Quest: Query-centric Data Synthesis Approach for Long-context Scaling of Large Language Model

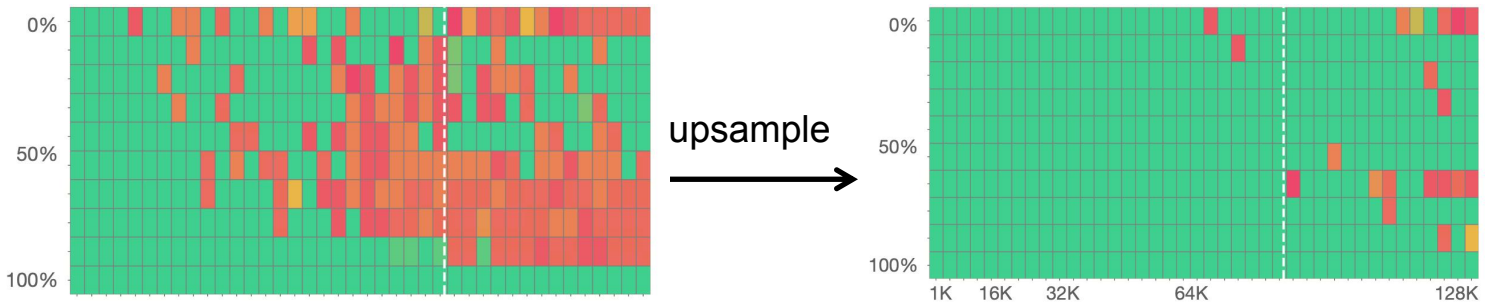
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Chaochen Gao · Xing W · Qi Fu · Songlin Hu

Motivation

- Large Language Models (LLMs) are typically pre-trained using fixed context lengths. Recent advancements, however, have highlighted the importance of extending the context lengths.

- Previous works select such data by filtering long documents from the training set that fit the target context length. For instance, Data engineering [1] demonstrated that increasing the proportion of long-text data can lead to better long-text recall capabilities.

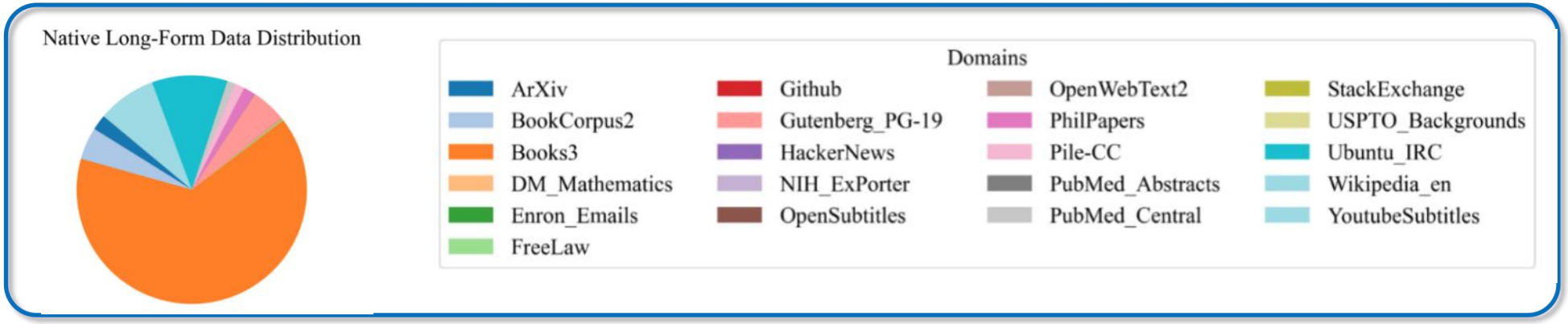


✓ **The model's performance in the needle-in-a-haystack task before and after upsampling long-text data, with green indicating the model's answers are entirely correct.**

[1] Fu Y, Panda R, Niu X, et al. Data engineering for scaling language models to 128k context. In Proceedings of the 41st International Conference on Machine Learning (ICML'24).

