Diverse and Private Synthetic Datasets Generation for RAG Evaluation: A Multi-agent Framework

Ilias Driouich[1,*], Hongliu Cao[1] and Eoin Thomas[1]

[1]AMADEUS France

Abstract

Retrieval-augmented generation (RAG) systems improve large language model outputs by incorporating external knowledge, enabling more informed and context-aware responses. However, the effectiveness and trustworthiness of these systems critically depends on how they are evaluated, particularly on whether the evaluation process captures real-world constraints like protecting sensitive information. While current evaluation efforts for RAG systems have primarily focused on the development of performance metrics, far less attention has been given to the design and quality of the underlying evaluation datasets, despite their pivotal role in enabling meaningful, reliable assessments. In this work, we introduce a novel multi-agent framework for generating synthetic QA datasets for RAG evaluation that prioritize semantic diversity and privacy preservation. Our approach involves: (1) a Diversity agent leveraging clustering techniques to maximize topical coverage and semantic variability, (2) a Privacy Agent that detects and mask sensitive information across multiple domains and (3) a QA curation agent that synthesizes private and diverse QA pairs suitable as ground truth for RAG evaluation. Extensive experiments demonstrate that our evaluation sets outperform baseline methods in diversity and achieve robust privacy masking on domain-specific datasets. This work offers a practical and ethically aligned pathway toward safer, more comprehensive RAG system evaluation, laying the foundation for future enhancements aligned with evolving AI regulations and compliance standards.

Keywords

Multi-Agent system, Privacy-preserving, Evaluation system, Synthetic Dataset Generation,

1. Introduction

Retrieval-augmented generation (RAG) aims to improve large language models (LLM) output by incorporating relevant information retrieved from external knowledge sources. It has been effectively applied in various scenarios, such as domain-specific chatbots [1, 2] and email/code completion [3]. A typical RAG system often operates in two stages: retrieval and generation. First, the system retrieves the relevant knowledge from an external database based on the user query. Then, the retrieved information is integrated with the query to form an input for the LLM in charge of the generation stage. The LLM uses its pre-trained knowledge and the retrieval data to generate a response, enhancing the overall quality of the output. As RAG sees wider adoption, ensuring robust performance evaluation becomes critical. While numerous automated methods, ranging from classic n-gram metrics (BLEU, ROUGE) and embedding-based measures (BERTScore)[4] to the "LLM-as-a-judge" approach leveraging GPT have been explored, an equally vital element is having the "golden" evaluation set with sufficiently diverse and representative samples that will serve as a complete benchmark to evaluate both the retrieval and generation processes [5]. Furthermore, despite the emergence of multiple RAG benchmarks [6, 7, 8] that span general-purpose and specialized domains, many still fall short of reflecting the complexity and variability of real-world use cases. In particular, traditional benchmarks often lack coverage of novel or underrepresented topics, limiting their ability to generalize [2, 9]. This gap poses a significant challenge for reliable evaluation, especially in domains requiring deep expertise and factual precision [10].

To address these challenges, a new line of work consisting in generating synthetic evaluation sets

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

ilias.driouich@amadeus.com (I. Driouich)

[11] has emerged and is very promising. Indeed, these methods offer a practical solution for generating datasets that mimic real human interactions by leveraging advanced LLM reasoning capabilities. Such synthetic datasets can include a broad range of scenarios, from straightforward factual questions to more nuanced domain-specific ones enabling robust and comprehensive evaluation of RAG systems. Additionally, synthetic data generation methods are increasingly recognized as vital components for the safe, transparent, and compliant evaluation of AI systems. In fact, regulatory frameworks, such as the European Union's AI Act [12], explicitly promote the use of synthetic datasets within AI compliance and auditing processes. However, maintaining both a high efficiency of privacy-preserving mechanisms in retrieval systems and adherence to privacy regulations remains essential to building reliable and ethical evaluation frameworks.

In fact, according to the existing literature [13, 14], retrieval systems may face serious privacy issues when the retrieval process involves sensitive data. For example, the authors in [13] observe that carefully designed user prompts are able to extract original sentences from the retrieval data or can also extract specific pieces of private information, potentially leading to the leakage of considerable amounts of retrieval data. The potential risk of information leakage can significantly limit the applications of retrieval systems. For example, a medical chatbot [15] using patient history diagnosis cases as a source of knowledge can improve response quality but raises concerns about exposing sensitive patient information.

