Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continued Pre-training of Large Language Models

Shengjie Ma[1,3*], **Xuhui Jiang**[1*], **Chengjin Xu**[1*],
Cehao Yang[2], **Liyu Zhang**[2], **Jiaxin Mao**[3†], **Jian Guo**[2†]
[1] DataArc Tech Ltd., Shenzhen, Guangdong, China
[2] IDEA Research, International Digital Economy Academy, Shenzhen, Guangdong, China
[3] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

Abstract

Large Language Models (LLMs) have achieved remarkable success but remain data-inefficient, especially when learning from small, specialized corpora with limited and proprietary data. Existing synthetic data generation methods for continue pre-training focus on intra-document content and overlook cross-document knowledge associations, limiting content diversity and depth. We propose Synthetic-on-Graph (SoG), a synthetic data generation framework that incorporates cross-document knowledge associations for efficient corpus expansion. SoG constructs a context graph by extracting entities and concepts from the original corpus, representing cross-document associations, and employing a graph walk strategy for knowledge-associated sampling. This enhances synthetic data diversity and coherence, enabling models to learn complex knowledge structures and handle rare knowledge. To further improve the quality of synthetic data, we integrate two complementary strategies, Chain-of-Thought (CoT) and Contrastive Clarifying (CC), to enhance both reasoning capability and discriminative power. Extensive experiments demonstrate that SoG surpasses state-of-the-art (SOTA) methods on multi-hop and domain-specific question answering, while achieving competitive performance on long-context reading comprehension. These results highlight the superior generalization ability of SoG. Our work advances the paradigm of synthetic data generation and offers practical solutions for efficient knowledge acquisition in LLMs, particularly for downstream tasks and domains with limited training data.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved groundbreaking advancements in the field of Natural Language Processing (NLP), demonstrating the ability to acquire knowledge from unstructured text and perform complex, knowledge-intensive tasks [1]. These models have exhibited exceptional performance across various applications, including question-answering systems, machine translation, and conversational agents. This success is largely attributed to the next-word prediction objective [2] combined with vast amounts of internet data [3]. However, despite these achievements, there remains a significant inefficiency in data utilization [4].

This data inefficiency becomes particularly pronounced when models need to learn from small-scale, high-value corpora. With the increasing demand for proprietary domain knowledge, models are required to efficiently acquire information from limited data sources. For instance, in specialized fields such as medicine, law, or specific technological domains, the available data is not only limited

[*]Both authors contributed equally to this research.
[†]Corresponding author

Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continued Pre-training of Large Language Models. *Proceedings of the Fourth Learning on Graphs Conference (LoG 2025)*, PMLR 269, Arizona State University, Phoenix, USA, December 10–12, 2025.

but often proprietary. In such cases, traditional large-scale pretraining methods are inapplicable due to the unavailability of sufficient training data [5].

Moreover, recent studies have revealed limitations in the current pretraining paradigm. For example, models struggle when learning simple relations and require a large number of repeated instances to effectively learn facts [6]. These issues become more acute when dealing with long-tail data or rare knowledge, as such information appears with extremely low frequency in large-scale corpora [5].

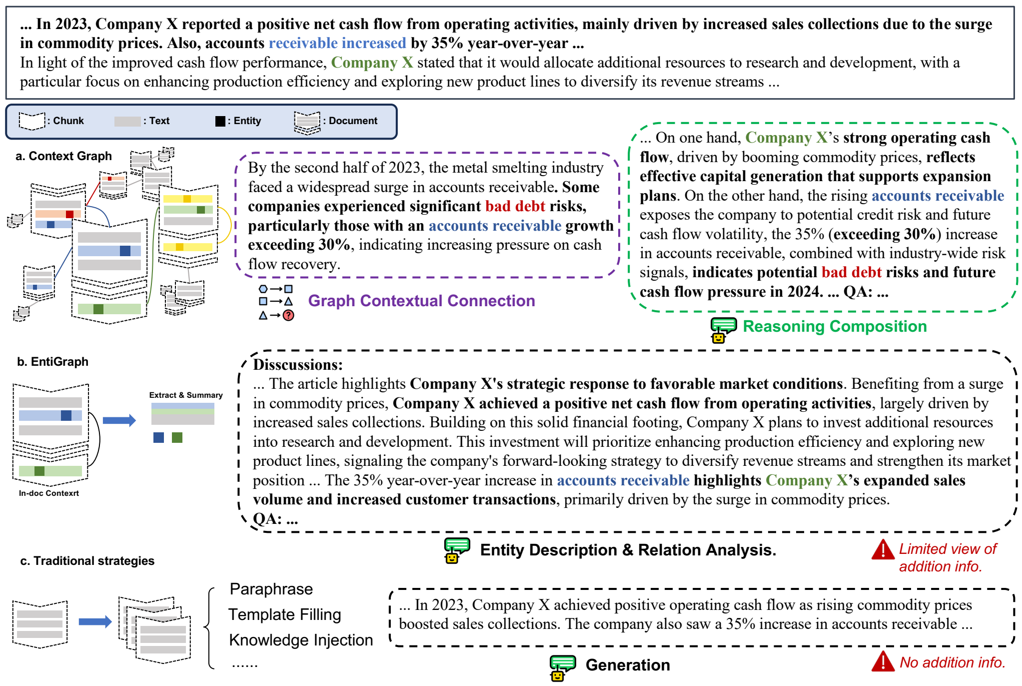


Figure 1: Comparison of the Proposed Context Graph for Synthetic Generation with Other Generation Strategies: a. Context Graph in SoG. b. Intra-document graph in EntiGraph, where the knowledge view is confined within a single document. c. Traditional synthetic generation methods, which struggle to incorporate extra knowledge.

To address the challenge of efficiently acquiring knowledge from small-scale corpora, synthetic data generation methods have been proposed for continued pretraining of models. They aim to expand the original limited data by generating diverse synthetic corpora, thereby improving the learning efficiency and performance of the models. For instance, the EntiGraph method decomposes the text corpus into a list of entities and generates descriptions about the relationships between entities, attempting to populate the underlying knowledge graph of the corpus [7]. However, as shown in Figure 1b this approach primarily focuses on intra-document content, neglecting inter-document knowledge associations. This leads to limitations in the content diversity and knowledge depth of the synthetic data. In reality, knowledge is often interconnected across documents and domains. Relying solely on entity combinations within a single document fails to capture the full spectrum of knowledge. Additionally, the lack of cross-document synthetic data constrains the model's ability to handle complex, multi-hop problems that require integrating information from multiple documents to derive an answer. For instance, in the context graph in Figure 1a, the first encountered literature primarily describes Company X's positive financial report and active market plans in 2023. However, relying on the across-document information associated with the entity "accounts receivable" —"companies with accounts receivable growth exceeding 30% face a special risk of bad debt" —we can derive a broader understanding of the literature: despite the positive net cash flow, people are suggested to be particularly cautious about the potential bad debt risk associated with Company X's 35% accounts receivable growth. Cross-document information can integrate multi-dimensional perspectives on a topic (both positive and negative), build a progressive chain of information, and uncover implicit phenomena — integrating knowledge in a way that uncovers more than what each document alone can offer, where "1+1>2".

To this end, we propose the Synthesize-on-Graph (SoG) framework—a context-graph-enhanced synthetic data generation method designed to provide an efficient solution for continued pretraining of LLMs. The core idea of SoG is to incorporate cross-document knowledge associations by constructing and leveraging a context graph to expand the original corpus effectively.

Specifically, SoG comprises two key components: (1) *Context Graph Construction and Cross-Document Sampling*: We build a context graph from entities and concepts extracted from the original corpus, representing cross-document knowledge associations. Using this graph, we apply a two-stage cross-document sampling strategy: first, random walks guided by document retrieval to achieve cross-document exploration, enhancing data diversity while preserving coherence and knowledge associations. This helps the model learn complex knowledge structures, especially for long-tail entities. Second, Secondary Sampling and Controlled Allocation help balance the knowledge distribution and support flexible data customization. (2)*Combined Chain-of-Thought and Contrastive Clarifying Synthesize*: We combine Chain-of-Thought and Contrastive Clarifying to enhance synthetic data quality. CoT guides the model to generate logical chains, improving depth and interpretability, while contrastive generation boosts the discriminative knowledge in the synthetic data.

Through extensive experiments, our approach outperforms existing state-of-the-art (SOTA) methods on multi-hop document and domain-specific question answering tasks, while achieving comparable results on long-context reading comprehension. Also, we demonstrate better generalization capability over the SOTA method. The introduction of the SoG framework marks a significant advancement in synthetic data generation and continued pertaining (CPT) for LLMs, providing new directions and possibilities for future research. Our work not only drives the development of synthetic data techniques but also offers new perspectives for optimizing the training of LLMs.