Motivation

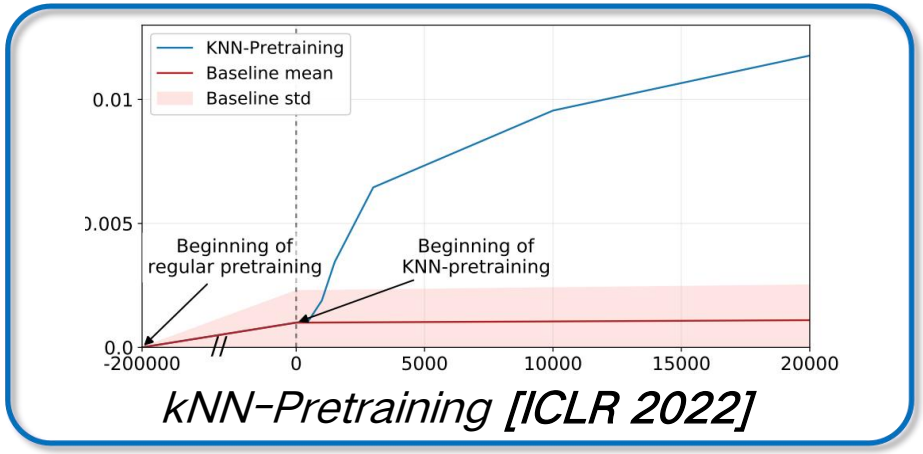
However, those documents often come from a few specific domains like Books3 or Arxiv, leading to a skewed distribution, which impacts model performance after continued training.



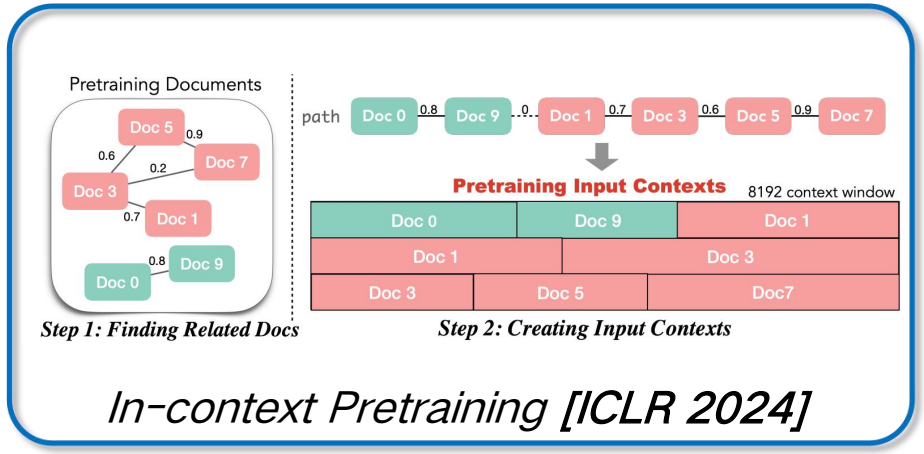
Previous studies synthesise long-context data by concatenating shorter documents to achieve a balanced domain distribution. Those methods can be classified into two categories: the Standard method and similarity-based methods like KNN.

Motivation

- *Standard* methods
  - The Standard method randomly concatenates short documents to meet a specified target length.
- Similarity-based methods
  - similarity-based methods aggregate semantically similar documents, e.g., by concatenating a document with its top-k most similar counterparts from the corpus.