In this work, we take the first step toward exploring the generation of diverse and privacy-aware synthetic QA datasets designed specifically to serve as evaluation ground truth for assessing RAG systems. Our main contributions can be summarized as follows:

- We introduce a modular and extensible multi-agent pipeline for synthetic QA dataset generation, designed specifically for evaluating RAG systems while ensuring a balance between diversity, privacy, and utility.
- We develop and perform a comprehensive twofold evaluation strategy: (1) diversity assessment combining qualitative judgments from LLM-based evaluators and quantitative diversity metrics, and (2) privacy assessment focused on the accuracy and effectiveness of entity masking across three specialized datasets.

2. Related works

2.1. Retrieval-augmented generation and privacy issues

Retrieval-Augmented Generation , introduced by [16], has gained substantial traction for enhancing LLM responses through external context. By retrieving relevant documents or passages and incorporating them into the prompt, RAG often yields output with improved accuracy and factual grounding [17], mitigating the well-documented "hallucination" problem in LLMs [18]. In addition to higher-quality outputs, RAG offers architectural flexibility by allowing independent upgrades to any of its components (e.g., data storage, retriever, or the core LLM) without requiring full model retraining [19, 20]. These advantages have led to the adoption of RAG in diverse settings, from personal chatbots to highly specialized expert systems [21].

Despite its clear benefits, the retrieval process introduces privacy risks, particularly in domains handling sensitive user data. For instance, [22] highlight privacy implications of retrieval-based language models, showing how training data and user inputs can be unintentionally exposed through retrieved passages. Other works have demonstrated that RAG models are susceptible to extraction attacks [13], including membership inference and reconstruction attacks, which exploit learned representations to infer whether a user's data was part of the training set [23]. Such vulnerabilities pose significant challenges for deploying RAG-based solutions in sensitive applications (e.g., healthcare, finance) where data privacy is paramount.

2.2. Synthetic data generation using Large Language Models

Recent advances in LLMs have created significant interest in using them to automatically generate synthetic data. For instance, [24, 25] utilize zero-shot prompting to produce synthetic samples for tasks such as text classification and question answering, subsequently training smaller models on this generated data. [26] introduce a noise-robust re-weighting framework to further refine data quality, while [27] propose mixing a set of soft prompts and applying prompt tuning to enhance diversity in the generated text. Beyond prompt-based methods, [28] examine specific attributes of the data itself, such as length and style, to diversify synthetic outputs. In parallel, a growing line of work has begun to address privacy concerns in synthetic data generation.

The authors in [29] propose a few-shot approach, generating private in-context demonstrations backed by differential privacy guarantees, and [30] design a private evolution algorithm that enforces differential privacy throughout the generation process. In [31], the authors propose a novel two-stage synthetic pipeline that includes attribute-based data generation, which aims to maintain key information, and iterative agent-based refinement, which further enhances the privacy of the input data in RAG systems.

2.3. Synthetic question answer generation (QAG) and RAG evaluation

A new wave of QAG and RAG evaluation approaches is redefining the field through synthetic data generation and dynamic assessment methods powered by LLMs. Rather than relying on static, human-curated datasets, recent efforts leverage LLMs to generate QA pairs and evaluate model outputs using automated scoring mechanisms, such as LLM-as-a-judge frameworks. Systems like RAGAS [11] support scalable, domain-adaptable evaluation by conditioning synthetic QA generation on retrieved context and employing flexible, model-driven evaluation criteria. These methods offer the advantage of tailoring evaluation to specific domains and evolving data distributions. However, they also introduce new challenges, including maintaining content diversity, ensuring output consistency, and protecting sensitive information, particularly when operating over proprietary or privacy-sensitive corpora. Our work builds upon this emerging paradigm by incorporating explicit mechanisms to address these concerns, contributing a privacy and and diversity-aware framework for RAG evaluation.

3. The proposed solution

Algorithm 1 formally outlines the multi-agent procedure, detailing the sequential interaction between the diversity agent, privacy Agent, and QA curation agent. Given an input dataset D and clustering hyperparameters, the process outputs a set of synthetic QA samples D_{QA}, enriched with semantic diversity and reinforced by privacy protections.