2 Related Work

This section presents an overview of recent developments in synthetic data generation for the pretraining of large language models (LLMs). Synthetic data generation has emerged as a crucial area of research, with various strategies proposed to enhance the diversity and effectiveness of training datasets. A significant trend in this domain is the adoption of hierarchical prompting to generate targeted synthetic content. For instance, [8] utilize API-based LLMs to create children's stories driven by specific keywords, illustrating that even smaller language models can yield fluent narratives when pre-trained on such datasets. [9] achieve automatic analysis and annotation on complex data in the legal domain by using a modular multi-process pipeline, along with the injection of expert knowledge in the form of few-shot learning into each submodule. This approach was used for both pretraining and fine-tuning. This underscores the potential of hierarchical prompting in producing effective and relevant training data.

In another vein, [10] generate diverse educational content, such as textbooks and coding exercises, by conditioning on attributes like topic, audience, and function names. The datasets generated from this method have supported the development of robust LLMs, as further explored in subsequent studies [11, 12]. However, these approaches are often hindered by a lack of public accessibility to the datasets and prompt strategies, limiting reproducibility and broader community progress. Similarly, [13] focus on rephrasing existing documents to generate new training data, reporting enhancements in training efficiency through these modified versions.

While these efforts have significantly advanced the field, they primarily focus on generating intra-document content, thereby overlooking the importance of cross-document knowledge associations. This oversight limits the diversity and depth of the synthetic content, which is crucial for developing LLMs capable of understanding and integrating complex knowledge structures. The prevailing focus on intra-document generation underscores the need for novel methodologies that can address these gaps by synthesizing data that not only maintains coherence but also captures broader, interconnected knowledge domains.

Current efforts [14, 15] explore synthetic QA generation for task-specific finetuning, reflecting an emerging interest in incorporating knowledge-aware strategies into data generation. Although such strategies have demonstrated benefits for specific QA tasks, their applicability remains limited for more general-purpose tasks, indicating a gap that could potentially be filled by new data generation approaches that are untethered to any particular downstream application.

Moreover, [16] explore continued pretraining of Llama 2 models using synthetic paraphrases of Wikipedia articles, with mixed results regarding performance improvements. This suggests limitations in relying solely on paraphrasing techniques to enhance model knowledge and underscores the need for research into more robust methods that can generate synthetic data with greater diversity and depth.

3 Methodology

We propose the SoG framework, a context-graph-enhanced synthetic data generation method designed to address limitations in content diversity and knowledge association found in existing approaches. The framework achieves this by leveraging cross-document, knowledge-associated sampling, enabling the integration of information across multiple sources. Additionally, it conducts a combined data synthesis approach based on Chain-of-Thought reasoning and Contrastive Clarifying analysis, which enhance generation models' ability to reason and distinguish between complex knowledge. The following sections provide a detailed overview of the SoG framework, highlighting three core components: Context Graph Construction, Cross-Document Sampling, and Generation Strategies.

$$\begin{aligned}\text{Corpus} &\xrightarrow{\text{Construction}} \text{Context Graph} \xrightarrow{\text{Context-graph Traversal}} \text{Path Set } \mathcal{P} \\ &\xrightarrow{\text{Secondary Sampling}} \text{Balanced Path Set } \mathcal{P}^* \xrightarrow{\text{Generation Strategies}} \text{Synthetic Data}\end{aligned} \tag{1}$$

The overall generation process and context graph building of SoG is shown in Figure 1a and Figure 2.

3.1 Context Graph Construction

3.1.1 Entity Extraction

First, given a corpus $\mathcal{C} = \{d_i\}, i \in [0, N)$, each document d_i is divided into several paragraphs $p_{i,j}$, where j denotes the j-th paragraph of document i. Subsequently, we prompt the LLM to identify key entities within each paragraph as $\mathcal{E}_{i,j} \in \mathcal{E}$, where $\mathcal{E}$ denote the extracted entities from the entire corpus $\mathcal{C}$.

3.1.2 Entity-Context Mapping

For each entity $e_k \in \mathcal{E}$, we collect all paragraphs in which it appears, denoted as $P_k = \{p_{i,j} \mid e_k \in \mathcal{E}_{i,j}, \forall i, j\}$. This forms an entity-paragraph mapping $M : e_k \mapsto P_k$, where M associates each entity e_k with its corresponding set of paragraphs P_k.

3.1.3 Context Graph

We define a context graph $\mathcal{G} = (\mathcal{E}, E)$, where $\mathcal{E}$ denotes the set of nodes corresponding to all identified entities. The edge set is given by

$$E = \{(e_x, e_y) \mid \exists i, j \ \text{ s.t. } \ e_x, e_y \in \mathcal{E}_{i,j}\},$$

where $\mathcal{E}_{i,j}$ represents the subset of entities co-occurring within a bounded textual unit (e.g., a paragraph or sentence). Thus, an edge between e_x and e_y is induced whenever the two entities are observed to co-occur within the same discourse context. In this way, the graph topology captures *implicit contextual associations* among entities, with co-occurrence serving as a distributional proxy for semantic relatedness.

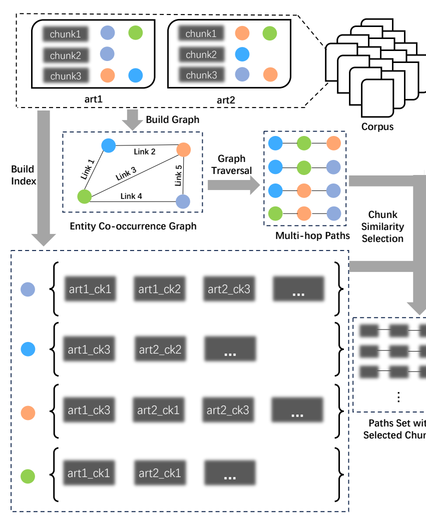


Figure 2: Context Graph Construction and Sampling

3.2 Cross-Document Sampling

3.2.1 Initialization

To enhance content diversity and knowledge association across multiple documents, we implement a cross-document sampling strategy that traverses the constructed context graph $\mathcal{G} = (\mathcal{G}, \mathcal{R})$. Starting from a root entity $e_{root} \in \mathcal{E}$, we perform a breadth-first search (BFS) traversal to collect multi-hop paths that link related entities and their associated text paragraphs across documents. We will traverse all nodes in $\mathcal{G}$ as a root entity.

In addition, for each e_{root}, we traverse all its paragraphs using the entity-context mapping M, which associates entities with the paragraphs in which they appear. If an entity occurs in a large number of paragraphs, we limit the number of starting paragraphs by randomly sampling up to S, a predefined hyperparameter. This step is crucial, as the selected starting paragraphs $p^{(0)}$ serve as references for computing embedding similarities during the traversal process.

Briefly, each entity serves as the root e_{root}. Then a graph traversal is performed up to a maximum of S steps according to the number of paragraphs P_{root} from the mapping M.

3.2.2 Context-graph Traversal

At each traversal, step up to a specified depth D, we explore neighboring entities of the current entity e. The neighbors are defined as:

$$N(e) = \{e' \mid (e, e') \in E\},$$

where (e, e') indicates an edge in the context graph signifying a contextual connection between entities e and e'.

To prioritize neighboring entities with relevant contexts, we introduce a similarity-based selection mechanism. For each neighboring entity e', we compute a similarity score $F_{sim}(q^{(0)}, c)$ between the root paragraph $q^{(0)}$ associated with e_{root} and candidate paragraphs c associated with e'. Using the average node degree d as an upper bound, when traversing the neighbors of an entity whose degree exceeds d, we randomly sample only d neighbors for traversal. This way, the majority of sparse entities are unaffected, while high-frequency entities are effectively suppressed, and traversal efficiency is improved. The similarity function $F_{sim}()$ can be based on semantic similarity measures such as the dot product of embeddings:

$$F_{sim}(q^{(0)}, c) = \text{dot}(\text{embed}(q^{(0)}), \text{embed}(c)).$$

We select the paragraph with the highest score, along with their corresponding entities, to include in the sampling paths.

After D steps, every traversal results in multiple paths originating from $(e_{root}, p^{(0)})$, each path representing a sequence of contextually connected entities and their associated text paragraphs across different documents. Formally, for the root entity e_{root}, we construct a set of paths $\mathcal{P} = \{P\}$, where each path P is defined as:

$$P = [(e_{root}, q^{(0)}), (e_1, c_1), \ldots, (e_n, c_n)], \quad n \leq D,$$

with $e_i \in \mathcal{E}$ and c_i being the associated paragraph of e_i.