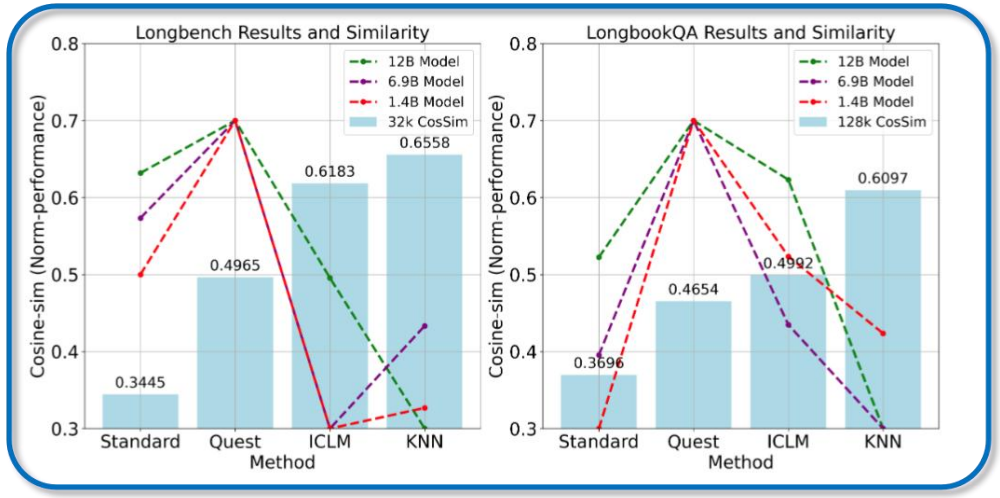
kNN-Pretraining [ICLR 2022]



In-context Pretraining [ICLR 2024]

Motivation

However, similarity-based methods overemphasize semantic correlation. **They are prone to falling into a narrow context (high redundancy) due to concatenating similar or even repeated documents.**

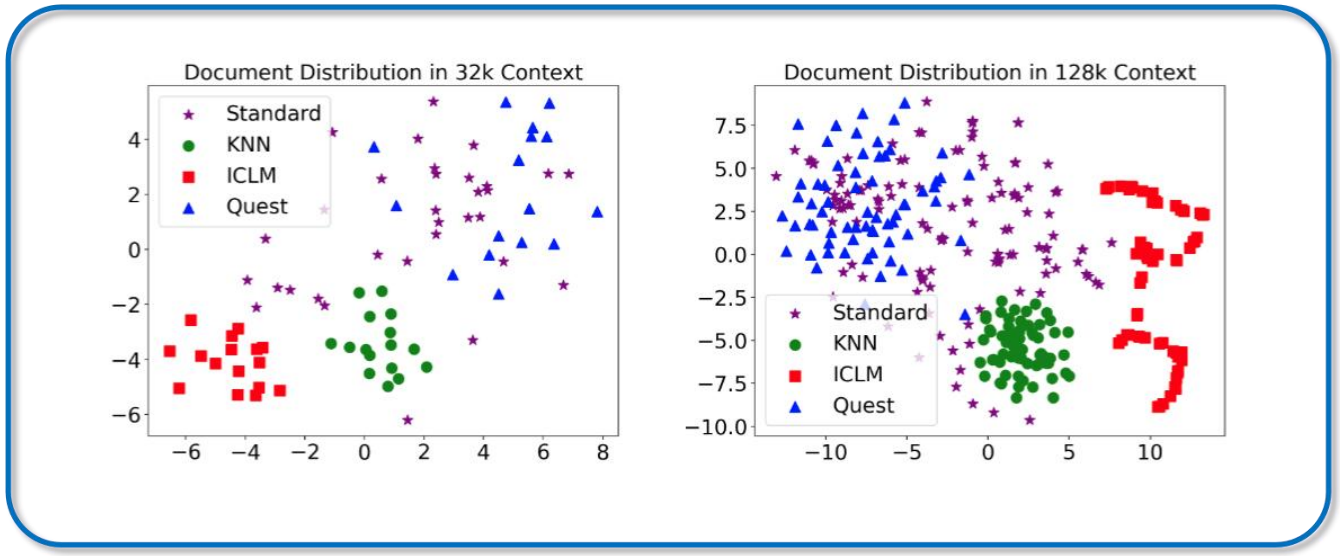


The results show that either prioritizing context diversity at the expense of semantic correlation (Standard) or overemphasizing semantic correlation while sacrificing context diversity (KNN and ICLM) leads to suboptimal performance.

