A Categories and Topic Distribution Analysis

We analyze the distribution of topics across BISAC (Martínez-Ávila 2016) categories and their length characteristics. As shown in Figure 1(a), the distribution of the top 15 BISAC categories utilized by LiteLong reveals a notable diversity and balance across domains such as science, economics, and technology. The strategic allocation of topics across these diverse categories ensures comprehensive knowledge coverage while maintaining proportional representation of critical domains. This deliberate topical diversity is a cornerstone of LiteLong's effectiveness, enabling robust performance across varied domains.

Figure 1(b) shows the distribution of topic lengths, which is right-skewed with most topics comprising 2 to 4 words (e.g., 11,048 topics of length 2). While shorter topics (1-word: 1,830) and longer topics (up to 11 words) are present, their frequencies are lower. This prevalence of moderately concise topics, complemented by a range of shorter and longer definitions, likely enhances LiteLong's ability to discern information at varying granularities within long documents. Such a distribution fosters both focused conceptual learning and the capacity to understand more complex thematic structures, contributing to improved comprehension.

B Task-Level Evaluation Details for Long-context benchmarks in Ablation Experiments

This section provides comprehensive task-level performance breakdowns and detailed experimental results that supplement the ablation studies presented in the main paper.

B.1 Detailed Comparison Between BISAC and Automatically Generated Categories

We present a detailed performance comparison of two topic category systems—GPT-4o-generated categories and the BISAC taxonomy—across the HELMET (Yen et al. 2025) and RULER (Hsieh et al. 2024) benchmarks. As shown in Table 1, the BISAC-based categories achieve the highest overall performance.

B.2 Detailed Results for Different Topic Retention Strategies

We compare three topic retention strategies—Keep-Accept, Keep-Fixed-K-Accept, and Filter-Reject—on the HELMET and RULER benchmarks. As shown in Table 2, the Filter-Reject strategy, which removes low-quality topics using a judge model, achieves the best overall performance.

B.3 Detailed Results for Different Topic Generators

We evaluate the average performance of different LLMs for topic generation on the HELMET and RULER benchmarks, without using the multi-agent debate mechanism. As shown in Table 3, Qwen2.5-7B (Yang et al. 2024) achieves the best overall performance across tasks.

B.4 Detailed Supervised Finetuning Results

We report the results of NExtLong (Gao et al. 2025b) and LiteLong+NExtLong on the LongBench v2 (Bai et al. 2024) benchmark after supervised fine-tuning with a 512K context length. Detailed metrics are provided in Table 4.

B.5 Detailed Short-context Performance

The detailed performance of LiteLong on short-context evaluation tasks is presented in Table 5.

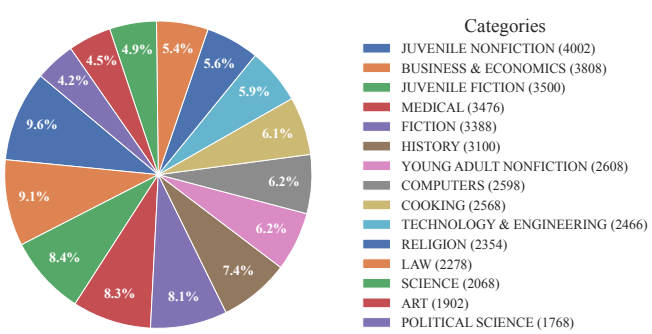
C Combined with UtK Long-dependency Enhancement Method

We further evaluate the performance of LiteLong combined with the Untie the Knots (UtK) (Tian et al. 2024) approach on the HELMET and RULER benchmarks. As shown in Table 6, the results indicate that while the combination of LiteLong and UtK performs well under a 64K context length—effectively enhancing long-range dependency—the addition of UtK does not yield improvements at the 128K length and may even lead to a decline in performance. This suggests that while both UtK and NExtLong are effective at improving long-range dependency at shorter context lengths, UtK does not maintain its effectiveness as the context length increases to 128K, unlike NExtLong.