By aggregating the information from these cross-document paths, we achieve greater diversity through a richer and more varied combination of cross-document knowledge. Additionally, the paths effectively capture and reflect the implicit contextual association between knowledge elements spanning multiple documents.

3.2.3 Secondary Sampling and Controlled Allocation

Before proceeding to the generation phase, it is crucial to consider the utilization rate and coverage of the original corpus during generation to balance the knowledge distribution, reduce redundancy, and compensate for long-tail knowledge. Therefore, we apply **secondary sampling** on $\mathcal{P}$ to selectively collect paths for generation. Specifically, we prioritize the inclusion of paths containing entities that appear less frequently in the secondary sampled path set, by accounting for the sum of utilization rate in every path. This strategy ensures a more uniform distribution of knowledge occurrences, which mitigates biases and promotes diversity within the sampled paths, thereby enhancing the overall generation quality and efficiency.

To further refine the control over synthetic data size, we iteratively allocate the secondary sampled paths into subsets according to the coverage of the original corpus, where each subset functions as an independent unit opted for the maximum corpus coverage ($> r$) and the most balanced paragraph frequency. This modular approach allows for seamless flexibility in data customization during the generation process: depending on the required volume of synthetic data, we can combine an appropriate number of subsets to support various experimental configurations. Specifically, due to the decreasing availability of sparse entities and associated texts as sampling iterations progress, the subset obtained in the first iteration should have the highest coverage r of the original corpus. As the number of iterations increases, the coverage of subsequent subsets will gradually decrease under the fixed sampling size. We use the sample size of the first subset and corpus coverage r as references and, based on the difference between the current iteration's sampling rate $\triangle r = \frac{r-r'}{r}$, re-sample and re-use texts of entities with the lowest utilization rate to complete the current path subset.

3.3 Generation Strategies

Given a path, we design prompts to guide the LLM in generating diverse and reliable synthetic data based on the text chunk of the entities along the path.

3.3.1 Generation Prompt

To produce coherent and informative content from the aggregated cross-document paths, we design two generation strategies: Chain-of-Thought (CoT) and a complementary strategy Contrastive Clarifying (CC), which are shown in Figure 6 and Figure 7.

We observed that the CoT generation method significantly improves training performance. CoT serves as a more general generation strategy, applicable to all entities with graph path connections. However, for entities with sparse graph connections—those lacking rich relationships within the graph—CoT's effectiveness can be limited, as fewer paths are available and may not provide enough context to generate comprehensive relationships with other entities.

To address this challenge, we apply CC synthetic to supplement CoT synthetic for these sparse entities. Unlike CoT, CC does not rely on graph path connections, enabling it to work effectively even with entities that have limited graph relationships. Specifically, in the secondary sampling process mentioned before, we continuously monitor the current corpus coverage rate r'. When the total number of samples exceeds a hyperparameter l and r' does not reach r, CC is triggered for the $\triangle r$ least sampled entities based on their utilization rate. CC will randomly pair these entities without replacement. If there are N least sampled entities, then the $N/2$ path will be built for CC generations. By doing so, we enrich the generation process, helping balance the model bias caused by the long-tail distribution of entities. Furthermore, CC can explicitly clarify the differences and similarities between entities in terms of their attributes and background knowledge. This can improve the model's discriminative power of sparse entities, providing deeper insights into their nuances.

CoT generation: We prompt the LLM to fully utilize the key information from each text fragment and build a step-by-step narrative where each text fragment logically leads to the next, forming a clear flow of cause and effect. The primary goal is to synthesize information from various sources into a logically connected storyline, which ensures that the generated content is coherent and that the relationships among the fragments are explicitly articulated.

Specifically, the narrative is structured into distinct phases—including initiation, development, turning points, and conclusion—with natural transitions that preserve the logical flow of causal relationships. Based on the constructed narrative, we prompt the LLM to formulate questions that require an understanding of the entire information chain to answer. The answers are provided in a chain-of-thought style, breaking down the reasoning process step by step to arrive at the final conclusion. This design can improve interpretability and provide deeper insight into the synthetic content.

Contrastive Clarifying: We prompt the LLM to generate a comparative analysis that contrasts and compares multiple text fragments. This approach is designed to prompt the LLM to explicitly analyze and highlight the implicit nuances or lack of direct connections between pieces of information, ensuring that such contrasts are clearly reflected in the synthetic data. By conducting a detailed comparative analysis, the model can effectively uncover and present discriminative information, enriching the groundedness and diversity of the synthetic content.

Specifically, the LLM is instructed to examine each entity or fragment individually, synthesize a thoughtful contrastive narrative, and summarize the comparative insights in a concluding section. When direct similarities are absent, the narrative shifts to highlighting the unique contributions or perspectives that each entity offers within its respective context. The generated output maintains an objective and analytical tone, avoiding any attempt to force connections between unrelated fragments.

4 Experiments

To comprehensively evaluate the effectiveness and applicability of the proposed Synthesize-on-Graph (SoG) framework, this section explores its performance through a series of carefully designed experiments. The experiments aim to assess SoG's contributions in four major aspects: First, to what extent does incorporating cross-document knowledge associations in SoG enhance the diversity and depth of synthetic data compared to intra-document-focused methods (**RQ1**)? Second, does SoG's synthetic data provide consistent performance gains across language models of different sizes (**RQ2**)? Third, to what extent can SoG mitigate the long-tail knowledge problem in the original corpus (**RQ3**)? In what scenarios is SoG synthesis applicable? (**RQ4**)?

4.1 Datasets

To address our research questions, we evaluate on three representative datasets: MULTIHOP-RAG, BIOASQ, and QUALITY. A detailed description of each dataset is provided in the Appendix A.1.

4.2 Baselines and Metrics

We choose Direct QA (directly answering by the base model), Rephrasing (back-translation and synonym replacement, following [17]) and the state-of-the-art methods EntiGraph [7] as baselines for evaluation. The evaluation metrics for MHRAG, BIOASQ and QUALITY are Exact Match (EM), model-based evaluation (MBE) approach using LLM-as-a-Judge[18], and Accuracy (Acc), respectively.

4.3 Experiment Details

In our generation setup, we used GPT-4o-mini as the generation model. The temperature was set at 0.7. We utilize semantic chunking[3] to split the long contexts. The semantic embedding was computed by bge-small-en-v1.5. In all experiments, we continued pretrain the LLMs with a context length of 2048 and a batch size of 64. We apply a linear learning rate warmup for 10% of the total steps, followed by a cosine decay with a peak learning rate of 5e-6. We perform full-parameter training for 2 epochs in BF16 precision, using a per-device batch size of 2 and accumulating gradients over 4 steps. In addition, within $4.5\times$ of the original corpus size, the sampling paths for CoT are of one-hop length, while beyond that, the sampling paths are of two-hop length. For QUALITY, we followed the evaluation setup in EntiGraph. For MHRAG, we evaluate the CPT models with zero-shot prompting on a sample of 1,000 QA pairs. For BIOASQ, we constructed a hard subset consisting of 1,114 questions that Qwen3-8B failed to answer correctly in a single attempt. This sampling criterion ensures that the selected questions reflect genuine challenges for strong LLMs, thereby providing a more rigorous evaluation of knowledge-intensive reasoning. For entity ambiguity issue, we rely on surface-form string matching combined with simple heuristics, including normalization of singular/plural forms and letter casing, alias matching, and Wikipedia-style redirect mappings to partially address this issue.

4.4 Main Experiment Results

To answer **RQ1** and **RQ2**, we compare the effectiveness of SoG, traditional Rephrasing augmentation and the intra-document-focused method EntiGraph in continued pre-training (CPT) with varying amounts of synthetic data on two datasets. The results are shown in Figure 3. For MHRAG and BIOASQ, model performance steadily improves as the amount of SoG synthetic data increases. In contrast, EntiGraph synthetic data provides limited gains. Especially in MHRAG, when the EntiGraph data size exceeds 1.5 times the original corpus, performance plateaus or even degrades due to its reliance on intra-document associations. This limitation prevents diverse and deeper generations, especially for complex tasks requiring cross-source knowledge integration. The sharp performance gap on MHRAG underscores the strength of SoG's cross-document knowledge integration in the

[3] https://python.langchain.com/docs/how_to/semantic-chunker/

context graph, which uncovers implicit entity relationships and enables richer reasoning. In addition, the most significant performance boost from SoG occurs when the synthetic data volume is within 0 to 1.5 times the original corpus, demonstrating that even a moderate amount of SoG data effectively enhances large model performance.

