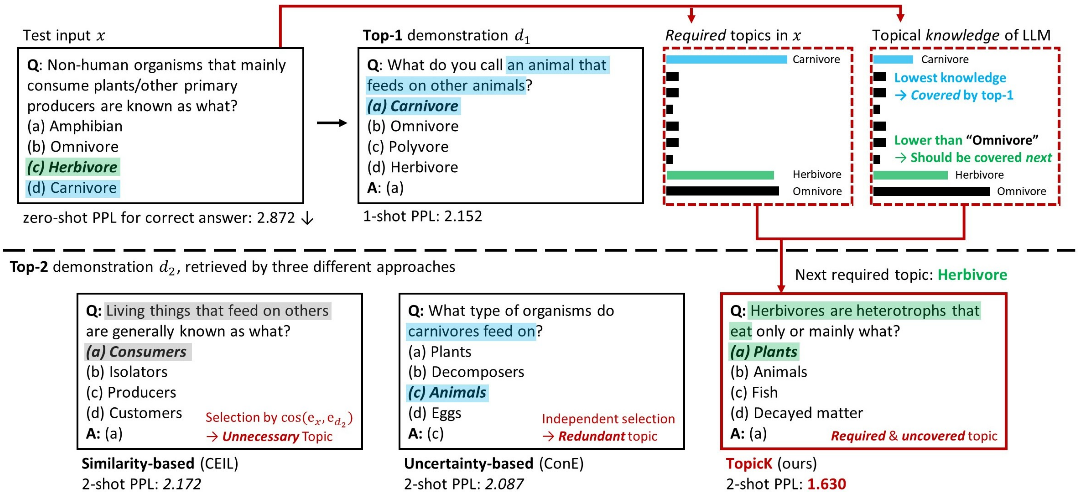

Figure 1: Case study on SciQ dataset and Llama-3.2-1B. The top-1 demonstration is given the same for all three methods. PPL denotes the perplexity (lower is better) for the correct answer (c).

the need for a novel approach that identifies fine-grained knowledge requirements (e.g., topics), and thoroughly covers them by retrieving relevant yet diverse demonstrations.

We propose **TopicK**, a topic coverage-based demonstration retrieval framework that explicitly captures the fine-grained informational demands of both the test input and the model. Specifically, TopicK estimates three key components: (1) *required topics* in the test input, (2) *covered topics* in candidate demonstrations, and (3) *topical knowledge* encoded in the model. These components are inferred via a lightweight topic predictor, without requiring human annotations or LLM inference at test time. TopicK then iteratively selects the demonstrations that introduce previously uncovered required topics, where the model shows low topical knowledge. As a result, in Figure 1, TopicK achieves the lowest perplexity by retrieving a demonstration with a new topic "`Herbivores`".

The key features of TopicK are summarized as:

- TopicK captures fine-grained topic-level knowledge requirements of test inputs, going beyond existing methods that rely solely on embedding similarity or generation probability.
- TopicK infers the required topics using a lightweight topic predictor, avoiding the need for LLM inference at test time as in previous uncertainty-based methods.
- TopicK consistently outperforms state-of-the-art approaches across diverse benchmarks and model scales, including both open- and closed-source LLMs.

2 Preliminary

2.1 In-Context Learning

In-context learning (ICL) is one of the core emergent capabilities of large language models (LLMs), enabling them to internalize and utilize novel information solely from contextual input, without requiring updates to model parameters.

Problem Formulation Given a test input x, an LLM generates the output $\hat{y}$, conditioned on a few in-context demonstrations, as follows:

$$\hat{y} \sim p_{\text{LM}}(\hat{y} \mid \underbrace{d_1, d_2, \ldots, d_K}_{\text{demonstrations}}, x), \tag{1}$$

where each demonstration $d_i = (x_i, y_i)$ is selected from a candidate pool $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$.

A variety of strategies have been proposed to improve ICL performance, including demonstration selection (Gao et al., 2021; Liu et al., 2022; Rubin et al., 2022; Wu et al., 2023b), demonstration ordering (Lu et al., 2022; Lee et al., 2024; Liang et al., 2025), and prompt template design (Deng et al., 2022; Xu et al., 2022; Prasad et al., 2023; Cheng et al., 2023). In this work, we focus on demonstration selection, which has been identified as the most critical factor influencing ICL effectiveness (Peng et al., 2024; Wan et al., 2024).

2.2 Similarity-based Approaches

Early work (Gao et al., 2021; Liu et al., 2022) employs embedding models (Reimers and Gurevych, 2019; Liu et al., 2019) to obtain vector representations $\mathbf{e}_x$ and $\mathbf{e}_d$ for the test input x and each demonstration $d \in \mathcal{D}$. The top-K demonstrations are then selected by ranking candidates in descending order of cosine similarity, i.e., $\cos(\mathbf{e}_x, \mathbf{e}_d)$. Subsequent studies further take account of diversity by employing majority voting (Su et al., 2022), determinantal point processes (Ye et al., 2023), or BERTScore (Gupta et al., 2023).

Limitation While these approaches offer fast retrieval, they remain model-independent and overlook the parametric knowledge of LLMs. Moreover, they consider diversity only at the surface-level embedding space. As a result, although the retrieved demonstrations may be semantically similar to the test input, they often provide limited utility and fail to meaningfully influence the model's decision-making process (Peng et al., 2024).

2.3 Uncertainty-based Approaches

Recent approaches argue that the utility of a demonstration is not solely determined by its similarity to the test input, but also by how it interacts with LLMs (Peng et al., 2024; Chen et al., 2024). These uncertainty-based approaches aim to select demonstrations that explicitly reduce the model's predictive uncertainty. For instance, Iter et al. (2023) select demonstrations that minimize the entropy of the model's output distribution: $\arg\min_{d_i \in \mathcal{D}} \ \mathbb{H}(\hat{y} \mid d_i, x)$. Similarly, Peng et al. (2024) select candidates that minimize the entropy of the test input: $\arg\min_{d_i \in \mathcal{D}} \ \mathbb{H}(x \mid d_i)$.

Limitation Although uncertainty-based objectives align demonstration retrieval with LLMs, they require a separate LLM inference to examine each demonstration. This incurs substantial computational overhead, severely limiting scalability and practical deployment. Furthermore, they rank the demonstrations based on independently computed probabilities, overlooking the diversity in the selected demonstrations.