In the first step, the Diversity agent clusters the original dataset using semantic embeddings and selects representative samples from each cluster, ensuring a broad coverage of topics. The second step, the privacy agent, operates over each cluster's representative samples, detecting and pseudonymizing sensitive entities to produce a private version of the data along with a structured privacy report. Finally, the QA curation agent synthesizes question-answer pairs from the private data, generating evaluation-ready samples along with a QA generation report that summarizes success rates and generation dynamics.

4. Experiments

4.1. Experimental Setup

Implementation details All components of our multi-agent framework were implemented in Python, using the LangGraph framework to orchestrate inter-agent communication and control flow. Each agent was instantiated as a node within a LangGraph workflow.

Algorithm 1 Multi-Agent Synthetic Evaluation Dataset Generation for RAG

Input: D & Original document
Output: D_{QA}: A diverse, privacy-compliant synthetic QA dataset
Initialization: $D_{\text{div}} \leftarrow \{\},\ D_{\text{priv}} \leftarrow \{\},\ D_{QA} \leftarrow \{\},\ Report_{\text{priv}}, Report_{QA} \leftarrow \emptyset$
Stage 1: Diversity Agent

1. **Clustering:** Apply k-means clustering algorithm to group D into k clusters $\{C_1, C_2, ..., C_k\}$ based on text embeddings.
2. **Representative Sampling:** For each cluster C_i, select a subset of representative samples S_i.
3. **Aggregate Diverse Samples:** $D_{\text{div}} \leftarrow \bigcup_{i=1}^{k} S_i$

Stage 2: Privacy Agent

1. **For each** S_i **in** D_{div}:
   - Detect PII in each sample $x \in S_i$
   - Pseudonymize each identified entity to produce x'
   - Accumulate private samples: $D_{\text{priv}_i} \leftarrow D_{\text{priv}_i} \cup \{x'\}$
2. **Aggregate Private Documents:** $D_{\text{priv}} \leftarrow \bigcup_{i=1}^{k} D_{\text{priv}_i}$
3. **Privacy Report:** Record types and frequencies of pseudonymized entities in $Report_{\text{priv}}$

Stage 3: QA Curation Agent

1. **For each** $x' \in D_{\text{priv}}$:
   - Generate n QAs pair (q, a)
   - Accumulate: $D_{QA} \leftarrow D_{QA} \cup \{(q, a)\}$
2. **Generate QA Report:** Log model settings, number of successful QA pairs, failures and generation procedure in $Report_{QA}$

return $D_{QA},\ Report_{\text{priv}},\ Report_{QA}$

Language models We employed models from Azure OpenAI services. Specifically, we used `GPT-4o` for the Diversity agent and the QA curation agent, due to its fast response time and strong generalization capabilities for content generation. For the privacy agent, we used `GPT-4.1`, which offers superior reasoning and tool-usage capabilities that are crucial for accurate PII detection and transformation tasks involving interaction with APIs or complex instructions. To ensure reproducibility and minimize variability in outputs, the temperature of all language models was fixed at 0 during inference. For the clustering process in the diversity agent, we generated embeddings using OpenAI's `text-embedding-3-small` (Ada 3) model with an embedding dimension of 1536. Input documents were preprocessed into chunks of 256 tokens before applying k-means clustering.

Agents tool configuration Each agent in the system operates with tailored tools suited to its objectives. First, the diversity agent uses a k-means clustering function to identify latent topic clusters within the input document. The optimal value of k is selected using intra-cluster distance scores.

Second, the privacy agent performs pseudonymization based on a predefined set of PII categories. It scans the generated content, identifies sensitive entities, and replaces them using context-aware transformations. In addition, it produces a structured privacy report detailing which PIIs were correctly detected, masked, or missed.

Last, the QA curator agent generates final QA pairs from the enriched, privacy-preserved inputs

by leveraging advanced prompting techniques. It also produces a comprehensive generation report summarizing the types of QA pairs created, their alignment with source content, and overall dataset characteristics.

4.2. Research question 1: How does the proposed multi-agent system enhance the diversity of the generated evaluation dataset?

4.2.1. Baselines

To assess the effectiveness of our proposed multi-agent approach in enhancing dataset diversity, we compare against two baselines:

(1) Evolutionary generation (RagasGen). Inspired by works such as Evol-Instruct and RAGAS [11], this baseline uses an evolutionary generation paradigm to produce QA pairs. It iteratively mutates and refines questions to maximize diversity along dimensions such as reasoning complexity, multi-hop dependencies, and topic breadth.