Although SoG exhibits slightly weaker performance on the QUALITY dataset, its results remain largely comparable EntiGraph. This modest decline stems primarily from SoG's design emphasis on flexibility and generalizability across tasks that rely on large, interconnected corpora. In contrast, QUALITY poses a distinct challenge: each document is an independent narrative with minimal shared knowledge or cross-document links. To better align with this task, we constrained SoG's path sampling strategy to operate strictly within individual documents. To align with this characteristic, we constrained SoG's sampling strictly within individual documents. Despite that SoG's core strength, cross-document knowledge aggregation, was not fully utilized on this dataset, it still performed comparably with the SOTA method. This underscores the better generalization capability of our SoG.

Moreover, the traditional Rephrasing augmentation method yields only marginal or negligible improvements across all datasets, further highlighting the necessity of structurally informed synthetic data construction.

Finally, another observation is that CPT solely on the original corpus yields at best limited gains and in some cases even degrades performance relative to the original model (see Raw CPT in Figure 3). We attribute this to the lack of diversity and distributional differences in the original corpus, which further emphasizes the critical role of Synthetic CPT.

4.5 Ablation Study

4.5.1 Influence Over Different Generation Strategy

Distribution of Synthetic Data of Different Generation Strategy (RQ3): The long-tail issue of entities in the original corpus may result in insufficient learning, thereby affecting the model's performance and accuracy. Additionally, the long-tail problem can cause the model to over-rely on high-frequency entities and further diminish its ability to recognize and understand rare entities. To investigate whether SoG synthetic data can alleviate the long-tail problem of entities in the original documents, we analyzed the entity distributions in the original corpus and in SoG synthetic corpora of varying sizes.

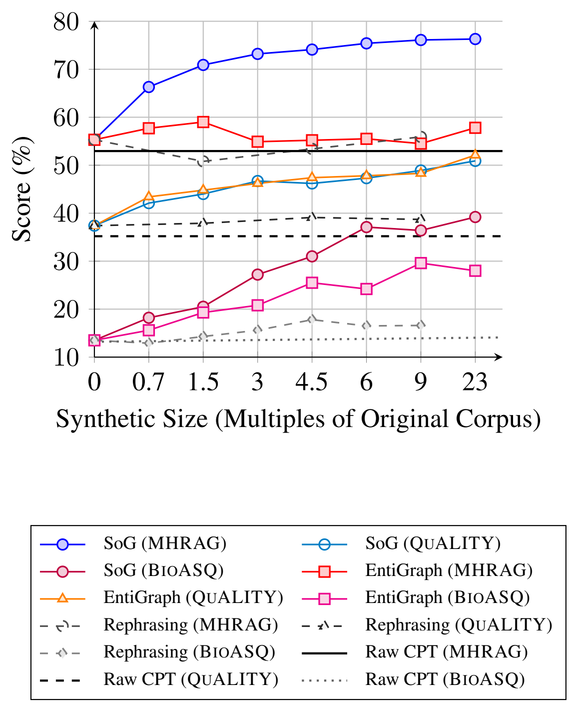
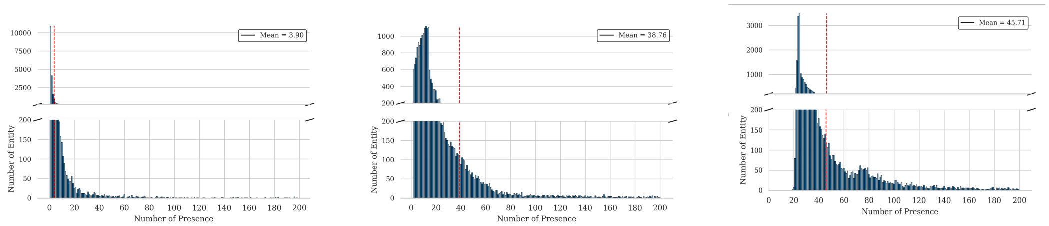


Figure 3: Performance trends of SoG and EntiGraph across three benchmarks with rephrasing and CPT baselines.

As illustrated in Figure 4b, 4a and 4c, entities in the original corpus exhibit a significant long-tail distribution. In the sampling process using only the CoT strategy (which selects paths by prioritizing entities with the lowest occurrence counts), the overall distribution becomes more concentrated. However, the long-tail trend still remains. When the Contrastive Clarifying (CC) strategy is introduced to supplement CoT (periodically enhancing long-tail knowledge based on sampling utilization rates), all long-tail entities are adequately covered, and the overall distribution begins to approximate a normal distribution. This significantly alleviates the issue of insufficient occurrences for most entities and improves diversity, demonstrating that our SoG framework can effectively balance the distribution of synthetic data.

Training Performance of Different Generation Strategy: CC is designed to specifically enhance the LLM's understanding of long-tail entities and is not suitable for standalone application to the entire



(a) Entity distribution: the original corpus (1.5M).

(b) Entity distribution: 6M SoG synthetic data with CoT generation.

(c) Entity distribution: 6M SoG synthetic data with CoT and CC generation.

Figure 4: Entity distributions for different data sets.

corpus. As a result, synthetic data solely through CC tends to be of lower quality compared to that produced by CoT. CoT primarily focuses on generating additional useful information by integrating knowledge across documents. Therefore, CoT alone can already achieve sufficient synthetic data quality. However, due to their low frequency, long-tail entities often receive less attention from CoT. As shown in the MHRAG results in Table 1, combining both generation strategies can further improves the effectiveness of synthetic data for CPT training. Interestingly, on the QUALITY dataset, using CoT alone outperforms the combined strategy. We believe this is because each QA pair in QUALITY is based on a single novel and does not involve cross-document knowledge. Such tasks tend to focus less on long-tail entities and more on the main plots and characters within the document. In this case, the CoT strategy naturally aligns with the primary content of the story. For different scenarios, our approach allows flexible adjustment of the sampling and synthesis strategies in SoG to better align with the feature of the original corpus and the specific task requirements. The specific SoG configuration adjustments for QUALITY are provided in the Appendix A.5.

Table 1: Performance of Different Approaches on Llama-3-8B-Instruct

Dataset	CoT + CC	CoT	CC	Direct QA
MHRAG(X1.5)	70.9	70.6	63.7	55.3
MHRAG(X4.5)	74.1	72.9	62.6	55.3
QUALITY(X1.5)	44.0	44.7	38.9	37.4
QUALITY(X4.5)	46.2	47.5	42.8	37.4

Table 2: CPT vs. RAG results: **Base LLM** denotes Llama-3-8B-Instruct. **CPT LLM** denotes the model CPT on the SoG data. **Zero-shot** denotes directly answering by the corresponding model. The RAG corpus consists of the raw corpus and the X3 synthetic data.

Model	RAG	Zero-shot
Base LLM	73.5	55.3
CPT LLM	70.7	73.2

4.6 CPT vs. RAG

In this experiment, we aim to answer whether non-parametric external knowledge in retrieval-augmented generation (RAG) can be replaced by parametric knowledge acquired through SoG-based CPT. Specifically, we adopt Llama-3-8B-Instruct as the base model and evaluate its performance on the MHRAG task under three configurations: LLM with SoG CPT, LLM with RAG, and LLM with both SoG CPT and RAG. From the results in Table 2, both RAG and CPT individually bring significant and similar performance gains to the LLM. Interestingly, applying RAG on top of the LLM already enhanced by synthetic CPT does not lead to further improvements. In fact, this combined setting performs worse than using either method alone. We argue that although RAG still holds a marginal advantage in performance, this advantage is outweighed by the broader benefits of synthetic CPT—including eliminating the need for retrieval, enabling shorter input windows for higher efficiency, and saving considerable computational costs in long term (**RQ4**). **Our findings highlight that incorporating SoG synthetic data into CPT enables parametric knowledge to streamline task adaptation and enhance output controllability, offering a more efficient alternative to reliance on inference-time retrieval.**

5 Conclusion

We propose Synthesize-on-Graph (SoG) framework, a context-graph-enhanced synthetic data generation method that effectively incorporates cross-document knowledge associations, which combine balanced sampling with Chain-of-Thought and Contrastive Clarifying generation strategies. Experimental results show that SoG achieves SOTA performance on multi-hop QA tasks while showing better generalization capability. Our work highlights the potential of SoG as a scalable and efficient solution for continued pretraining, offering new directions for optimizing large language model training in knowledge-intensive domains.