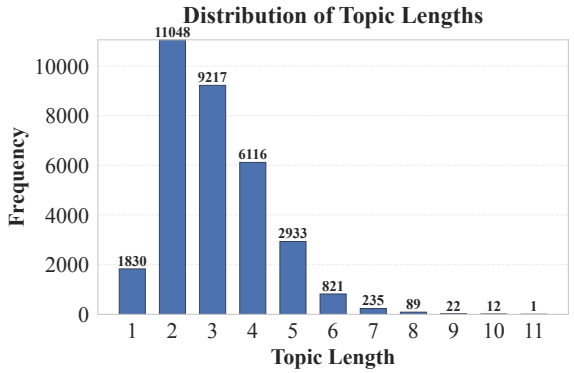
D Combined with NExtLong Long-dependency Enhancement Method

We further evaluate the performance of LiteLong combined with BM25-NExtLong, a computationally efficient variant of NExtLong that uses BM25 for similarity scoring instead of dense embeddings, thereby avoiding the need to generate and store document embeddings. As shown in Table 7, both LiteLong+BM25-NExtLong and LiteLong+NExtLong yield consistent improvements over the LiteLong baseline at the 64K context length, with LiteLong+NExtLong achieving the highest average score (50.69), followed closely by LiteLong+BM25-NExtLong (50.67). This suggests that both methods effectively enhance long-dependency understanding in medium-length contexts, and BM25-NExtLong offers a resource-friendly alternative without sacrificing performance at this scale.

However, at 128K context length, a divergence emerges: LiteLong+NExtLong continues to improve and achieves the best overall result (63.04), while LiteLong+BM25-NExtLong slightly underperforms relative to the LiteLong baseline (61.37 vs. 61.90). This indicates that while BM25-NExtLong is effective at shorter lengths, its reliance on surface-level lexical matching may limit its utility in longer contexts, where deeper semantic understanding becomes increasingly important. In contrast, the embedding-based NExtLong method demonstrates greater robustness and scalability, maintaining its benefits even as the context window expands.



(a) Top 15 category distribution



(b) Topic length distribution

Figure 1: Category and topic distribution analysis. Figure (a) shows diverse and balanced category coverage, while figure (b) shows a right-skewed topic length distribution with most topics comprising 2–4 words.

Method	Recall	RAG	ICL	Rerank	LongQA	RULER	Avg
GPT4-o Categories	80.6	**60.65**	77.36	25.84	**33.37**	81.85	59.94
BISAC	**83.23**	60.43	**80.12**	**30.73**	33.01	**83.88**	**61.90**

Table 1: Task-level HELMET performance and overall RULER scores under different category systems.

Method	Recall	RAG	ICL	Rerank	LongQA	RULER	Avg
Keep-Accept	82.06	60.4	79.12	29.29	**33.92**	83.27	61.34
Keep-Fixed-K-Accept	82.59	**61.0**	**82.0**	27.48	33.55	82.65	61.54
Filter-Reject	**83.23**	60.43	80.12	**30.73**	33.01	**83.88**	**61.90**

Table 2: Task-level HELMET performance and overall RULER scores using different topic retention strategies.

Method	Recall	RAG	ICL	Rerank	LongQA	RULER	Avg
Mistral	82.26	**62.08**	**82.88**	26.58	33.87	82.18	61.31
Gemma	81.04	61.80	78.40	**28.92**	**34.08**	82.06	61.05
Qwen	**84.40**	60.71	81.96	26.05	31.93	**83.65**	**61.45**

Table 3: Task-level HELMET performance and overall RULER scores using different single-model topic generators without the multi-agent debate mechanism.

Model	Easy	Hard	Short	Medium	Long	Avg
SOTA	33.3	**28.6**	32.2	**30.7**	26.9	30.4
+LiteLong	**34.9**	28.0	**33.3**	29.7	**28.7**	**30.6**

Table 4: Detailed performance comparison on LongBench v2 benchmarks after supervised fine-tuning to 512k context length.

Model	ARC-c	ARC-e	HellaSwag	Logiqa	PIQA	WinoGrande	Avg
Llama3-8B-base	50.34	80.18	**60.13**	**27.5**	79.60	72.85	**62.05**
+ LiteLong	**51.20**	**80.39**	59.56	**27.5**	**79.87**	**73.24**	61.92

Table 5: Detailed performance comparison on short-context benchmarks. The results demonstrate that LiteLong maintains comparable performance to the base model across multiple tasks.