Motivation

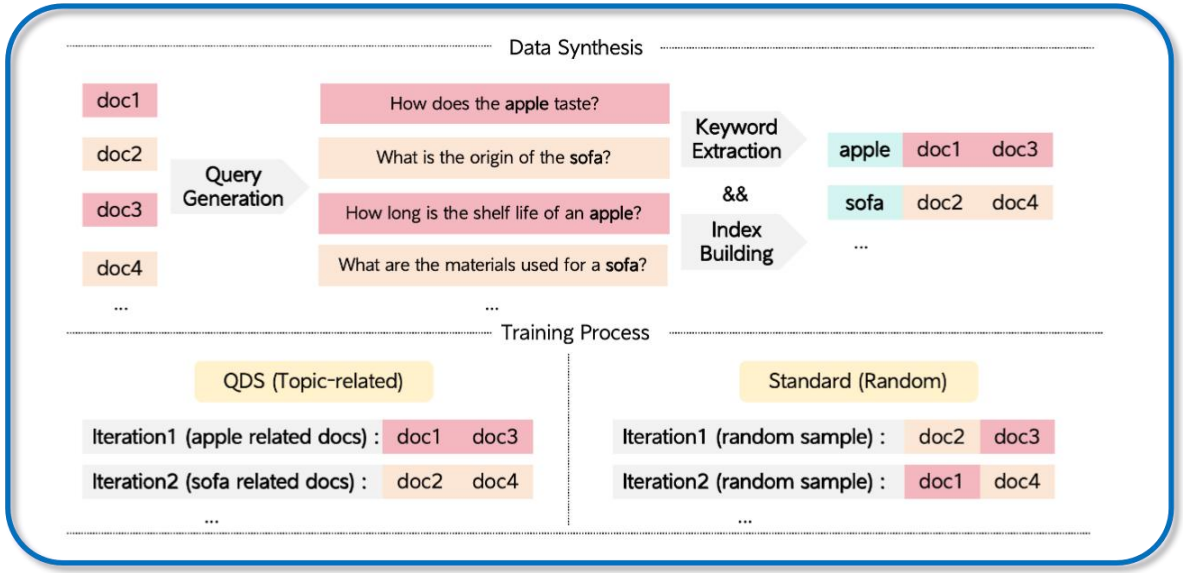
- To achieve the balance effectively, this paper proposes Quest, a query-centric data synthesis approach that simultaneously ensures semantic correlation and context diversity within the long-context data.

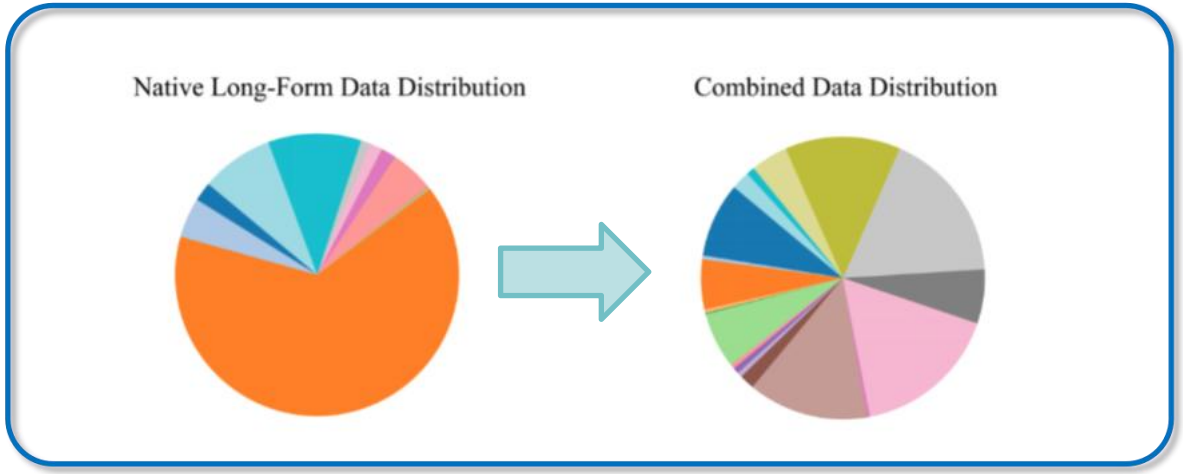
- Our inspiration stems from the observation that similar queries can aggregate semantic relevant but low-redundancy documents via search engines.



