Generation of Synthetic Data for the Improvement and Evaluation of RAG Systems: Bibliographical Report

Yassine Zanned[*]

January 9, 2026

Abstract

This bibliographical report explores the state-of-the-art methods for generating synthetic data to improve and evaluate Retrieval-Augmented Generation (RAG) systems. We systematically review approaches across five key areas: graph-based generation methods, evolutionary and iterative refinement techniques, agentic solutions, evaluation metrics, and taxonomy/multimodality considerations. The report synthesizes recent advances in the field and identifies promising directions for developing novel synthetic data generation pipelines that address the challenges of limited labeled data in specialized domains.

Contents

[*]Role: Graph-Based Methods

1 Introduction

Retrieval-Augmented Generation (RAG) systems have emerged as a powerful approach to extend Large Language Model (LLM) knowledge beyond their training data by combining them with external databases. However, evaluating and fine-tuning LLMs to improve RAG systems remains a significant challenge, particularly due to the diversity of use cases and the scarcity of labeled data. Most publicly available datasets are generic and unsuitable for specific industrial applications, while creating and annotating datasets by human experts is both expensive and time-consuming.

A promising solution is to leverage LLMs themselves to generate synthetic datasets adapted to specific use cases. This report reviews the current literature on synthetic data generation methods for RAG systems, with particular emphasis on graph-based approaches, evolutionary techniques, and their integration with other methodologies.

1.1 Context and Motivation

Traditional RAG evaluation faces several key challenges:

- **Domain Specificity:** Generic public datasets fail to capture the nuances of specialized industrial applications
- **Data Scarcity:** Limited availability of labeled data in proprietary domains
- **Annotation Costs:** High expenses associated with expert annotation and dataset creation
- **Diversity Requirements:** Need for varied question types and reasoning patterns in evaluation datasets

1.2 Report Organization

- Section 2 examines graph-based synthetic data generation methods

2 Graph-Based Synthetic Data Generation Methods

2.1 Overview of Graph-Based Approaches

Graph-based methods for synthetic data generation leverage knowledge graphs and relationship structures to create contextually connected synthetic datasets. Unlike traditional generation approaches that focus on isolated documents, graph-based methods capture the intricate web of relationships between entities, concepts, and documents, enabling the generation of more diverse and knowledge-rich synthetic data for RAG system evaluation.

2.2 Evolution from Intra-Document to Cross-Document Methods

Early graph-based approaches, such as EntiGraph [2], primarily focused on intra-document content by decomposing text corpora into entity lists and generating descriptions about entity relationships within individual documents. While this approach attempted to populate the underlying knowledge graph of a corpus, it suffered from significant limitations in content diversity and knowledge depth due to its confinement to single-document boundaries.

The critical limitation of intra-document methods lies in their inability to capture cross-document knowledge associations. In reality, knowledge is inherently interconnected across documents and domains, and relying solely on entity combinations within a single document fails to capture the full spectrum of knowledge complexity. This constraint particularly affects the model's ability to handle multi-hop reasoning problems that require integrating information from multiple sources.

2.3 Synthesize-on-Graph (SoG): A State-of-the-Art Framework

2.3.1 Core Methodology

Jiang et al. [1] proposed Synthesize-on-Graph (SoG), a context-graph-enhanced synthetic data generation framework that addresses the limitations of intra-document methods by incorporating cross-document knowledge associations. SoG represents a significant advancement in graph-based synthetic data generation through its comprehensive two-component architecture:

Context Graph Construction and Cross-Document Sampling. The framework begins by constructing a context graph $G = (E, \mathcal{E})$ where nodes E represent entities and concepts extracted from the original corpus, and edges $\mathcal{E}$ represent cross-document knowledge associations. The edge set is defined as:

$$\mathcal{E} = \{(e_x, e_y) \mid \exists i, j \text{ such that } e_x, e_y \in E_{i,j}\} \tag{1}$$

where $E_{i,j}$ represents the entities extracted from paragraph j of document i.

The framework employs a sophisticated two-stage sampling strategy:

1. **Graph Traversal with Similarity-Based Selection:** Starting from a root entity e_{root}, the system performs breadth-first search (BFS) traversal up to depth D. At each step, neighboring entities are prioritized using a similarity function:

   $$F_{\text{sim}}(q^{(0)}, c) = \text{dot}(\text{embed}(q^{(0)}), \text{embed}(c)) \tag{2}$$

   where $q^{(0)}$ is the root paragraph and c is a candidate paragraph from neighboring entities.

2. **Secondary Sampling and Controlled Allocation:** To balance knowledge distribution and address long-tail entities, the framework applies secondary sampling that prioritizes paths containing less frequently appearing entities, ensuring more uniform knowledge coverage across the synthetic corpus.

Combined Chain-of-Thought and Contrastive Clarifying Synthesis. To enhance synthetic data quality, SoG integrates two complementary generation strategies:

- **Chain-of-Thought (CoT) Generation:** This strategy guides the language model to construct step-by-step narratives where text fragments from different documents are logically connected through causal relationships. The narrative is structured into distinct phases—initiation, development, turning points, and conclusion—with natural transitions that preserve logical flow. Based on the constructed narrative, the system generates questions requiring multi-hop reasoning, with answers provided in a chain-of-thought style.