(2) Direct Prompting (DirPmpt). This baseline uses direct LLM prompting with few shot examples. A GPT-4o model is prompted with handcrafted instructions to produce diverse QA pairs.

4.2.2. Diversity evaluation dataset

To evaluate our multi-agent framework's ability to generate diverse QA pairs, we use the official EU AI Act as input. Its rich structure and varied content provide a realistic and challenging testbed for assessing diversity in synthetic evaluation sets.

4.2.3. Diversity evaluation methodology

To assess the diversity of the generated QA sets, we use the LLM-as-a-Judge approach, where GPT-4.1 is prompted to act as an expert evaluator. The model receives pairs of evaluation sets, our generated set and baseline sets, along with instructions to judge question diversity based on semantic variety, topical coverage, and phrasing differences. It then assigns diversity scores on a scale from 1 to 10. Additionally, we use the **CosineSimilaritytoDiversity**[32], which inverts the average pairwise cosine similarity of sentence embeddings, lower values indicate greater semantic spread.

4.2.4. Findings discussion

First, we observe that our multi-agent system outperforms RagasGen and DirPrmpt in all evaluated settings, with consistent gains observed across both qualitative and quantitative metrics.

Furthermore, we observe a consistent trend across all diversity measures: as the test set size increases, so does the diversity of the generated questions. The LLM-as-a-Judge scores (GPT-4.1) rise from 7.8 at 10 samples to 9 at 100 samples, indicating that the generated question sets increasingly exhibit richer topic coverage and variation in structure. Quantitatively, the `CosineSim2toDiversity` score becomes less negative (closer to zero), reflecting that questions are increasingly dissimilar to each other, a direct proxy for higher diversity. These results demonstrate that our multi-agent system enhances question diversity, particularly at larger scales.

QA set size	GPT-4.1 Diversity Rating			Cosine Sim. to Diversity		
	Ours	RagasGen	DirPmpt	Ours	RagasGen	DirPmpt
10	**7.8**	7.0	6.2	**-0.36**	-0.40	-0.45
25	**8.2**	7.3	6.3	**-0.31**	-0.38	-0.43
50	**8.6**	7.4	6.9	**-0.26**	-0.36	-0.38
75	**8.9**	8.0	7.5	**-0.18**	-0.34	-0.35
100	**9.0**	8.1	7.6	**-0.15**	-0.33	-0.33

Table 1
Diversity and similarity metrics comparison between question sets generated by our method, RagasGen, and DirPmpt.

4.3. Research Question 2: How does the proposed multi-agent solution preserve the overall privacy of the system?

4.3.1. Privacy evaluation datasets

To evaluate the effectiveness of the privacy agent, we use three benchmark datasets provided by AI4Privacy[1]: **PII-Masking-200K**, **PWI-Masking-200K**, and **PHI-Masking-200K**. These tabular datasets contain long-form sentences annotated with private entities from different domains. The PWI dataset includes job titles, company names, and salary information. The PHI dataset focuses on medical diagnoses, genetic information, and gender. The PII dataset contains names, locations, dates of birth, and contact details. Each dataset also provides additional metadata such as entity type, position, and frequency.

To simulate realistic input conditions for our pipeline, we concatenated individual samples from each dataset into longer text paragraphs. We refer to each resulting dataset as PWI, PHI, or PII. Each consists of domain-specific long sentences containing private entities and their corresponding masked versions. Table 2 summarizes the main statistics for each dataset, including document length and the number of private entities.

Dataset	Dataset length (sentences)	Total entities number	Avg entities per sentence
PWI	1800	451	3.99
PHI	1700	422	4.02
PII	1600	591	2.71

Table 2
Statistics of the constructed privacy evaluation datasets.

4.3.2. Experimental results

In Figure 1 we present label-wise accuracy across the PHI, PWI, and PII datasets The privacy agent shows strong overall performance, with most labels achieving accuracies between 0.75 and 0.90. On the PHI dataset, the highest scores are observed for DISABILITYSTATUS (0.91), HOSPITALNAME (0.90), and MENTALHEALTHINFO (0.90), indicating robust handling of sensitive medical information. On the PWI dataset, the model performs best on JOBTYPE (0.94), TELEPHONENUM (0.90), and DATE, GENDER, SALARY, ORGANISATION, DBAREA (all 0.88), demonstrating high reliability in identifying entities related to the workplace. Moreover, results on the PII dataset highlight strong performance for LASTNAME (0.91), CARDNUMBER and CITY (0.87), and FIRSTNAME, STATE, and JOBAREA (all 0.86), confirming the agent's effectiveness in detecting general personal identifiers.