Method

- 1. Query Prediction: Predict a query for each document in the dataset.
- 2. Keyword Extraction: Extract keywords for each query.
- 3. Index Construction: Build an inverted index based on the keywords.
- 4. Index Partitioning: Oversample low-frequency keywords to achieve a more balanced data distribution.
- 5. Synthetic Long-Text Data Training: Concatenate documents under each keyword to form long-text data for training.





Result

Quest demonstrates better performance on average.

- ✓ On the 32k Longbench test set, Quest achieved the best average performance across different model sizes.
- ✓ On the 128k LongBookQA test set, Quest also attained the highest performance.

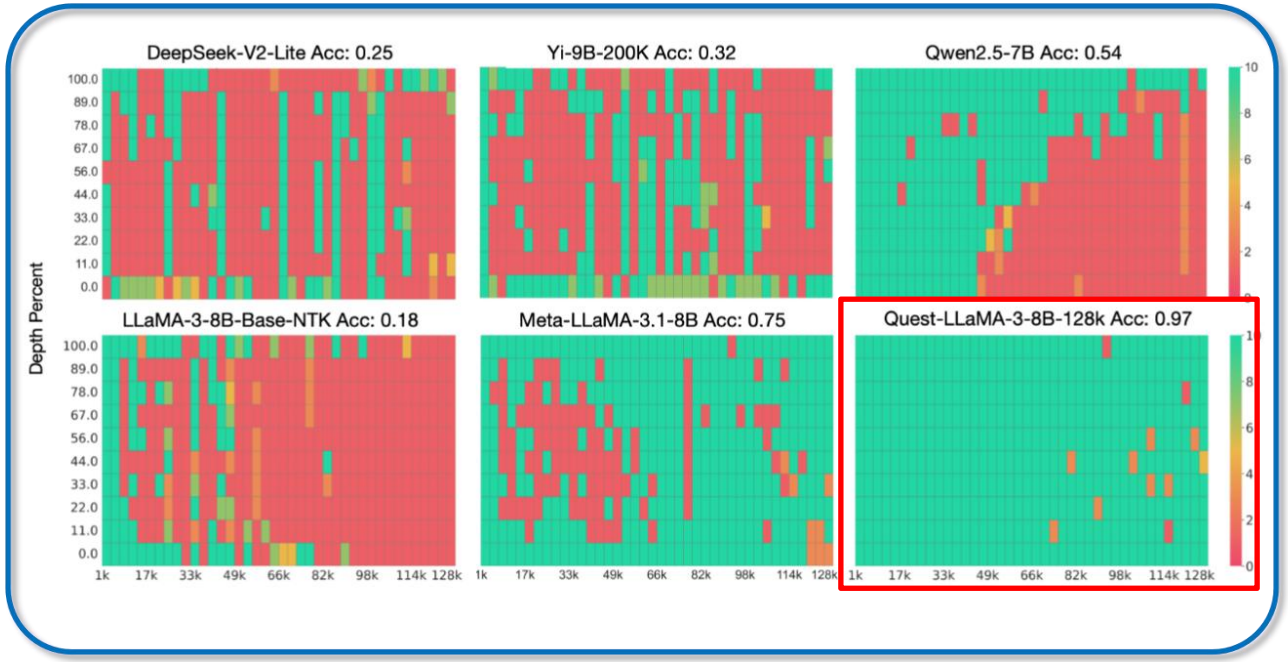
Table 1: Comparison of Longbench results across methods. "Avg." represents the average over multiple test sets. The proposed Quest consistently outperforms baseline methods across various model sizes in the 32K context length setting. Detailed results can be found in Appendix B.