Length	Model	Recall	RAG	ICL	Rerank	LongQA	RULER	Avg
64K	UtK	66.36	49.94	**64.6**	17.18	25.82	61.6	47.58
	LiteLong	68.24	49.53	61.72	19.25	25.74	66.87	48.56
	LiteLong+UtK	**72.35**	**49.98**	64.44	**19.49**	**25.95**	**68.01**	**50.04**
128K	UtK	60.49	60.98	63.08	16.96	35.7	73.73	51.82
	LiteLong	**83.23**	60.43	**80.12**	**30.73**	33.01	**83.88**	**61.90**
	LiteLong+UtK	78.23	**61.52**	77.72	25.34	**37.03**	81.19	60.17

Table 6: Task-level HELMET performance and overall RULER scores : LiteLong+UtK enhances long-dependency at 64K but not at 128K, where performance may decline.

Length	Model	Recall	RAG	ICL	Rerank	LongQA	RULER	Avg
64K	LiteLong	68.24	49.53	61.72	19.25	**25.74**	66.87	48.56
	LiteLong+BM25-NExtLong	73.74	50.18	**64.92**	**22.92**	25.21	**67.03**	50.67
	LiteLong+NExtLong	**74.95**	**50.38**	64.32	22.51	25.44	66.52	**50.69**
128K	LiteLong	**83.23**	60.43	**80.12**	30.73	33.01	**83.88**	61.90
	LiteLong+BM25-NExtLong	80.73	**61.37**	77.96	28.28	**37.87**	82.04	61.37
	LiteLong+NExtLong	82.93	60.81	**80.12**	**33.68**	36.97	83.73	**63.04**

Table 7: Task-level HELMET performance and overall RULER scores : LiteLong+BM25-NExtLong enhances long-dependency at 64K but not at 128K, where performance may decline.The NExtLong method continues to improve model performance at 64K and 128K context lengths

Structured Topic Generation Prompt

In the field of {primary_category}, list {num_to_generate} subtopics in {secondary_category} and provide a brief explanation of each.
Return the output strictly as a JSON array of objects, using the following format:

```
[
  {
    "topic": "...",
    "explanation": "..."
  },
  {
    "topic": "...",
    "explanation": "..."
  }
  ...
]
```

Figure 2: Prompt structure for the topic generate stage.

E Resource Consumption Calculation Details

We compare disk usage and computational time across several long-context data construction methods, including LiteLong, Quest, KNN/ICLM, NExtLong, and LiteLong-NExtLong.

LiteLong. Before applying BM25 retrieval, LiteLong builds an inverted index, which is slightly larger than the raw corpus. The index occupies 0.92 TB in total, consisting of 0.77 TB from FineWeb-Edu and 0.15 TB from Cosmopedia V2.

For topic generation using a vLLM-deployed model, the overall time is approximately 6 hours, which includes 2.5 hours for topic generation, 2.5 hours for topic critique, and 1 hour for judge evaluation.

Quest. Similarly to LiteLong, Quest (Gao et al. 2025a) builds an inverted index of size 0.92 TB. Additionally, it uses a doc2query model to generate one synthetic query per document. The corpus contains 229,302,005 documents from FineWeb-Edu and Cosmopedia V2.

To generate queries for document embedding, processing the FineWeb-Edu dataset requires 702 hours, while the CosmopediaV2 dataset takes 104 hours. In total, query generation across both datasets requires 806 hours.

KNN and ICLM. These methods (Guu et al. 2020; Levine et al. 2022; Shi et al. 2024) do not require inverted indices. Instead, they encode each document into a 1024-dimensional float32 embedding. The total storage required is 0.89 TB.

Document embedding for the FineWeb-Edu dataset requires 565 hours, while the CosmopediaV2 dataset takes 52 hours. In total, query generation across both datasets requires 617 hours.

NExtLong. NExtLong (Gao et al. 2025b) segments each document into 2048-character chunks, without inverted indexing. A total of 644,275,250 chunks are produced, which requires 2.4 TB of storage. The embedding computation takes approximately 928 hours.