Interestingly, overlapping labels such as GENDER appear across PHI (0.83), PWI (0.88), and PII (0.83), and yield consistently strong scores. This suggests that the privacy agent generalizes well across

[1] https://huggingface.co/ai4privacy

domains.

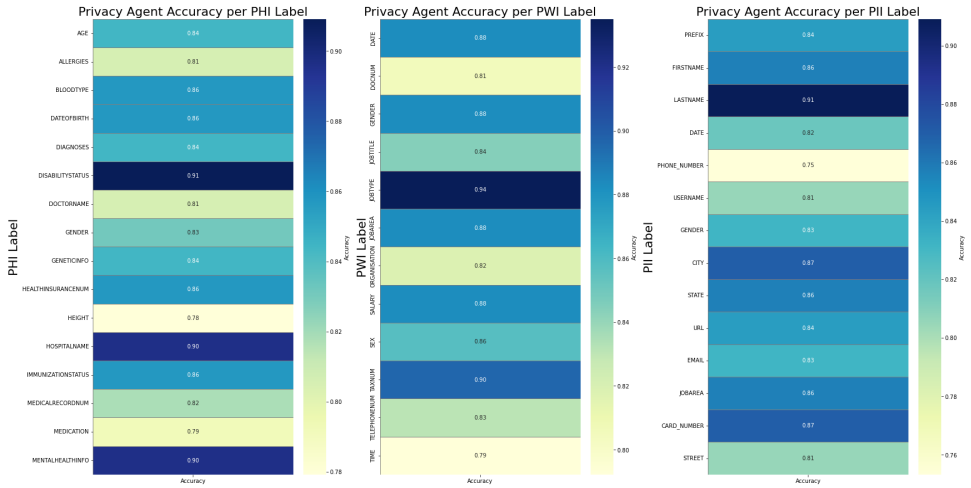

Figure 1: Privacy agent accuracy per entity type across the PHI, PWI, and PII datasets.

5. Conclusion and future work

In this work, we introduced a modular multi-agent framework for the generation of synthetic QA datasets tailored to the rigorous evaluation of RAG systems. Our approach decomposes the dataset construction process into distinct, specialized agents, each focused on enriching semantic diversity, enforcing privacy safeguards, and curating high-quality QA pairs. Through comprehensive experiments we highlight the framework's effectiveness in producing evaluation datasets that are both representative and privacy-preserving, addressing critical challenges in real-world RAG evaluation.

Looking ahead, we aim to enhance the autonomy and collaboration of individual agents by leveraging tool-augmented foundation models. For example, the diversity agent could dynamically infer optimal clustering structures, while the privacy agent could adaptively identify and transform PIIs beyond static entity lists. In addition, we plan to explore agent-to-agent communication protocols and effective independent agent collaboration, potentially through frameworks like model context protocol, to improve coordination and task delegation among agents.

Future work will also include rigorous evaluation of the framework's resilience to privacy attacks, helping to clarify its defensive boundaries and inform improvements. As AI regulations such as the EU AI Act continue to evolve, subsequent versions of our framework will further align synthetic evaluation set generation not only with principles of technical trustworthiness, but also with emerging regulatory requirements.

6. Declaration on Generative AI

The authors have not employed any generative AI tools.

References

[1] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, S. Nanayakkara, Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering, Transactions of the Association for Computational Linguistics 11 (2023) 1–17.

[2] H. Cao, Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark, arXiv preprint arXiv:2406.01607 (2024).

[3] M. R. Parvez, W. Ahmad, S. Chakraborty, B. Ray, K.-W. Chang, Retrieval augmented code generation and summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2719–2734.

[4] A. Chen, G. Stanovsky, S. Singh, M. Gardner, Evaluating question answering evaluation, in: A. Fisch, A. Talmor, R. Jia, M. Seo, E. Choi, D. Chen (Eds.), Proceedings of the 2nd Workshop on Machine Reading for Question Answering, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 119–124. URL: https://aclanthology.org/D19-5817/. doi:10.18653/v1/D19-5817.