References

[1] 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. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL `https://arxiv.org/abs/2005.14165`. 1

[2] Claude E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 30(1):50–64, 1951. 1

[3] Common Crawl Foundation. Common crawl dataset. 2007. URL `http://commoncrawl.org/`. 1

[4] Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. Efficient large language models: A survey, 2024. URL `https://arxiv.org/abs/2312.03863`. 1

[5] Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. Knowledge mechanisms in large language models: A survey and perspective, 2024. URL `https://arxiv.org/abs/2407.15017`. 2

[6] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1067–1077. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/allen-zhu24a.html`. 2

[7] Zitong Yang et al. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024. URL `https://arxiv.org/abs/2409.07431`. 2, 7

[8] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023. 3

[9] Shengjie Ma, Chong Chen, Qi Chu, and Jiaxin Mao. Leveraging large language models for relevance judgments in legal case retrieval, 2024. URL `https://arxiv.org/abs/2403.18405`. 3

[10] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. 3

[11] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023. 3

[12] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023. 3

[13] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*, 2024. 3

[14] Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, et al. Injecting new knowledge into large language models via supervised fine-tuning. *arXiv preprint arXiv:2404.00213*, 2024. 3

[15] Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data, 2025. URL https://arxiv.org/abs/2502.12583. 3

[16] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*, 2023. 4

[17] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019. URL https://arxiv.org/abs/1901.11196. 7

[18] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594. 7

[19] Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, 2024. URL https://arxiv.org/abs/2401.15391. 11

[20] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10 (1):170, 2023. 11

[21] Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States, July 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.naacl-main.391. 12

[22] Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation, 2025. URL https://arxiv.org/abs/2407.10805. 13

[23] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pre-training of language models, 2023. URL https://arxiv.org/abs/2302.03241. 14

A Appendix

A.1 Datasets

- MultiHop-RAG (MHRAG) [19] is specifically designed to challenge the multi-hop reasoning capabilities of LLMs. It consists of queries constructed from news articles published between September and December 2023, which include information beyond the training cutoff of existing LLMs, ensuring that synthetic data is required to fill knowledge gaps. In addition, each query requires models to integrate evidence from multiple documents, mimicking real-world scenarios where knowledge is dispersed across sources. Existing LLMs, even RAG systems, often struggle with such tasks, underperforming in tasks that demand integrating and reasoning over scattered evidence. This dataset serves as an ideal benchmark to evaluate how SoG-generated synthetic data equips LLMs to utilize their internal knowledge for handling complex multi-hop reasoning effectively.
- BioASQ [20]: The BioASQ question answering (QA) benchmark dataset contains questions in English, along with golden standard (reference) answers and related material. The dataset has been designed to reflect real information needs of biomedical experts, assess the comprehensive understanding of professional knowledge, and is therefore more realistic and challenging than most existing datasets. We aim to explore challenging problems in professional domains that require highly specialized expertise, and investigate to what extent SoG can provide models with better learning corpora.

- QUALITY [21] is a multiple-choice question-answering dataset for long document comprehension. Unlike in prior work with passages, the questions are written and validated by contributors who have read the entire passage, rather than relying on summaries or excerpts. For a fair comparison with the state-of-the-art CPT synthetic data method, EntiGraph, we also chose this dataset for evaluation.

A.2 Long-tail Balance Analysis

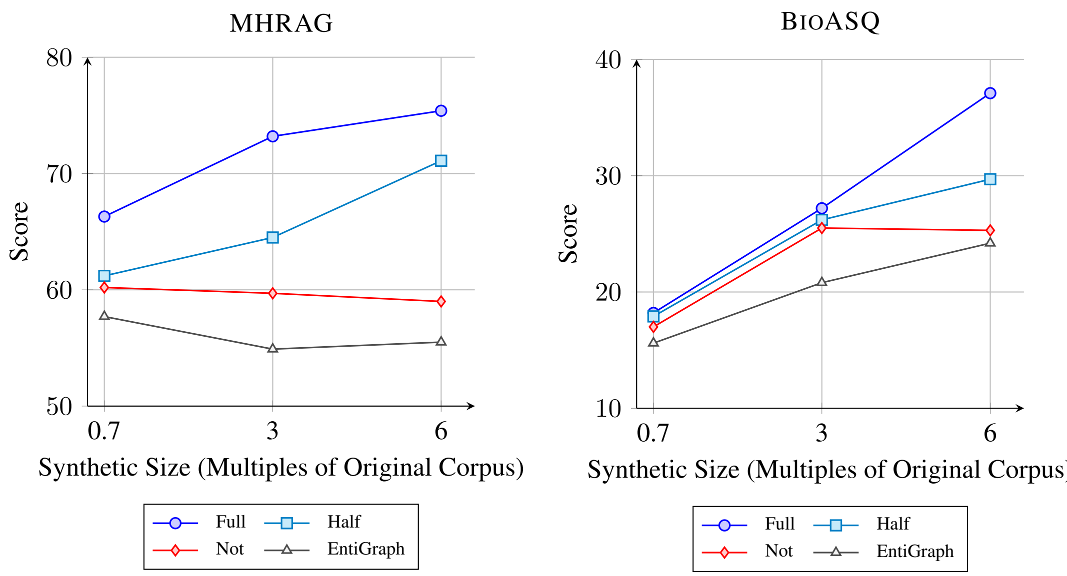


Figure 5: Performance of different sampling strategies (Full, Half, Not).

To evaluate the impact of secondary sampling in mitigating long-tail bias, we compare four strategies under different corpus scaling factors ($0.7\times$, $3\times$, $6\times$). **Not** denotes random path selection from $\mathcal{P}$ without balancing; **Half** mixes random sampling with secondary sampling at a 1:1 ratio; **Full** applies secondary sampling for all synthetic data, enforcing explicit long-tail balancing;

According to the results in Figure 5, without long-tail balancing (**Not**), the benefit of synthetic data to downstream models tends to degrade as the data scale increases, although it still outperforms EntiGraph. We argue that: a) Compared with EntiGraph, this again demonstrates that cross-document information aggregation is more valuable than intra-document synthesis alone. If the long-tail distribution is not balanced, the bias from the long-tail will gradually intensify as sampling grows, making the quality of synthetic data more prone to degradation.

Furthermore, under the **Half** setting, the gains from synthetic data diminish rapidly as the scale increases. We believe this indicates that retaining half random sampling continues to accumulate the inherent long-tail bias of the corpus, thereby limiting the scalability of synthetic data.

These observations show that secondary sampling with long-tail balancing is essential for scalable synthetic data generation. Without balancing, additional data may amplify corpus bias and even degrade quality, whereas **Full** secondary sampling consistently delivers stable improvements as the corpus scales.

A.3 Performance on More Backbone Models

Evaluating across more base models is crucial for assessing the robustness and generalizability of SoG. To this end, we have conducted additional experiments using **Qwen2.5-7B-Instruct** and **Qwen2.5-32B-Instruct** on the MHRAG dataset. The results, presented below, show that, with SoG CPT, smaller models tend to yield closer performance to the larger model:

Table 3: Performance of SoG on the MHRAG dataset across different backbone models.

Model	Direct QA	3×	6×
Qwen2.5-3B-Instruct	46.7	67.1 (+43.7%)	73.0 (+56.4%)
Qwen3-8B	50.7	70.5 (+39.0%)	76.4 (+50.7%)
LLaMA-3-8B-Instruct	48.7	70.9 (+45.6%)	75.4 (+54.8%)
Qwen2.5-32B-Instruct	55.6	73.4 (+32.0%)	81.3 (+46.2%)

Table 4: Performance of SoG on the BIOASQ dataset across different backbone models.

Model	Direct QA	3×	6×
Qwen2.5-3B-Instruct	11.8	21.7 (+83.9%)	29.4 (+149.2%)
Qwen3-8B	10.3	20.9 (+103.9%)	28.5 (+176.7%)
LLaMA-3-8B-Instruct	13.5	26.2 (+94.1%)	35.1 (+159.3%)
Qwen2.5-32B-Instruct	27.8	44.5 (+60.1%)	57.3 (+106.1%)

A.4 Influence of Path Length

We conduct a comparison to assess the impact of different sampling path length choices on the performance of CPT training in Table 5. The 1-hop paths can generate up to $5\times$ the data volume; therefore, only the $4.5\times$ result is reported. In general, the 1-hop setting achieves the best performance. The data synthesized from 2-hop paths also show significant performance. However, the 3-hop paths perform considerably weaker. We believe that this may be related to the inherent difficulty of the dataset's tasks. Furthermore, considering the challenges of constructing multi-hop reasoning tasks, most reasoning tasks are designed within two hops [22].

Table 5: Impact of Sampling Path Length on CPT Training Performance

Scale	1-Hop	2-Hop	1+2-Hop $(1:1)$	3-Hop
$4.5\times$	74.0	71.9	72.5	69.3
$9\times$	-	73.5	76.1	70.7

A.5 Configuration Adjustment Detail for QUALITY

Since each question in QUALITY focuses on a single article, we impose a constraint during multi-hop path sampling: All entities along the sampled path must be mapped to the same article ID to ensure that the retrieved texts come from the same article. We prioritize sampling the 1-hop paths. Additionally, during synthesis, we explicitly inform the LLM of the article title to which each input chunk belongs.