3 Methodology

We propose **TopicK**, a novel demonstration retrieval framework that leverages topics to explicitly examine the fine-grained informational demands of both the test input and the target LLM. TopicK consists of two major stages as follows:

- **Topical Knowledge Assessment** (§3.1): TopicK estimates three key components (1) *required topics* in the test input, (2) *covered topics* in candidate demonstrations, and (3) *topical knowledge* encoded in the LLM's parameters.
- **Topic Coverage-based Retrieval** (§3.2): TopicK selects demonstrations that introduce previously uncovered required topics, where the model exhibits low topical knowledge.

3.1 Topical Knowledge Assessment

We first identify core topics of each demonstration, without relying on external data or human annotations. Then, a lightweight topic predictor is devised based on the identified topics, and utilized to estimate the topic distributions of both the test input and candidate demonstrations.

3.1.1 Topic Identification

Given a candidate pool $\mathcal{D}$ and a topic set $\mathcal{T}$, our objective is to identify the core topics of each demonstration. While any pre-defined topic set (Shen et al., 2018) can be employed, for broader applicability, we construct $\mathcal{T}$ from scratch using topic mining tools (Shang et al., 2018; Zhang et al., 2023).

Candidate Topic Matching After constructing $\mathcal{T}$, we find a candidate topic set for each demonstration $d \in \mathcal{D}$ with two types of matching:

- **Lexical Overlap:** Select the top-10 topics based on BM25 scores (Robertson et al., 2009).
- **Semantic Similarity:** Select the top-10 topics based on cosine similarity $\cos(\mathbf{e}_d, \mathbf{e}_t)$. $(t \in \mathcal{T})$

To ensure coverage, topics matched by the lexical overlap are excluded from the semantic similarity matching. The candidate topic set $\mathcal{T}'_d \subset \mathcal{T}$ is obtained by merging those two results.

Core Topic Matching with LLMs Topics can exhibit varying semantics depending on the context. To consider this, we leverage the contextualization capabilities of LLMs. Specifically, we prompt GPT-4o (Hurst et al., 2024) to select the core topics from $\mathcal{T}'_d$ and identify any missing but relevant topics. This process yields the finalized core topic set $\mathcal{T}_d \subset \mathcal{T}$ for each demonstration $d \in \mathcal{D}$.

3.1.2 Topic Predictor

Using the identified core topics, we devise a lightweight topic predictor that maps each demonstration embedding $\mathbf{e}_d$ to a topic distribution $\hat{\mathbf{t}}_d \in [0, 1]^{|\mathcal{T}|}$. Each element $\hat{\mathbf{t}}_{d,t} \in \hat{\mathbf{t}}_d$ represents the degree of membership of topic t in the core topic set $\mathcal{T}_d$. In this work, we employ a three-layer MLP $\hat{\mathbf{t}}_d = f(\mathbf{e}_d)$, as the simplest choice. We note that the topic predictor not only generalizes to unseen

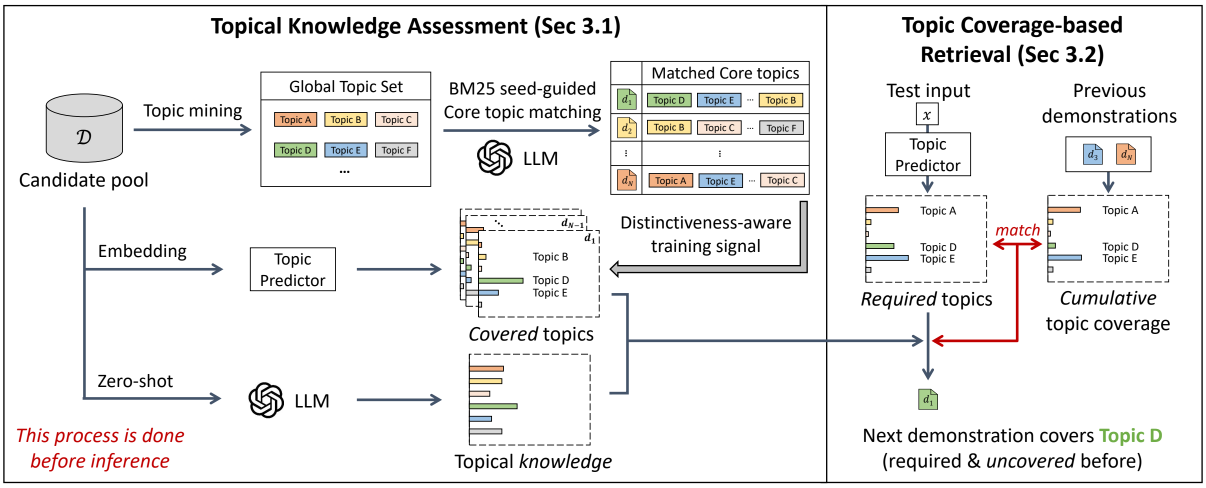

Figure 2: Overview of topic coverage-based demonstration retrieval (TopicK) framework.

test inputs (i.e., $\mathbf{e}_x$), but also *enriches* topic distributions by inferring related topics beyond the initial core topic set.

Distinctiveness-aware Training Signal A naive training signal for $\hat{\mathbf{t}}_d$ would be a binary vector $\mathbf{t}_d \in \{0,1\}^{|\mathcal{T}|}$, where $\mathbf{t}_{d,t} = 1$ if $t \in \mathcal{T}_d$ and 0 otherwise. However, not all topics contribute equally; some topics are more distinctive to a given demonstration. To capture this, we adopt a distinctiveness metric inspired by Lee et al. (2022):

$$\mathrm{DST}(d,t) = \frac{\exp\left(\mathrm{BM25}(d,t)\right)}{1 + \sum_{d' \in \mathcal{D}_d} \exp\left(\mathrm{BM25}(d',t)\right)}, \tag{2}$$

where $\mathcal{D}_d$ denotes the set of 100 demonstrations nearest to d in the embedding space. We then normalize the distinctiveness scores to produce a soft target vector $\mathbf{t}_d \in [0,1]^{|\mathcal{T}|}$:

$$\mathbf{t}_{d,t} = \frac{\mathrm{DST}(d,t)}{\max_{t' \in \mathcal{T}_d} \mathrm{DST}(d,t')}. \tag{3}$$