LiteLong + NExtLong. In this hybrid approach, LiteLong performs document-level retrieval, and NExtLong encodes

Topic Critique Prompt

You are an expert in semantic analysis, topic modeling, and educational content classification.
You are given a list of subtopics generated by another model for the following BISAC topic:

- **Top Category**: {primary_category}
- **Subcategory**: {secondary_category}

Each subtopic contains:

- Topic Name: name
- Topic Explanation: generate explanation

For example, for each generated topic i:

```
generate topic i:
Topic Name: <topic name>
Topic Explanation: <topic explanation>
```

Your task is to critically evaluate each proposed subtopic for its quality, distinctiveness, and strategic usefulness in semantic clustering and topic seeding for educational datasets. Assess each subtopic using the following rigorous criteria:

1. **Topical Relevance** — Is the subtopic clearly and directly aligned with its assigned primary or secondary educational category?
2. **Semantic Distinctiveness** — Does it introduce a meaningfully different angle (e.g., audience, concept, application, cultural lens), or is it redundant with existing entries?
3. **Diversity & Complementarity** — Does this subtopic expand the conceptual space covered by the full set, providing non-overlapping, complementary perspectives?
4. **Anchor Value for Clustering** — Can this subtopic act as a strong semantic node or seed in downstream clustering tasks for large-scale educational content?

Reject vague, overly broad, redundant, or weakly anchored subtopics. Prioritize those that enhance overall topical diversity, fill semantic gaps, and support meaningful partitioning in topic models. Be selective and analytical—do not default to acceptance.

—

Return your critique in the following JSON structure:

```
[
    {
        "accepted": [
            {
            "topic": "generate name",
            "explanation": "generate explanation",
            "reason": "accept reason"
            },
            ...
        ],
        "rejected": [
            {
            "topic": "generate explanation",
            "explanation": "generate explanation",
            "reason": "reject reason"
            },
            ...
        ]
    }
]
```

Figure 3: Prompt structure for the topic critique stage.

Judge Model Prompt

You are a fair and insightful topic evaluation judge.
Your task is to assess a set of topic candidates generated by AI models for the following category:

- **Primary Category (BISAC)**: {primary_category}
- **Secondary Category**: {secondary_category}

Each topic has been reviewed by peer models. Use their evaluations as **reference only** — do not blindly accept or reject based on their suggestions. Apply your own comprehensive judgment to determine which topics are **worth keeping**.
Here are the topic candidates and critiques:

```
Topic 1: Title: <title>
Explanation: <explanation>
Suggested Action: <Keep / Reject>
Critique Reason: <reason>

... (similar entries for other topics)
```

Evaluation Guidelines:

- You are not required to select a fixed number of topics. Instead, remove any topics that meet the criteria below.
- Do not fully accept or reject topics solely based on peer critique. Use critique as insight, not as ground truth.
- The goal is to retain a set of topics that is diverse, well-balanced, and representative of the overall category.

Removal Criteria:

1. Redundancy / Lack of Diversity — Remove topics too similar to others without offering a distinct angle or sub-theme.
2. Insufficient Contribution to Coverage — Remove topics that do not expand the content space meaningfully or are too narrow to add value.
3. Low Relevance — Remove topics not clearly aligned with the specified categories.
4. Lower Quality in Similar Groups — Among similar topics, retain the one with clearer wording, stronger conceptual value, or broader applicability.

Return your decision (remove topics you select) using the following JSON format:

```
{
  "rejected_topics": [
    {
      "title": "...",
      "explanation": "...",
      "source": "Generated Model source",
      "critique reason": "...",
      "judge reason": "Reason for removal"
    },
    ...
  ],
  "summary": "Summarize your overall decision-making approach. "
}
```

Figure 4: Prompt structure for the topic judge stage.

the resulting chunks. The inverted index still occupies 0.92 TB. The 117,522,683 retrieved chunks require 0.44 TB, yielding a total of 1.36 TB storage. The total embedding time is 170 hours.

F Prompts Used for Topic Generation and Retention

We present the prompt templates used in our system, covering three functional roles: topic generation, topic critique, and topic quality judgment. These prompts are designed to support both diversity and precision in topic selection through multi-agent collaboration and filtering.

F.1 Prompt Used for Topic Generation

The prompt used for topic generation is directly adopted from CosmopediaV2 (Ben Allal et al. 2024) without modification. We also follow the same inference hyperparameters as in CosmopediaV2, with temperature set to 0.6 and top-p set to 0.95. As shown in Figure 2, the model is expected to return a JSON array of topic–explanation pairs.