A.6 Implementation Cost

Our method does not rely on the strongest or most expensive LLMs. All generations are conducted with **GPT-4o-mini**, a fast and cost-efficient model (pricing: \$0.15 per 1M input tokens, \$0.08 per 1M cached input tokens, and \$0.60 per 1M output tokens). In the synthetic generation stage, the average input and output token counts per instance are approximately $1{,}700$ and 900, respectively. Based on our experiments, expanding the corpus by $3\times$–$4.5\times$ (i.e., $\approx$ 2–3M tokens for Enti-Graph) is already sufficient to yield substantial performance improvements. Consequently, the overall cost of SoG remains modest, making it a practical and accessible choice even under limited computational or financial resources.

A.7 Limitations

While our method shows promising results, several limitations remain. First, although we conducted experimental analysis on the setting of sampling path length in MHRAG, this setting is task-dependent, and determining an appropriate setting for different datasets may require empirical tuning.

Second, continued pretraining may introduce unstable LLM output, which requires additional training techniques [23]. We leave these for future work.

A.8 Ablation on Balanced Sampling without Synthesis

To address the concern that data synthesis could be avoided by directly sampling raw corpora and performing continued pre-training (CPT) to save LLM inference cost, we conducted an ablation on two benchmarks: MHRAG and BioASQ. Specifically, we compared the following settings: (i) *Zero-shot*: direct answering without CPT; (ii) *Raw CPT*: CPT on the unprocessed raw corpus; (iii) *Balanced Sample Only*: long-tail balancing only, where sampled raw chunks are concatenated and used directly for CPT without synthesis; and (iv) *SoG*: our proposed synthesis with balanced sampling.

Dataset	Zero-shot	Raw CPT	Balanced Sample Only	SoG
MHRAG	55.3	52.9	54.5	**67.9**
BioASQ	13.5	13.2	14.3	**20.3**

Table 6: Ablation on balancing without synthesis.

While our previous long-tail analysis showed that balancing helps, under the "balancing-only, no synthesis" setting we observe only **marginal gains on BioASQ** (13.5 $\rightarrow$ 14.3) and even a **drop on MHRAG** (55.3 $\rightarrow$ 54.5). This indicates that balancing alone, without synthesis, is insufficient.

This phenomenon can be explained by two factors: (1) **Distributional mismatch:** the small domain-specific raw corpus departs significantly from the original pre-training distribution. Directly continuing pre-training on such a narrow corpus reduces generalization capability. (2) **Lack of expression diversity:** most facts in the raw corpus appear only a few times with narrow phrasing. Under the next-token prediction objective, the model suffers from the *reversal curse* (seeing "A is B" does not imply learning "B is A"), making knowledge injection highly inefficient.

Why SoG Works: Unlike raw balancing, SoG does not fabricate new facts but leverages a context graph to rearrange the corpus into a **balanced and learnable** form. Through integration, the same facts are presented in more diverse, compositional expressions. This provides richer supervision under the next-token objective, leading to significantly better knowledge absorption. Overall, the results support our recipe: **balanced sampling + necessary synthesis (SoG)** is indispensable for effective knowledge injection.

A.9 Balanced Secondary Sampling

Algorithm 1 SECONDARYSAMPLING

Input: `PathSet`, target coverage rate r, standard length l, and entity to chunk index `EntityToChunk`.
`RemainingPaths` $\leftarrow$ `PathSet`
`SampledPathsCollections` $\leftarrow \emptyset$
INITIALIZE (`EntityUtilizationDict`) with default value 0
while $\mathcal{R} \neq \emptyset$ **do**
  $\mathcal{P}^*, \mathcal{R},$ `EntityUtilizationDict` $\leftarrow$
    `BalancedSampling`($\mathcal{R}, r,$ `EntityUtilizationDict`$, l,$ `EntityToChunk`)
  ADD $\mathcal{P}^*$ to `SampledPathsCollections`
end while
SAVE(`SampledPathsCollections`) for synthetic generation

Algorithm 2 BALANCEDSAMPLING

```
Input: remaining paths set R = {P}, target coverage rate r, EntityUtilizationDict,
    standard length l, and entity to chunk index EntityToChunk.
Output: sampled paths set P*, R, EntityUtilizationDict
P* ← {"cot": ∅, "cc": ∅}
r' ← 0
while R ≠ ∅ do
    R ← SORT(R, descending, by
            PATHUTILIZATIONCOUNT(P) = ∑_{node∈P} EntityUtilizationDict[node])
    # In default, l = TotalNumberOfChunksInCorpus/(hop + 1) and r = 100%
    P' ← POP(R)
    ADD P' to P*["cot"]
    # Remove the path with the least node Utilization count.
    UPDATE EntityUtilizationDict and r' based on P'
    if r' ≥ r then
        BREAK
    end if
    if LEN(P*["cot"]) ≥ l then
        △r ← (r − r')/r
        SORT EntityUtilizationDict in ascending order
        k ← ⌊△r × l⌋
        cut ← ⌊(1 − △r) × l⌋
        ADD(P*["cot"][cut :]) back to R
        REVERSE EntityUtilizationDict based on P*["cot"][cut :]
        P*["cot"] ← P*["cot"][0 : cut]
        SparseEntities ← EntityUtilizationDict[0 : k]
        for each pair (e_x, e_y) ∈ SAMPLEPAIRS(SparseEntities) do
            c_x ← SAMPLECHUNKS(EntityToChunk[e_x])
            c_y ← SAMPLECHUNKS(EntityToChunk[e_y])
            # SAMPLEPAIRS: Random combinations without replacement.
            # SAMPLECHUNKS: Random sample one chunk.
            ADD [(e_x, c_x), (e_y, c_y)] to P*["cc"]
        end for
        UPDATE EntityUtilizationDict based on P*["cc"]
        BREAK
    end if
end while
return P*, R, EntityUtilizationDict
```

A.10 Prompt

You are tasked with constructing a coherent narrative that builds a causal relationship among several text fragments.
Your role involves generating an information chain that fulfills the following criteria:

1. **Causal Narrative Development**: Use the information from each text fragment to build a step-by-step narrative that establishes a causal relationship. Develop a storyline where each fragment logically leads to the next, creating a clear flow of cause and effect.
2. **Use Provided Information Fully**: Ensure that the generated narrative makes full use of the key information from each text fragment. The causal relationships should be based directly on the details provided in the fragments.
3. **Logical Structure with Transitions**: Structure the narrative to include distinct phases such as initiation, development, turning point, and conclusion. Ensure that transitions between phases are natural and maintain the causal flow.
4. **Chain-of-Thought Question and Answer**: Based on the causal narrative, formulate a question that requires an understanding of the entire information chain to answer. Provide a detailed answer in a Chain-of-Thought style, breaking down the reasoning process step-by-step to arrive at the final answer.

Chain-of-Thought Question and Answer Example Format:

- **Question**: [Design a question based on the causal relationships in the narrative]
- **Chain-of-Thought Answer**:
  1. [First step: Identify key information relevant to the question]
  2. [Second step: Describe how this information leads to the next conclusion]
  3. [Third step: Connect this step to the following causal point]
  4. [Final step: Arrive at the answer]
  - **The answer is**: {{Answer}}

Example:

- **Generated Narrative**: Pramatha Chaudhuri influenced Bharoto Bhagyo Bidhata. Bharoto Bhagyo Bidhata wrote Jana Gana Mana.
- **Question**: Who lists Pramatha Chaudhuri as an influence and wrote Jana Gana Mana?
- **Chain-of-Thought Answer**:
  1. First, Bharoto Bhagyo Bidhata wrote Jana Gana Mana.
  2. Second, Bharoto Bhagyo Bidhata lists Pramatha Chaudhuri as an influence.
  - **The answer is**: {{Bharoto Bhagyo Bidhata}}

The output should be a natural, flowing narrative that effectively links the given fragments in a cause-and-effect chain, followed by a thoughtfully constructed question and a detailed Chain-of-Thought answer.

INPUT:{}

Figure 6: CoT Synthetic Prompt

You are tasked with generating a comparative analysis based on several text fragments. In this scenario, the text fragments may be unrelated to each other in certain aspects, and your role involves generating a thoughtful contrastive narrative that fulfills the following criteria:

1. **Entity-Focused Comparative Analysis**: The analysis should focus on comparing and contrasting the given text fragments. Do not attempt to force a connection between unrelated fragments.
2. **Maximize Use of Provided Information**: Ensure that the generated analysis makes full use of the key information provided in each text fragment. Drawing on the distinct points presented in the fragments.
3. **Highlight Differences and Similarities**: Identify and highlight the differences and any possible similarities between the key entities appear in the given text fragments. If no direct similarity exists, focus on how each entity contributes to its unique perspective or domain.
4. **Objective and Analytical Tone**: Maintain an objective and analytical tone throughout the narrative, ensuring that the analysis is insightful and grounded in the information provided.
5. **Structured and Cohesive Presentation**: Present the analysis in a structured way, such as by examining each entity in separate sections and then providing a comparative summary. This will help ensure clarity and cohesiveness in the final output.