Finally, we train the topic predictor $\hat{\mathbf{t}}_d = f(\mathbf{e}_d)$ by using a binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{TP}} = -\sum_{d \in \mathcal{D}} \Big(\sum_{t \in \mathcal{T}_d} \mathbf{t}_{d,t} \log \hat{\mathbf{t}}_{d,t} + \sum_{t \notin \mathcal{T}_d} \log(1 - \hat{\mathbf{t}}_{d,t}) \Big). \tag{4}$$

3.1.3 Topical Knowledge Assessment

Required & Covered Topics To assess relevant topics for each sample, we utilize the trained topic predictor described earlier. Given embeddings of a test input $\mathbf{e}_x$ and a demonstration $\mathbf{e}_d$, we predict their topic distributions $\hat{\mathbf{t}}_x = f(\mathbf{e}_x) \in [0,1]^{|\mathcal{T}|}$ and $\hat{\mathbf{t}}_d = f(\mathbf{e}_d) \in [0,1]^{|\mathcal{T}|}$. Here, $\hat{\mathbf{t}}_x$ represents the *required* topics needed to understand and answer the test input x, while $\hat{\mathbf{t}}_d$ indicates the *covered* topics in the candidate demonstration d. These distributions allow fine-grained assessment of how well a demonstration aligns with the informational needs of a test input.

Topical Knowledge In addition to the required and covered topics, we also consider the model's inherent knowledge on each topic, defined as $\hat{\mathbf{t}}_{\mathrm{LM}} \in [0,1]^{|\mathcal{T}|}$. We estimate the topical knowledge by aggregating the model's *zero-shot* accuracy on candidate demonstrations:

$$\hat{\mathbf{t}}_{\mathrm{LM},t} = \frac{\sum_{d \in \mathcal{D}} \hat{\mathbf{t}}_{d,t} \cdot \text{zero-shot}(d)}{\sum_{d \in \mathcal{D}} \hat{\mathbf{t}}_{d,t}}, \quad \text{zero-shot}(d) = \mathbf{1}[y = \arg\max_{\hat{y}} p_{\mathrm{LM}}(\hat{y}|x)], \tag{5}$$

where zero-shot$(d) \in \{0,1\}$ indicates the zero-shot accuracy on demonstration $d = (x,y) \in \mathcal{D}$. That is, we measure how reliably the LLM answers instances associated with each topic without any demonstrations. This prior provides insights into which topics the model has already internalized, allowing us to avoid selecting demonstrations for topics that the model already knows well.

3.2 Topic Coverage-based Retrieval

3.2.1 Topic Coverage-aware Relevance

We define a novel relevance score between a test input x and a candidate demonstration d as follows:

$$r(x,d) = \sum_{t \in \mathcal{T}} \frac{\hat{\mathbf{t}}_{x,t} \cdot \hat{\mathbf{t}}_{d,t}}{\hat{\mathbf{t}}_{\mathrm{LM},t}} = \langle \hat{\mathbf{t}}_x \oslash \hat{\mathbf{t}}_{\mathrm{LM}}, \hat{\mathbf{t}}_d \rangle, \tag{6}$$

where $\oslash$ denotes the element-wise division and $\langle \cdot, \cdot \rangle$ is the inner product. This relevance score captures three critical aspects:

- **Required Topics**: $\hat{\mathbf{t}}_{x,t}$ prioritizes topics highly relevant to the test input.
- **Covered Topics**: $\hat{\mathbf{t}}_{d,t}$ promotes demonstrations that provide high coverage of required topics.
- **Topical Knowledge**: $\hat{\mathbf{t}}_{\text{LM},t}$ down-weights topics that the model already knows well.

By our design, TopicK assigns a high relevance score for a demonstration d, whose covered topics $\hat{\mathbf{t}}_d$ align well with the knowledge-weighted required topics (i.e., $\hat{\mathbf{t}}_x \oslash \hat{\mathbf{t}}_{\text{LM}}$). It is worth noting that $\hat{\mathbf{t}}_{\text{LM}}$ is pre-computed before the test time, while $\hat{\mathbf{t}}_x$ and $\hat{\mathbf{t}}_d$ are inferred via a lightweight topic predictor. Therefore, TopicK enables LLM-aware demonstration selection without LLM inference at test time. We use the final relevance score as $r(x, d) + \lambda \cdot \cos(\mathbf{e}_x, \mathbf{e}_d)$, incorporating both topical and semantic relevance.

3.2.2 Cumulative Topic Coverage

We further incorporate cumulative topic coverage to avoid retrieving redundant demonstrations. Given a set of previously selected demonstrations $\mathcal{D}'_x$, we update the covered topics $\hat{\mathbf{t}}_d$ in Eq. 6 as:

$$\hat{\mathbf{t}}_d \leftarrow (\hat{\mathbf{t}}_{d\cup\mathcal{D}'_x} - \hat{\mathbf{t}}_{\mathcal{D}'_x}). \tag{7}$$

Here, $\hat{\mathbf{t}}_{\mathcal{D}'_x}$ and $\hat{\mathbf{t}}_{d\cup\mathcal{D}'_x}$ represent the cumulative topic coverage before and after adding d. These are also obtained by the topic predictor using mean-pooled embeddings, e.g., $\hat{\mathbf{t}}_{d\cup\mathcal{D}'_x} = f(\mathbf{e}_{d\cup\mathcal{D}'_x})$ and $\mathbf{e}_{d\cup\mathcal{D}'_x} = (\mathbf{e}_d + \sum_{d'\in\mathcal{D}'_x} \mathbf{e}_{d'})/(1 + |\mathcal{D}'_x|)$. This formulation encourages the selection of the next demonstration that introduces novel topic coverage beyond what has already been covered by $\mathcal{D}'_x$. After iteratively selecting K demonstrations, the final set $\mathcal{D}_x = \{d_i\}_{i=1}^K$ is prepended to the test input to generate the output $\hat{y} \sim p_{\text{LM}}(\hat{y} \mid \mathcal{D}_x, x)$. To reduce computational overhead, we retain only the top-300 candidates of the first iteration.