F.2 Prompt Used for Topic Critique

The prompt used for generating the *critique* is shown in Figure 3. We set the generation hyperparameters to a temperature of 0.6 and a top-p of 0.95, following the same configuration used in other components.

F.3 Prompt Used for Topic Judge

The prompt used for the judge model is shown in Figure 4, with generation parameters set to temperature 0.6 and top-p 0.95.

G Baseline Methods Details

- **Standard Method** shuffles and concatenates documents randomly in the input context and has been the mainstream practice in pre-training (Ouyang et al. 2022). This approach is computationally efficient but often produces incoherent transitions between documents, which may limit the model's ability to develop robust long-range understanding.
- **KNN** (Retrieval-augmented Language Model Pretraining) (Guu et al. 2020; Levine et al. 2022) places each document along with the top k most similar retrieved documents in the same input context. Although this method improves semantic coherence in training samples, it requires building large embedding indices that demand substantial computational resources.
- **ICLM** (Shi et al. 2024) is a recently proposed method that utilizes a traveling salesman algorithm to alleviate the document redundancy problem in the KNN method by ranking similarities and determining the optimal training path. This approach aims to maximize contextual similarity while ensuring unique document inclusion, but still requires significant computational resources for retrieval and indexing.
- **Quest** is a query-centric data synthesis method aggregating semantically relevant yet diverse documents (Gao et al. 2025a). Quest uses a generative model to predict potential queries for each document, grouping documents with similar queries and keywords. While effective for creating coherent long-context training data, this approach introduces substantial computational overhead for query generation and embedding-based retrieval.
- **NExtLong** is a novel framework for synthesizing long-context data through Negative document Extension (Gao et al. 2025b). NExtLong decomposes a document into multiple meta-chunks and extends the context by interleaving hard negative distractors retrieved from pretraining corpora. This approach compels the model to discriminate long-range dependent context from distracting content, enhancing its ability to model long-range dependencies. However, it requires building extensive vector retrieval databases from the entire pretraining corpus, demanding significant computational and storage resources.

References

Bai, Y.; Tu, S.; Zhang, J.; Peng, H.; Wang, X.; Lv, X.; Cao, S.; Xu, J.; Hou, L.; Dong, Y.; et al. 2024. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*.

Ben Allal, L.; Lozhkov, A.; Penedo, G.; Wolf, T.; and von Werra, L. 2024. SmolLM-Corpus.

Gao, C.; W, X.; Fu, Q.; and Hu, S. 2025a. Quest: Query-centric Data Synthesis Approach for Long-context Scaling of Large Language Model. In *The Thirteenth International Conference on Learning Representations*.

Gao, C.; Wu, X.; Lin, Z.; Zhang, D.; and Hu, S. 2025b. Nextlong: Toward effective long-context training without long documents. *arXiv preprint arXiv:2501.12766*.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, 3929–3938. PMLR.

Hsieh, C.-P.; Sun, S.; Kriman, S.; Acharya, S.; Rekesh, D.; Jia, F.; and Ginsburg, B. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? In *First Conference on Language Modeling*.

Levine, Y.; Wies, N.; Jannai, D.; Navon, D.; Hoshen, Y.; and Shashua, A. 2022. The Inductive Bias of In-Context Learning: Rethinking Pretraining Example Design. In *International Conference on Learning Representations*.

Martínez-Ávila, D. 2016. BISAC: Book Industry Standards and Communications. *Knowledge Organization*, 43(8).

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Gray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Shi, W.; Min, S.; Lomeli, M.; Zhou, C.; Li, M.; Lin, X. V.; Smith, N. A.; Zettlemoyer, L.; tau Yih, W.; and Lewis, M.

2024. In-Context Pretraining: Language Modeling Beyond Document Boundaries. In *The Twelfth International Conference on Learning Representations*.

Tian, J.; Zheng, D.; Cheng, Y.; Wang, R.; Zhang, C.; and Zhang, D. 2024. Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models. *arXiv preprint arXiv:2409.04774*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; and et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yen, H.; Gao, T.; Hou, M.; Ding, K.; Fleischer, D.; Izsak, P.; Wasserblat, M.; and Chen, D. 2025. HELMET: How to Evaluate Long-context Models Effectively and Thoroughly. In *The Thirteenth International Conference on Learning Representations*.