The output should be a natural, flowing analysis that effectively contrasts the various fragments, making it easy for a reader to understand each topic entity's unique aspects and how they differ from or relate to the others.

INPUT:{}

Figure 7: CC Synthetic Prompt

A.11 Cases

INPUT ## :
Text Fragment 1: Even before the fighting started, it was voting with its feet, and some companies and investors are already refusing to do business in the country with the prospect of the reforms getting pushed through. (That effort, in turn, has been brought to a halt: People on the two sides of that debate are now standing together for what they see as the significantly bigger fight and threat.) Even before last weekend's assault, startup investment in Israel was way down from the prior year. As TechCrunch's Anna Heim reported in July, "With a provisional tally of $3.2 billion for the first half of the year, funding activity in Israel dropped by 73% compared to the same period in 2022, IVC data shows." Other countries and regions have reported similar declines in their private-market capital flows this year. As the war enters its sixth day, many people are being called up to the reserves, or volunteering to help in other ways. With the tech industry skewing younger, it's estimated that anywhere between 10% and 30% of all of Israel's tech employees are getting mobilized. More than 500 VCs have pledged their support of the efforts, too. "In the spirit of peace and unity, we encourage the global venture community to support and engage with Israeli startups, entrepreneurs, and investors as they navigate through these challenging times," the signatories wrote. To better understand the on-the-ground situation for Israeli entrepreneurs, we spoke to a number of startup founders, PR people who work with startups and investors. The predominant response was that the businesses have to continue to operate as best they can, to help ensure the economic viability of the country. Making sure everyone is OK The founders and operators that we spoke to stressed after they ensured the safety of their teams, they have gingerly started to approach how to resume work.
Text Fragment 2:It's not a simple formula. "Currently, the emphasis is on finding a balance between supporting Israeli team members, some of whom want to keep working to clear their minds a little from what's happening, and some who cannot even think of working," Omer Davidi, CEO & co-founder, BeeHero, an Israeli agritech startup, told TechCrunch. "The priority is to focus on core company operations, especially in the short term, as we await a clearer understanding of the unfolding situation," Davidi said. Yonatan Cohen, co-founder and CTO of Quantum Machines, says that the war is clearly affecting people, but they are returning to work. "I can't deny that we all had trouble concentrating on work during the first few days of the war, " he said. He points out that several of his employees have been called to active duty, but those who weren't are beginning to concentrate on work again in spite of the circumstances. Shuly Galili, founding partner at venture capital firm UpWest, says the situation is complex, but companies are doing their best to navigate the challenges of operating in war conditions. "People are being called up," she said, adding that there are some startup denizens, who are in the U.S.\n
OUTPUT ## :
Generated Narrative
Initiation
Before the outbreak of active conflict, Israel's startup ecosystem was already facing significant challenges. Tensions around political reforms led to companies and investors distancing themselves from the country, with startup investment seeing a stark decline—dropping by 73% in the first half of the year compared to the same period in 2022, according to IVC data. This downward trajectory mirrored global trends in private-market capital flows. However, the situation took a dramatic turn with the onset of war. With the fighting escalating, many Israeli tech employees—estimated between 10% and 30%—were being mobilized for active duty, either through reserve call-ups or voluntary efforts. Meanwhile, over 500 venture capitalists pledged their support, urging the global venture community to engage with Israeli startups to help sustain their operations during this crisis.
Development
As the war continued, Israeli startups faced the dual challenge of ensuring the safety of their teams while trying to maintain business continuity to support the nation's economic viability. Many founders and operators prioritized the well-being of their employees, carefully navigating the balance between personal and professional responsibilities. While some team members returned to work to regain a semblance of normalcy, others found it impossible to focus due to the gravity of the situation. For example, BeeHero CEO Omer Davidi emphasized the need to focus on core operations in the short term, while waiting for clarity on the evolving circumstances. Similarly, Yonatan Cohen, CTO of Quantum Machines, noted the difficulty of concentrating during the early days of the war but highlighted that employees who were not mobilized were gradually resuming their work. These efforts underscored the resilience of the Israeli tech community amidst the turmoil.
Turning Point
The war brought about a unique dynamic where startups had to adapt their operations to accommodate employees involved in active duty while providing support to those remaining on the ground. Venture capital firms like UpWest played a critical role in helping startups navigate these challenges. Founding partner Shuly Galili acknowledged the complexity of operating under war conditions, noting that some Israeli startup professionals had relocated to the United States, while others were finding ways to contribute domestically despite personal hardships. This collective effort—balancing operational demands with individual needs—showcased the determination of the tech community to persevere.
Conclusion
The narrative of Israel's startup ecosystem during the war reveals an intricate balance of resilience, adaptation, and unity. Despite the challenges of declining investments, mobilization of employees, and the strain of conflict, Israeli entrepreneurs, employees, and investors demonstrated a commitment to sustaining their businesses. This collective effort aimed to ensure the economic viability of the country while supporting team members in navigating the emotional and logistical complexities of the situation. The global venture capital community's solidarity further reinforced the importance of maintaining operations as a symbol of resilience and hope.
Chain-of-Thought Question and Answer
Question: How did Israeli startups adapt their operations amidst the war, and what role did the global venture capital community play in supporting them?
Chain-of-Thought Answer:
Step 1: Identify the initial challenges faced by Israeli startups: Before the war started, Israeli startups were already dealing with reduced investments due to political tensions and reform debates, with funding dropping 73% compared to the previous year. The situation worsened as the war led to mobilization of 10% to 30% of tech employees for active duty, creating operational disruptions.\n\n
Step 2: Examine how startups prioritized their response during the war: Founders and operators emphasized the safety of their teams while attempting to resume operations. Some employees returned to work to regain normalcy, while others struggled to focus due to the circumstances. Leaders like Omer Davidi and Yonatan Cohen focused on maintaining core company operations and supporting employees in balancing work and personal challenges.
Step 3: Highlight the adaptations made by startups: Startups adjusted their operations to accommodate employees called to active duty and provided support to those remaining. Venture capital firms, such as UpWest, assisted startups in navigating the complexities of operating under war conditions, with some professionals relocating to the U.S. while others contributed domestically.
Step 4: Explore the role of the global venture capital community.:More than 500 venture capitalists pledged their support, encouraging engagement with Israeli startups as a way to sustain their operations and bolster economic viability during the crisis. Their solidarity reinforced the resilience of the tech community amidst the war.
The Answer is: Israeli startups adapted by prioritizing employee safety, focusing on core operations, and supporting team members both at home and abroad. They navigated the challenges of mobilization and work disruptions, with venture capital firms providing critical support. The global venture capital community played a key role by pledging solidarity and encouraging engagement with Israeli startups, ensuring their continued operation as a symbol of resilience and economic stability."

Figure 8: CoT Case 1: 1-hop.