3.2.3 Theoretical Justification

Lastly, we provide a theoretical justification for how our topic coverage-aware relevance is derived. We start from $\mathbb{H}(x|d)$, the uncertainty regarding the test input x given the demonstration d. Since x is known at test time, minimizing this uncertainty is equivalent to maximizing the generation probability $p(x|d)$. While ConE (Peng et al., 2024) estimates this probability through expensive LLM inference at test time, we instead decompose it via topic modeling (Blei et al., 2003):

$$\begin{aligned} p(x|d) &= \sum_{t\in\mathcal{T}} p(x|t) \cdot p(t|d) \\ &= \sum_{t\in\mathcal{T}} (p(t|x) \cdot p(x) \,/\, p(t)) \cdot p(t|d) \\ &= p(x) \cdot \sum_{t\in\mathcal{T}} \underbrace{p(t|x)}_{\substack{\text{required}\\\text{topics}}} \cdot \underbrace{p(t|d)}_{\substack{\text{covered}\\\text{topics}}} \,/\, \underbrace{p(t)}_{\substack{\text{topical}\\\text{knowledge}}} . \end{aligned} \tag{8}$$

Here, $p(x)$ is constant across demonstrations. The terms $p(t|x)$, $p(t|d)$, and $p(t)$ correspond to $\hat{\mathbf{t}}_{x,t}$, $\hat{\mathbf{t}}_{d,t}$, and $\hat{\mathbf{t}}_{\text{LM},t}$ in Eq. 6, respectively. Thus, our topic coverage-based retrieval is equivalent to minimizing the model's uncertainty on the test input.

4 Experiment

4.1 Experimental Setup

Due to a lack of space, please refer to Appendix B for further details.

Models We conduct experiments using two widely adopted model families, **Llama3.2** (Grattafiori et al., 2024) and **Qwen2.5** (Yang et al., 2024), covering a range of model sizes: Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B, Qwen-2.5-0.5B, Qwen-2.5-3B, and Qwen-2.5-7B. All models are instruction-tuned. Additionally, we adopt **Gemini-2.0-Flash-Lite** (Google, 2023) and **Claude-3.0-Haiku** (Anthropic, 2024) for evaluation on closed-source LLMs.

Datasets We evaluate our method on 6 datasets spanning a variety of domains. For general-domain tasks, we use **CommonsenseQA** (Talmor et al., 2019) and **SciQ** (Welbl et al., 2017) for natural language understanding, as well as **QNLI** (Wang et al., 2018) and **MNLI** (Williams et al., 2018) for natural language inference. To assess question-answering performance in specialized domains, we include **MedMCQA** (Pal et al., 2022) from the medical domain and **Law** (Cheng et al., 2024) from the legal domain. Each dataset has a demonstration pool of input-output pairs, and we evaluate the accuracy on the test set with *three* different random seeds. If the test set is private, we report the results on the validation set as done in Peng et al. (2024).

Baselines We compare **TopicK** (ours) with various conventional and state-of-the-art approaches. Specifically, we adopt two basic methods:

- **Zero** uses no demonstration and serves as a zero-shot baseline.

	CommonsenseQA			SciQ			QNLI			MNLI			MedMCQA			Law		
Llama3.2	1B	3B	8B	1B	3B	8B	1B	3B	8B	1B	3B	8B	1B	3B	8B	1B	3B	8B
Zero	37.67	55.13	64.70	64.10	81.50	91.90	51.11	51.73	53.73	42.14	43.45	44.11	37.71	52.45	58.26	41.55	79.50	87.40
Rand	41.03	56.59	63.47	64.90	82.00	92.00	53.54	60.34	72.82	37.87	42.21	46.42	35.76	52.31	59.23	42.70	90.70	93.60
BM25	42.37	52.99	64.05	67.20	82.90	92.30	55.27	68.15	75.68	41.80	46.54	50.79	38.64	56.36	61.57	44.10	91.20	94.70
TopK	43.14	56.33	65.59	71.20	89.00	92.90	60.18	71.05	77.95	50.58	58.94	66.25	39.80	59.65	67.89	47.00	91.80	96.30
CEIL	44.18	57.78	66.68	72.20	89.20	93.30	61.06	71.46	78.63	51.22	60.04	67.02	40.09	61.25	68.10	48.10	92.25	96.80
Set-BSR	44.72	58.48	67.49	72.90	90.20	94.40	61.80	72.32	79.59	51.84	60.77	67.84	40.27	61.79	68.93	48.70	93.00	97.35
MDL	44.51	58.23	67.10	72.60	89.80	94.30	61.34	72.17	79.81	51.76	60.34	67.97	40.16	61.51	68.79	48.50	93.10	97.25
MDR	44.46	57.78	66.88	72.40	89.60	94.10	61.22	72.14	79.37	51.66	60.27	67.88	40.25	60.88	68.63	48.35	92.85	97.10
ConE	44.34	58.40	66.91	72.80	90.10	94.50	61.56	72.20	80.14	51.89	60.53	**68.11**	40.45	62.07	69.03	48.60	93.15	97.45
TopicK	**46.19***	**60.52***	**68.63***	**74.60***	**91.20***	**95.20***	**62.51***	**73.55***	**81.35***	**52.81***	**61.67***	68.06	**41.80***	**62.36**$^{+}$	**70.21***	**51.70***	**93.80**$^{+}$	**97.60**
Qwen2.5	0.5B	3B	7B	0.5B	3B	7B	0.5B	3B	7B	0.5B	3B	7B	0.5B	3B	7B	0.5B	3B	7B
Zero	43.41	63.44	69.45	65.10	93.00	93.80	57.97	64.31	54.15	47.20	47.78	49.54	34.16	51.24	55.47	41.10	69.70	81.85
Rand	44.72	64.50	69.21	71.00	92.60	94.60	55.39	70.52	65.41	48.61	48.14	50.02	35.19	51.28	59.59	42.00	86.90	92.10
BM25	45.62	66.49	70.35	72.30	93.00	94.90	58.13	72.96	74.18	50.55	62.46	64.39	37.07	51.71	60.83	45.50	93.40	95.20
TopK	48.14	65.23	70.42	78.30	93.30	95.10	59.02	76.36	79.57	51.59	67.61	73.55	38.70	53.54	62.88	46.90	95.70	96.30
CEIL	49.64	66.39	71.28	80.50	93.70	95.50	60.55	77.57	80.68	52.28	68.15	74.41	40.00	55.96	64.69	47.45	96.35	96.65
Set-BSR	50.24	66.99	72.15	81.50	94.80	96.20	61.29	78.51	81.66	52.91	68.98	75.31	40.48	56.64	65.48	48.00	96.60	97.50
MDL	49.80	66.34	71.68	81.10	94.30	95.70	61.18	78.03	81.41	53.17	68.66	75.14	39.78	56.48	65.34	48.55	96.80	97.55
MDR	49.63	66.31	71.54	79.80	94.10	95.30	61.07	78.11	81.23	53.23	68.61	75.09	40.12	56.12	65.31	48.50	96.65	97.40
ConE	50.11	66.75	71.91	81.30	94.50	95.90	61.31	78.35	81.75	53.30	68.74	75.28	40.29	56.93	65.52	48.65	96.95	97.70
TopicK	**51.84***	**67.32**$^{+}$	**72.97***	**81.80**$^{+}$	**94.90**	**96.40**	**62.04***	**79.63***	**82.68***	**53.46**$^{+}$	**69.35***	**75.59**$^{+}$	**41.30***	**57.85***	**66.34***	**49.85***	**97.15**	**98.30**$^{+}$