- **Contrastive Clarifying (CC) Generation:** Designed specifically for sparse entities with limited graph connections, CC generates comparative analyses that contrast and compare multiple text fragments. This approach explicitly highlights discriminative information and nuances between entities, even when direct similarities are absent. CC is triggered adaptively when entity utilization rates fall below a threshold, helping to balance model bias caused by long-tail entity distribution.

2.3.2 Key Technical Contributions

The SoG framework introduces several notable technical innovations:

1. **Entity-Context Mapping:** For each entity $e_k \in E$, the system maintains a mapping $M : e_k \mapsto P_k$ that associates entities with all paragraphs in which they appear, enabling efficient cross-document relationship discovery.

2. **Multi-Hop Path Construction:** Each traversal results in paths of the form:

$$P = [(e_{\text{root}}, q^{(0)}), (e_1, c_1), \dots, (e_n, c_n)], \quad n \leq D \tag{3}$$

   where $e_i \in E$ and c_i is the associated paragraph, capturing contextually connected knowledge across multiple documents.

3. **Adaptive Strategy Selection:** The framework dynamically chooses between CoT and CC generation based on entity graph connectivity and utilization rates, optimizing for both common and rare knowledge elements.

4. **Long-Tail Entity Mitigation:** Through secondary sampling and CC generation, SoG effectively addresses the long-tail distribution problem prevalent in most corpora, where a few entities dominate while many remain underrepresented.

2.3.3 Experimental Validation

Jiang et al. conducted comprehensive experiments on two benchmark datasets:

- **MultiHop-RAG (MHRAG):** A dataset specifically designed to evaluate multi-hop reasoning capabilities, consisting of queries that require integrating evidence from multiple documents. SoG demonstrated substantial performance improvements over the state-of-the-art EntiGraph method, with performance gains increasing proportionally with synthetic data volume up to 9× the original corpus size.

- **QUALITY:** A long document comprehension dataset where each question focuses on a single narrative document. While SoG showed slightly weaker performance compared to EntiGraph on this dataset, the results remained largely comparable, demonstrating SoG's better generalization capability across different task types.

The experiments revealed several important findings:

1. Cross-document knowledge integration significantly enhances performance on complex multi-hop reasoning tasks, with the most substantial gains occurring when synthetic data volume is within 0 to 1.5× the original corpus size.

2. Entity distribution analysis showed that combining CoT and CC strategies transforms the long-tail distribution of the original corpus into a more balanced, near-normal distribution in the synthetic data.

3. SoG-based continue pretraining (CPT) achieved performance comparable to retrieval-augmented generation (RAG) on certain tasks, suggesting that parametric knowledge acquired through high-quality synthetic data can potentially reduce or eliminate the need for external retrieval systems.

2.4 Implications for RAG System Evaluation

The SoG framework has significant implications for generating evaluation datasets for RAG systems:

- **Enhanced Diversity:** Cross-document sampling ensures that synthetic evaluation data covers a broader spectrum of knowledge relationships, better reflecting real-world information-seeking scenarios.
- **Multi-Hop Reasoning:** The graph-based approach naturally generates questions requiring multi-hop reasoning, which is critical for evaluating advanced RAG systems that must integrate information from multiple sources.
- **Coverage of Rare Knowledge:** The explicit handling of long-tail entities ensures that evaluation datasets adequately test system performance on less common but potentially important knowledge elements.
- **Quality-Diversity Trade-off:** The dual generation strategy (CoT + CC) balances the need for high-quality, coherent synthetic data with the requirement for diverse coverage across the knowledge space.

2.5 Limitations and Future Directions

Despite its strengths, the SoG framework has several acknowledged limitations that represent opportunities for future research:

1. **Task-Dependent Hyperparameters:** The optimal path length settings (1-hop, 2-hop, or 3-hop) are task-dependent and require empirical tuning for different datasets and use cases. Developing adaptive methods for automatic hyperparameter selection could improve the framework's applicability.
2. **Computational Requirements:** The framework relies on large language models (GPT-4o-mini in the reported experiments) for synthetic data generation, which may be computationally expensive for large-scale applications. Exploring more efficient generation models or caching strategies could reduce costs.
3. **Output Stability:** Continue pretraining may introduce instability in LLM outputs, potentially requiring additional training techniques or regularization strategies to ensure consistent performance.
4. **Dataset-Specific Adjustments:** As demonstrated by the different configurations required for MHRAG versus QUALITY, the framework may need customization for different types of evaluation tasks, suggesting that developing more robust, task-agnostic sampling strategies is an important direction.

2.6 Summary

Graph-based methods, particularly the recent SoG framework, represent a significant advancement in synthetic data generation for RAG systems. By moving beyond intra-document relationships to capture cross-document knowledge associations, these methods enable the creation of more diverse, comprehensive, and realistic evaluation datasets. The explicit handling of graph structures, combined with sophisticated sampling strategies and adaptive generation techniques, positions graph-based approaches as a promising direction for addressing the challenges of evaluating and improving RAG systems in specialized domains with limited labeled data.

References

[1] Xuhui Jiang, Shengjie Ma, Chengjin Xu, Cehao Yang, Liyu Zhang, and Jian Guo. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pre-training of large language models. *arXiv preprint arXiv:2505.00979*, 2025.

[2] Zitong Yang et al. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024.