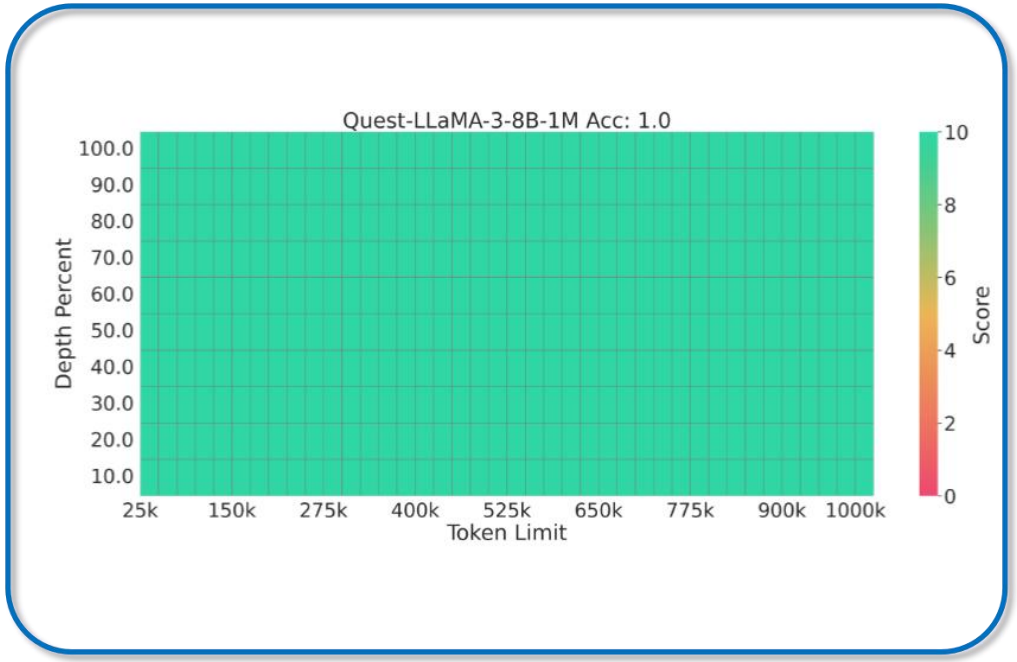
Train&Test	Model size	Method	Avg.	Sgl.	Multi.	Sum.	Few.	Syn.	Code.
32k	1.4B	*Standard*	20.94	19.24	17.46	20.65	26.75	2.04	36.41
		KNN	19.97	17.26	13.01	**22.97**	24.16	2.33	39.22
		ICLM	19.82	**20.01**	14.71	21.95	23.09	1.94	35.31
		Quest	**22.06**	17.97	**17.98**	21.91	**28.06**	**2.33**	**42.25**
32k	6.9B	*Standard*	22.48	18.07	**16.83**	22.33	**30.23**	**3.86**	40.91
		KNN	21.65	18.5	13.64	22.56	28.18	3.76	41.88
		ICLM	20.86	17.82	15.34	22.35	25.86	1.21	41.15
		Quest	**23.23**	**19.21**	14.13	22.45	30.14	2.96	**50.55**
32k	12B	*Standard*	24.85	22.18	21.94	22.30	**32.05**	**3.78**	43.73
		KNN	22.95	20.55	20.48	23.51	29.19	2.47	37.44
		ICLM	24.07	22.67	**23.29**	23.41	30.99	1.5	37.09
		Quest	**25.24**	**22.34**	21.08	**23.74**	31.91	3.22	**46.8**

Train&Test	Model size	Method	Longbook QA
128k	1.4B	*Standard*	9.94
		KNN	10.36
		ICLM	10.70
		Quest	**11.30**
128k	6.9B	*Standard*	14.47
		KNN	13.38
		ICLM	14.92
		Quest	**17.95**
128k	12B	*Standard*	17.81
		KNN	16.42
		ICLM	18.44
		Quest	**18.92**

Result

- ✓ Quest-LLaMA3-8B achieves a 97\% accuracy on the Needle-in-a-Haystack task (retrieving a text sentence).
- ✓ Quest is the first base model (without instruction tuning) to achieve 100% accuracy with a 1M context length. (retrieving a number string)





中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

Thanks!