Table 1: Performance (accuracy) of ICL with different demonstration selection strategies. "-B" indicates the model size, and the best result in each column is highlighted in **bold**. $*$ and $+$ indicate $p \leq 0.01$ and $p \leq 0.05$ for the paired t-test with the best competitor.

- **Rand** randomly selects demonstrations for each test example.

and four similarity-based approaches:

- **BM25** (Robertson et al., 2009) selects demonstrations based on lexical overlap.
- **TopK** (Liu et al., 2022) selects the K-nearest-neighbors using dense retriever embeddings.
- **CEIL** (Ye et al., 2023) adopts DPP (Chen et al., 2018) to enhance diversity. For a fair comparison, we exclude the retriever fine-tuning and apply only the DPP-based inference
- **Set-BSR** (Gupta et al., 2023) selects demonstrations based on BERTScore-Recall (BSR) (Zhang et al., 2019), to cover the tokens in the test input.

and three uncertainty-based approaches:

- **MDL** (Iter et al., 2023) selects demonstrations that minimize predictive uncertainty.
- **MDR** (Wang et al., 2024a) selects demonstrations where the model exhibits minimum predictive error.
- **ConE** (Peng et al., 2024) selects demonstrations that minimize uncertainty on test input.

For all methods compared, we adopt the **8-shot** setting, following ConE. We would like to note that all baselines, like ours, freeze both the retriever and the LLMs. For a fair comparison, we exclude methods utilizing retriever update for selecting prompts (Rubin et al., 2022) or demonstrations (Chen et al., 2024; Wang et al., 2024b,c).

Implementation Details Our evaluation setup, including prompt templates and inference procedures, is based on the OpenICL library (Wu et al., 2023a). For the embeddings, we use `all-mpnet-base-v2` (SBERT) (Reimers and Gurevych, 2019), which has shown strong retrieval performance in ConE (Peng et al., 2024). For similarity-based baselines, we utilize the FAISS library (Douze et al., 2024) to perform efficient nearest-neighbor search. For uncertainty-based methods, we retrieve 30 candidate demonstrations with TopK to narrow the search space, following their implementations. All baselines are implemented using publicly available author code, and we strictly follow the documented configurations and hyperparameters.

4.2 Main Results

Table 1 shows the ICL performance with different demonstration selections. We first observe that different model families possess varying levels of domain knowledge. For instance, Llama-3.2 models outperform Qwen-2.5 models on MedMCQA but underperform on CommonsenseQA. This supports our motivation to examine the topical knowledge of each model before retrieving demonstrations. Additionally, we find that Rand occasionally performs worse than Zero, highlighting the importance of appropriate demonstration selection.

Similarity-based approaches (e.g., Set-BSR) generally perform well on general-domain tasks such as CommonsenseQA and SciQ. In contrast, uncertainty-based methods (e.g., ConE) perform

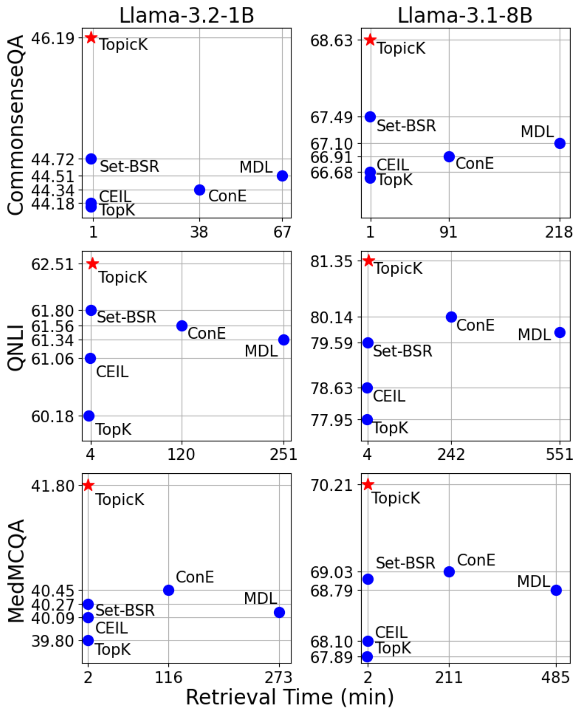

Figure 3: Time-accuracy trade-off. The y-axis represents ICL accuracy, and the x-axis indicates the time elapsed for retrieval with a single A100 GPU.

better in specialized domains like MedMCQA and Law. This is because similarity-based methods rely on surface-level relevance, which is often sufficient for general-domain tasks, whereas uncertainty-based methods incorporate the model's internal knowledge and uncertainty, making them more effective in handling complex or domain-specific reasoning required in specialized tasks.