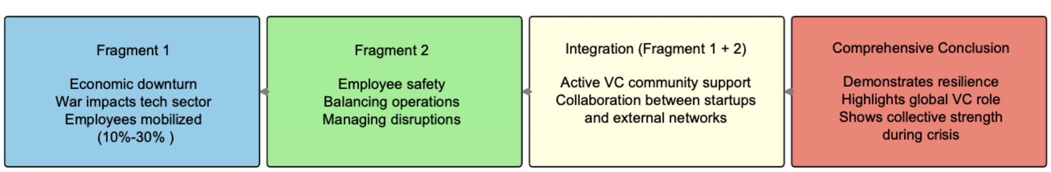


Figure 9: Brief View: CoT Case 1

INPUT ## :
Fragment 1: CMG had undergone a few leadership changes within a short period of time, and that turmoil was compounded by the economic volatility of the pandemic. As a result, the company faced eroding market share, low morale, and a stagnant artist roster. Because I had spent almost a decade here, I was uniquely situated to bring both a fresh perspective to the job and build a modern culture within the company. Major kudos go to [President of Capitol Music Group] Arjun [Pulijal], my partner in all of this, who is a brilliant creative marketer and equally adept at strategizing innovative paths for artists that range from Ice Spice and Offset to Troye Sivan and Paul McCartney. He's a great motivator and I'm proud of how he's stepped into this leadership role and excelled over the past two years. And I'm really glad you mentioned the TikTok-in-2023 wrap-up. Overall, Capitol was the top-performing label across the global and US lists, which tells yet another facet of our success story.
Text Fragment 2: But I'll spare you an exhaustive rant and focus on one issue. We need more diversity at all levels and within all facets of the music industry, particularly on the label side. "along with doing what I can to increase the number of women at our company. There just aren't enough of us, particularly within the most senior ranks at the labels." Diversity is a superpower. It's an ongoing process that I've put a lot of focus on, along with doing what I can to increase the number of women at our company. There just aren't enough of us, particularly within the most senior ranks at the labels. UMG has [Nashville Chair & CEO] Cindy Mabe and me running standalone label groups, and I know that Lucian agrees with me that there needs to be more. I will continue to do what I can to increase representation at Capitol Music Group and within the music business at large.
Text Fragment 3: What were your biggest priorities for change on day one as Chair of Capitol Music Group – and what do they tell us more generally about modernized major label groups in 2023? First, signing a truly diverse array of new artists, with specific attention on building a strong presence in hip-hop and R&B. Our President of A&R, Jeremy Vuernick, has been crucial to working with me in this regard, as well as overseeing the making of our records. The company hadn't been consistently competitive in these areas for a very long time, and, remember: Motown had been spun off on its own long before I took the reins of the company. [Motown was spun out in early 2021, then returned to CMG in early 2023]. Second, we took a hard look at every aspect of our business.
OUTPUT ## :
Generated Narrative
Initiation
The narrative begins with Capitol Music Group (CMG) experiencing significant challenges, including leadership turmoil, economic instability caused by the pandemic, eroding market share, low team morale, and a stagnant artist roster. These issues left the company struggling to remain competitive in the music industry. However, the newly appointed Chair of Capitol Music Group, who had spent nearly a decade at the company, brought a unique perspective and a plan to modernize its culture and operations. Working closely with Capitol Music Group President Arjun Pulijal, a skilled marketer and strategist, the Chair focused on revitalizing the label's artist roster and rebuilding its competitiveness. Arjun's leadership contributed to Capitol's remarkable success, as evidenced by its strong performance on TikTok's 2023 wrap-up lists.
Development
The Chair's immediate priorities included signing a diverse range of new artists, particularly focusing on hip-hop and R&B, genres where the company had lagged behind for years. This effort was aided by Jeremy Vuernick, the President of A&R, who played a critical role in overseeing the development of records and ensuring Capitol's renewed focus on these genres. The reintegration of Motown into CMG in early 2023 further exemplified the company's efforts to strengthen its artist portfolio and market presence
Turning Point
Beyond artist recruitment, the Chair recognized the need for systemic change within the music industry, particularly in promoting diversity at all levels. Acknowledging the underrepresentation of women in senior positions across record labels, the Chair committed to increasing representation within Capitol Music Group and the broader music business. This commitment aligned with efforts across Universal Music Group (UMG), where leaders such as Cindy Mabe and the Chair herself were among the few women running standalone label groups. The Chair emphasized that diversity is a \"superpower\" and an ongoing process essential for driving innovation and success in the modern music industry.
Conclusion
By addressing both internal challenges and systemic industry issues, Capitol Music Group emerged as a revitalized and forward-looking organization, committed to artist development, diversity, and long-term sustainability.
Chain-of-Thought Question and Answer
Question: How did the leadership at Capitol Music Group address its challenges, and what broader industry issues did they seek to impact?
Chain-of-Thought Answer:
Step 1. Capitol Music Group faced leadership instability, economic volatility from the pandemic, eroding market share, low morale, and a stagnant artist roster. These challenges necessitated a comprehensive overhaul of the company's strategy.
Step 2. Second, the newly appointed Chair, leveraging almost a decade of experience at CMG, collaborated with President Arjun Pulijal to modernize the company's culture and enhance competitiveness. Arjun's strategic and creative leadership contributed to Capitol's strong performance, as evidenced by its TikTok success.
Step 3. Third, the Chair prioritized signing a diverse range of new artists, with a focus on hip-hop and R&B, genres where CMG had historically been weak. This effort was supported by Jeremy Vuernick, President of A&R, and further bolstered by the reintegration of Motown into CMG in 2023.
Step 4. Fourth, the Chair recognized broader systemic issues in the music industry, particularly the lack of diversity and the underrepresentation of women in senior leadership roles. She committed to increasing representation at Capitol Music Group and across the industry, aligning with similar efforts at Universal Music Group.
Step 5. Finally, by addressing both internal challenges and systemic industry issues, Capitol Music Group emerged as a revitalized organization committed to diversity, artist development, and long-term sustainability
The Answer is: The leadership at Capitol Music Group addressed its challenges by modernizing its culture, focusing on artist development in underserved genres, and promoting diversity within the company. They also sought to impact broader industry issues by increasing representation of women and diversity in leadership roles.

Figure 10: CoT Case 2: 2-hop.

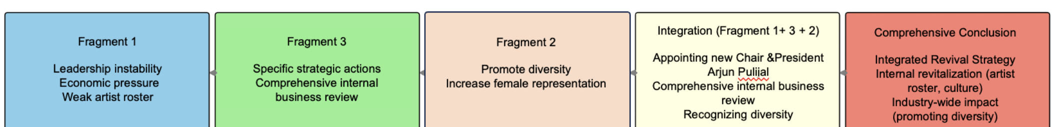


Figure 11: Brief View: CoT Case 2

INPUT ## :
Fragment 1: It was even worse for disabled attendees. Streamer Steve Saylor, a blind gamer and disability advocate, posted that he struggled to get around Las Vegas, even with rideshares, and that those with motor disabilities had even more difficulties because of the lack of adequate accommodations in the city's infrastructure. Finally back from #TwitchCon and it was such a cool vibe that I really enjoyed everything about it.","The people there were amazing!","Vegas tho, just the absolute worst. Especially for those with disabilities trying to get anywhere.
Text Fragment 2: Disney Animation's newest movie, Wish, is the culmination of the studio's celebration of its own centennial. That project stretched throughout 2023, through projects like the short Once Upon a Studio, specialized merchandise drops centered around each decade the company has been around, and the rebrand of the iconic Disney castle logo. But Wish leans so hard on celebrating the better Disney films that came before it that it doesn't have much identity of its own.
OUTPUT ## :
This analysis will explore the key themes and entities presented in two distinct text fragments. The first fragment addresses accessibility challenges in Las Vegas, the second discusses Disney's centennial celebration and its reliance on nostalgia in its film projects. Each fragment presents a unique perspective, highlighting different issues and themes.
Text Fragment 1:#### Accessibility in Las Vegas\n. The first fragment expresses a strong negative sentiment about Las Vegas, particularly concerning its accessibility for individuals with disabilities. The phrase "absolute worst" underscores the frustration and challenges faced by this demographic in navigating the city. This highlights broader societal issues regarding urban planning, inclusivity, and the need for improved infrastructure to accommodate all citizens. The focus here is on the tangible difficulties posed by the environment, emphasizing a need for systemic change to enhance accessibility.
Text Fragment 2: #### Disney's Nostalgia and Identity Crisis\nThe second fragment provides a comprehensive critique of Disney's approach to its centennial celebration through its latest movie, "Wish." The discussion centers on Disney's heavy reliance on nostalgia, which is portrayed as a double-edged sword. While nostalgia can evoke fond memories and generate revenue, the analysis suggests that it may also hinder the company2019s ability to innovate and create new, original content. The text points out that nostalgia-driven projects, while profitable, may lead to oversaturation and a lack of lasting legacy. This highlights a tension between commercial success and artistic integrity, calling into question the sustainability of such a strategy.
Conclusion
****Similarities:****
Dissatisfaction with Quality: Both fragments express a form of dissatisfaction. This indicates a broader theme of consumer frustration with quality in entertainment.
Cultural Reflection: Each fragment reflects on different aspects of culture: the text about Las Vegas addresses social inclusivity in urban environments, while the text about Disney reflect on the cultural products of the entertainment industry, particularly in relation to consumer expectations and nostalgia.
****Differences:****
Subject Matter: The first fragment focuses on accessibility issues in a physical context (Las Vegas), while the second centers around cultural products (Disney films and possibly other media). This creates a distinct divide between a societal issue and entertainment critique.
Tone and Depth: The tone of the first fragment is explicitly critical of a societal failing, while the second provides a nuanced analysis of a major corporation's strategic missteps.
Contextual Focus: The first fragment (Las Vegas) is grounded in real-world issues affecting individuals with disabilities, emphasizing practical challenges. In contrast, the second fragment Disney films and possibly other media) critiques a corporate strategy and its implications for creativity and legacy.

Figure 12: CC Case: 1-hop.