TopicK consistently outperforms the state-of-the-art similarity-based (Set-BSR) and uncertainty-based (ConE) methods by integrating semantic similarity with topic coverage. Across all datasets, TopicK achieves relative improvements of 1.59% over Set-BSR and ConE. Notably, TopicK yields larger improvement in specialized domains (MedMCQA, Law) by up to 6.38% over ConE with Llama-3.2-1B. This indicates that TopicK selects demonstrations that comprehensively cover the topics in the test input, enabling better leveraging of domain-specific knowledge for unseen tasks.

4.3 Time-Accuracy Trade-off

Figure 3 illustrates the time-accuracy trade-off of Llama-3.2-1B and Llama-3.1-8B across three datasets. Similarity-based methods (TopK, CEIL, Set-BSR) benefit from efficient retrieval through dual-encoder architectures, offering low latency. However, their reliance solely on surface-level similarity often leads to suboptimal performance, particularly in specialized domains (i.e., MedMCQA). On the other hand, uncertainty-based methods (ConE) attain higher accuracy on MedMCQA, by leveraging the LLM itself to evaluate the informativeness of demonstrations. However, they require separate LLM inference for each demonstration, incurring significant computational overhead at test time; ConE is 37× slower than TopicK on QNLI.

TopicK strikes the best balance between accuracy and efficiency. TopicK consistently achieves the best performance by comprehensively covering fine-grained topic-level knowledge, while maintaining low retrieval latency. This advantage arises from its use of a lightweight topic predictor to estimate the knowledge required for each test input. Importantly, the topic predictor operates independently of the LLM size, making TopicK highly scalable and effective across both small and large LLMs

Model	Method	Common	QNLI	MedMCQA
Gemini-2.0 -Flash-Lite	Zero	62.33	74.17	70.31
	Rand	65.10	76.66	71.02
	TopK	67.98	77.54	74.29
	CEIL	68.06	79.97	75.04
	Set-BSR	68.23	80.36	75.47
	TopicK	**69.37**	**84.20**	**78.59**
Claude-3.0 -Haiku	Zero	57.00	72.34	53.73
	Rand	58.97	74.83	59.48
	TopK	63.64	75.71	66.80
	CEIL	64.78	78.14	67.65
	Set-BSR	65.02	78.36	67.81
	TopicK	**67.40**	**82.37**	**69.11**

Table 2: Performance of 5-shot ICL with closed-source LLMs. "Common" represents CommonsenseQA.

4.4 Results with Closed-Source LLMs

Since TopicK estimates the topical knowledge via a topic predictor and zero-shot accuracy, it can be applied to closed-source LLMs as well. We note that uncertainty-based methods (MDL, MDR, ConE) rely on generation probabilities, and therefore, are generally incompatible with closed-source LLMs (Google, 2023; Anthropic, 2024).[1] Table 2 presents the 5-shot ICL performance of two closed-source LLMs. TopicK consistently outperforms all baselines across all tasks and models, demonstrating its effectiveness and generality on restricted APIs.

[1]As of submission, logprobs for those two closed-source LLMs in Table 2 are unavailable.

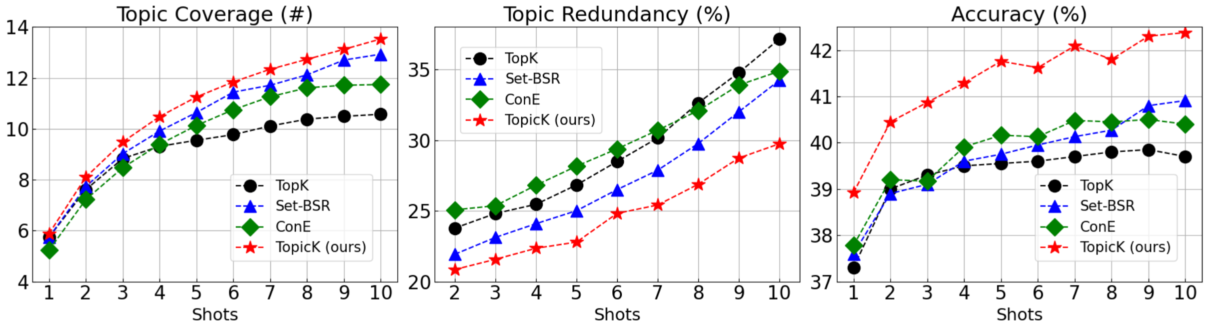

Figure 4: Analysis on topic coverage with Llama-3.2-1B and MedMCQA.

4.5 Topic Coverage Analysis

Figure 4 presents an in-depth analysis of demonstration diversity as the number of shots increases from $K = 1$ to 10. For this analysis, we introduce two metrics:

- **Topic Coverage**: the number of topics covered by demonstrations $(d_1, \ldots, d_K)$, among the top-20 required topics in the test input.
- **Topic Redundancy**: the proportion of topics covered by the current demonstration (d_K) that have already been covered by previous demonstrations $(d_1, \ldots, d_{K-1})$.

We observe that TopK suffers from low topic coverage and high redundancy due to its reliance on similarity ranking without diversity control. Set-BSR improves diversity by adopting setwise BERTScore, but remains limited by surface-level embedding similarity and implicit token-level coverage. ConE, despite considering model uncertainty, shows high redundancy and low coverage as it evaluates each demonstration independently. In contrast, TopicK explicitly targets fine-grained topic coverage, achieving the highest coverage and lowest redundancy. This demonstrates its effectiveness in retrieving demonstrations that are not only relevant and informative but also comprehensively cover a broader range of necessary topics, enhancing overall ICL performance.

4.6 Ablation Study

Table 3 presents an ablation study of TopicK with three variations:

- **"w/o Core Topic"** replaces the LLM-matched core topic set with a BM25-generated candidate topic set.
- **"w/o Soft Label"** trains the topic predictor using a binary vector, rather than the distinctiveness-aware soft label (Eq. 3).

	Common	**QNLI**	**MedMCQA**
TopicK	**46.19**	**62.51**	**41.80**
w/o Core Topic (§3.1.1)	44.72	62.03	41.17
w/o Soft Label (§3.1.2)	45.21	62.38	41.56
w/o Cumulative Coverage (§3.2.2)	44.41	61.47	40.12

Table 3: Ablation study of TopicK with Llama-3.2-1B. "Common" represents CommonsenseQA.

- **"w/o Cumulative Coverage"** omits the update of covered topics (Eq. 7), selecting demonstrations independently.

Removing any component reduces performance, confirming their utility. Removing core topic matching leads to performance degradation across all datasets, confirming the importance of aligning demonstrations with the central topic of the test input. Eliminating distinctiveness-aware labeling slightly reduces accuracy, suggesting that filtering out popular topics improves selection precision. Lastly, the removal of the cumulative topic coverage consistently causes the largest degradation, especially on MedMCQA (-4.02%), indicating that capturing a wide range of subtopics is crucial for complex knowledge-intensive tasks.

5 Conclusions

We argue that an effective set of demonstrations should provide comprehensive coverage of fine-grained aspects (i.e., topics) required by the test input and models. We propose TopicK, which identifies the required topics in the test input and retrieves demonstrations that maximize cumulative topic coverage. By assessing the model's informational needs through topic-level signals, TopicK relies solely on a lightweight topic predictor and avoids any LLM inference at test time. Extensive experiments across diverse domains and both open- and closed-source LLMs demonstrate that TopicK consistently outperforms state-of-the-art methods.

Limitations

Model scale Due to computational constraints, we evaluate TopicK on open-source LLMs ranging from 0.5B to 8B parameters. It would be valuable to scale our experiments to larger models such as Llama-3.3-70B. Instead, we validate TopicK on large-scale closed-source models, including Gemini-2.0-Flash-Lite and Claude-3.0-Haiku.

Flat topic set In this work, we construct a flat topic set and devise a flat topic predictor. Exploring hierarchical topic structures (e.g., topical taxonomy) remains a promising direction for future work, potentially enabling a richer understanding of topic coverage.

References

Anthropic. 2024. Claude 3 model family. https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf. Accessed: 2025-05-06.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31.

Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, and Benjamin Van Durme. 2024. Learning to retrieve iteratively for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7156–7168, Miami, Florida, USA. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337, Singapore. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Google. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162, Singapore. Association for Computational Linguistics.

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, pages 2819–2829.

Yoonsang Lee, Pranav Atreya, Xi Ye, and Eunsol Choi. 2024. Crafting in-context examples according to LMs' parametric knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2069–2085, Mexico City, Mexico. Association for Computational Linguistics.

CHEN Liang, Li Shen, Yang Deng, Xiaoyan Zhao, Bin Liang, and Kam-Fai Wong. 2025. Pearl: Towards permutation-resilient llms. In *The Thirteenth International Conference on Learning Representations*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. GrIPS: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3845–3864, Dubrovnik, Croatia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

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. In *The Eleventh International Conference on Learning Representations*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Xingchen Wan, Ruoxi Sun, Hootan Nakhost, and Sercan Arik. 2024. Teach better or show smarter? on instructions and exemplars in automatic prompt optimization. *Advances in Neural Information Processing Systems*, 37:58174–58244.

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. In *International Conference on Learning Representations*.

Huazheng Wang, Jinming Wu, Haifeng Sun, Zixuan Xia, Daixuan Cheng, Jingyu Wang, Qi Qi, and Jianxin Liao. 2024a. MDR: Model-specific demonstration retrieval at inference time for in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4189–4204, Mexico City, Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2024b. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Song Wang, Zihan Chen, Chengshuai Shi, Cong Shen, and Jundong Li. 2024c. Mixture of demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 37:88091–88116.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhenyu Wu, Yaoxiang Wang, Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Jingjing Xu, and Yu Qiao. 2023a. OpenICL: An open-source framework for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 489–498, Toronto, Canada. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023b. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Wang Yanggang, Haiyu Li, and Zhilin Yang. 2022. Zero-Prompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4235–4252, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 429–437.

A Implementation Details for TopicK

Our source code, including the core topic set for each demonstration, is available at `https://anonymous.4open.science/r/TopicK_ARRmay`

Topic Mining We employ two topic mining tools: SeedTopicMine (Zhang et al., 2023) for extracting single-word topics and AutoPhrase (Shang et al., 2018) for multi-word phrases. We then merge the outputs to construct the topic set $\mathcal{T}$ for each dataset.

Core Topic Matching with GPT-4o We prompt GPT-4o to select the core topics from the candidate topic set $\mathcal{T}'_d$ and identify any missing but relevant topics, as follows:

> You will receive a question-answer demonstration along with a candidate topic set. Your task is to output relevant topics of the demonstration. You may choose topics from the candidate topic set, or you can create new relevant topics. You must provide at least five topics. Do not include any explanation or numbers. Please just output the list of relevant topics, separated by commas. Demonstration: $\{d\}$, Candidate topic set: $\{\mathcal{T}'_d\}$

This process yields the finalized core topic set $\mathcal{T}_d \subset \mathcal{T}$ for each demonstration d.

Topic Predictor In this work, we employ a three-layer MLP $\hat{\mathbf{t}}_d = f(\mathbf{e}_d)$ for the topic predictor. The input embedding $\mathbf{e}_d \in \mathbb{R}^{768}$ is extracted using the `all-mpnet-base-v2` model (Reimers and Gurevych, 2019). The last classification layer is initialized with the embeddings of topic names. The model is optimized using the distinctiveness-aware soft label (Eq.3) and binary cross-entropy (Eq.4).

B Experiment Details

Datasets Table 4 shows the statistics of each dataset. "#Topics" indicates the number of mined topics from each dataset. Since Law (Cheng et al., 2024) dataset does not provide an explicit data split, we randomly partition the 10k input-output pairs into training, validation, and test sets using a 7:1:2 ratio. All datasets are sourced from their official HuggingFace repositories (Wolf et al., 2019).

Dataset	Data Split	#Classes	#Topics
CommonsenseQA	9,741 / 1,221 / 1,140	5	3,781
SciQ	11,679 / 1,000 / 1,000	4	11,451
QNLI	104,743 / 5,463 / 5,463	2	51,809
MNLI	392,702 / 19,647 / 19,643	3	109,390
MedMCQA	120,765 / 6,150 / 4,183	4	49,925
Law	7,000 / 1,000 / 2,000	4	5,296

Table 4: Dataset statistics.

Task	Prompt	Class
CommonsenseQA	Question: $< x >$ Answer: <A>	A
	Question: $< x >$ Answer: <B>	B
	Question: $< x >$ Answer: <C>	C
	Question: $< x >$ Answer: <D>	D
	Question: $< x >$ Answer: <E>	E
MNLI	$< x_1 >$ Can we know $< x_2 >$? Yes.	Entailment
	$< x_1 >$ Can we know $< x_2 >$? Maybe.	Neutral
	$< x_1 >$ Can we know $< x_2 >$? No.	Contradiction
QNLI	$< x_1 >$ Can we know $< x_2 >$? Yes.	Entailment
	$< x_1 >$ Can we know $< x_2 >$? No.	Contradiction
SciQ MedMCQA Law	Question: $< x >$ Answer: <A>	A
	Question: $< x >$ Answer: <B>	B
	Question: $< x >$ Answer: <C>	C
	Question: $< x >$ Answer: <D>	D

Table 5: Templates of tasks. $< x >$ is a placeholder for test inputs.

Templates We adopt the OpenICL library (Wu et al., 2023a) for the prompt templates and inference procedures. Table 5 shows the templates in OpenICL for datasets in the experiment. For a stable evaluation, following ConE (Peng et al., 2024), we adopt the perplexity-based inference in OpenICL.

Hyperparameters All hyperparameters of TopicK and baselines are selected with a grid search on the validation set. If the test set is private and the validation set is used for evaluation, we reserve 10% of the training set as a held-out validation set. For CEIL, the scale factor λ for the DPP-based inference is selected from [0, 0.5]. For uncertainty-based methods (MDL, MDR, ConE), we retrieve 30 candidate demonstrations with TopK to narrow the search space, following their implementations. For MDL, the select time is set to 10 to constrain the time limit. For MDR, the coefficient C is selected from $[0, 1]$. For TopicK, the learning rate of the topic predictor is selected from $[1\mathrm{e}{-5}, 1\mathrm{e}{-4}]$. For the final relevance score $r(x, d) + \lambda \cdot \cos(\mathbf{e}_x, \mathbf{e}_d)$, λ is selected from $[0.1, 1]$ and z-score normalization is applied for $r(x, d)$ and $\cos(\mathbf{e}_x, \mathbf{e}_d)$ to ensure their scales are matched.

Resources For open-source LLMs (i.e., Llama3.2 and Qwen2.5 families), all experiments were conducted on a single NVIDIA A100 80GB GPU with an AMD EPYC™ 7513 2.60GHz CPU. For closed-source LLMs (i.e., Gemini-2.0-Flash-Lite and Claude-3.0-Haiku), all experiments were performed via their respective APIs, subject to usage-based pricing.

Test input Question: Non-human organisms that mainly consume plants/other primary producers are known as what? (A) Amphibian (B) Omnivore (C) Herbivore (D) Carnivore	**Inferred required topics** **carnivore** (0.91), **omnivore** (0.90), **herbivore** (0.87), **plant** (0.34), **ecosystem** (0.28), **food chain** (0.23), **animal** (0.18), **food web** (0.09), **insectivore** (0.07), **vegetarian** (0.06), **organism** (0.05)	**Topical knowledge of LLM** carnivore (0.72), omnivore (0.85), herbivore (0.75), plant (0.77), ecosystem (0.69), food chain (0.74), animal (0.78), food web (0.89), insectivore (0.73), vegetarian (0.76), organism (0.73) **Zero-shot PPL:** 2.872
Top-1 demonstration Question: What do you call an animal that feeds on other animals? (A) Carnivore (B) Omnivore (C) Polyvore (D) Herbivore Answer: (A)	**Inferred covered topics** **carnivore** (0.87), **ecosystem** (0.32), **animal** (0.19), **food chain** (0.19), polyvore (0.13), **organism** (0.11), omnivore (0.08), herbivore (0.07)	**1-shot PPL**: 2.152
Top-2 demonstration Question: Herbivores are heterotrophs that eat only or mainly what? (A) Plants (B) Animals (C) Fish (D) Decayed matter Answer: (A)	**Inferred covered topics** **herbivore** (0.96), heterophile (0.51), **plant** (0.27), xerophyte (0.23), **vegetarian** (0.21), decayed matter (0.19), rotifer (0.15), **food web** (0.08), eutroph (0.07), moss (0.06)	**2-shot PPL**: 1.630
Top-3 demonstration Question: Omnivores are animals that eat both plant- and? (A) Biofuel (B) Liquid diets (C) Recycled food (D) Animal-derived food Answer: (D)	**Inferred covered topics** **omnivore** (0.90), liquid diet (0.33), recycled food (0.17), animal (0.13), biofuel (0.11), carnivore (0.07), plant (0.07), **insectivore** (0.06), food chain (0.03), vegetarian (0.03)	**3-shot PPL**: 1.369

Table 6: Detailed case study on SciQ dataset and Llama-3.2-1B (extended from Figure 1). PPL denotes the perplexity (lower is better) for the correct answer ((C) Herbivore). Scores for inferred topics represent importance according to the topic predictor. i.e., $\hat{\mathbf{t}}_{x,t}$ and $\hat{\mathbf{t}}_{d,t}$.

C Case Study

Table 6 presents a detailed case study with TopicK, extending from Figure 1. First, we observe that the topic predictor not only generalizes to the unseen test input but also enriches the topic distribution by inferring semantically related concepts. For example, TopicK identifies relevant but implicit topics such as ecosystem, food chain, and food web, which are not explicitly mentioned in the input question but enhance the model's understanding.

Second, TopicK retrieves a diverse and relevant set of demonstrations that comprehensively cover the required topics in the test input. By considering cumulative topic coverage, TopicK avoids retrieving redundant demonstrations and prioritizes those that address uncovered but important topics.

Third, TopicK incorporates the model's understanding of each topic via topical knowledge. For instance, although **omnivore** (0.90) has a higher importance than **herbivore** (0.87), the model exhibits weaker topical knowledge for herbivore. As a result, TopicK selects the herbivore-related demonstration (2-shot PPL: 1.630) over the omnivore-related one (2-shot PPL: 1.820, if the Top-2 and Top-3 demonstrations are swapped).