[5] H. Cao, I. Driouich, R. Singh, E. Thomas, Multi-agent llm judge: automatic personalized llm judge design for evaluating natural language generation applications, arXiv preprint arXiv:2504.02867 (2025).

[6] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1601–1611.

[7] J. Chen, H. Lin, X. Han, L. Sun, Benchmarking large language models in retrieval-augmented generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 17754–17762.

[8] Y. Lyu, Z. Li, S. Niu, F. Xiong, B. Tang, W. Wang, H. Wu, H. Liu, T. Xu, E. Chen, Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models, arXiv preprint arXiv:2401.17043 (2024).

[9] H. Cao, Enhancing negation awareness in universal text embeddings: A data-efficient and computational-efficient approach, Proceedings of the 28th European Conference on Artificial Intelligence (ECAI-2025) (2025).

[10] T. Bruckhaus, Rag does not work for enterprises, 2024. URL: https://arxiv.org/abs/2406.04369. arXiv:2406.04369.

[11] S. Es, J. James, L. Espinosa Anke, S. Schockaert, RAGAs: Automated evaluation of retrieval augmented generation, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 150–158. URL: https://aclanthology.org/2024.eacl-demo.16.

[12] E. Commission, Eu artificial intelligence act (ai act), https://artificialintelligenceact.eu, 2024. Accessed: 2025-07-15.

[13] S. Zeng, J. Zhang, P. He, Y. Xing, Y. Liu, H. Xu, J. Ren, S. Wang, D. Yin, Y. Chang, et al., The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag), ACL Findings (2024).

[14] Y. Ding, W. Fan, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, Q. Li, A survey on rag meets llms: Towards retrieval-augmented large language models, arXiv preprint arXiv:2405.06211 (2024).

[15] L. Yunxiang, L. Zihan, Z. Kai, D. Ruilong, Z. You, Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge, arXiv preprint arXiv:2303.14070 (2023).

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[17] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997 (2023).

[18] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, arXiv preprint arXiv:2104.07567 (2021).

[19] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, W. Chen, Enhancing retrieval-augmented large language xmodels with iterative retrieval-generation synergy, arXiv preprint arXiv:2305.15294 (2023).

[20] X. Cheng, D. Luo, X. Chen, L. Liu, D. Zhao, R. Yan, Lift yourself up: Retrieval-augmented text generation with self memory, arXiv preprint arXiv:2305.02437 (2023).

[21] D. P. Panagoulias, M. Virvou, G. A. Tsihrintzis, Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis, Electronics 13 (2024) 320.

[22] Y. Huang, S. Gupta, Z. Zhong, K. Li, D. Chen, Privacy implications of retrieval-based language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023.

[23] Z. Qi, H. Zhang, E. Xing, S. Kakade, H. Lakkaraju, Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems, arXiv preprint arXiv:2402.17840 (2024).

[24] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, L. Kong, Zerogen: Efficient zero-shot learning via dataset generation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 11653–11669.

[25] Y. Meng, J. Huang, Y. Zhang, J. Han, Generating training data with language models: Towards zero-shot language understanding, Advances in Neural Information Processing Systems 35 (2022) 462–477.

[26] J. Gao, R. Pi, L. Yong, H. Xu, J. Ye, Z. Wu, W. Zhang, X. Liang, Z. Li, L. Kong, Self-guided noise-free data generation for efficient zero-shot learning, in: International Conference on Learning Representations (ICLR 2023), 2023.

[27] D. Chen, C. Lee, Y. Lu, D. Rosati, Z. Yu, Mixture of soft prompts for controllable data generation, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 14815–14833.

[28] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. J. Ratner, R. Krishna, J. Shen, C. Zhang, Large language model as attributed training data generator: A tale of diversity and bias, Advances in Neural Information Processing Systems 36 (2024).

[29] X. Tang, R. Shin, H. A. Inan, A. Manoel, F. Mireshghallah, Z. Lin, S. Gopi, J. Kulkarni, R. Sim, Privacy-preserving in-context learning with differentially private few-shot generation, arXiv preprint arXiv:2309.11765 (2023).

[30] C. Xie, Z. Lin, A. Backurs, S. Gopi, D. Yu, H. A. Inan, H. Nori, H. Jiang, H. Zhang, Y. T. Lee, et al., Differentially private synthetic data via foundation model apis 2: Text, in: ICLR 2024 Workshop on Secure and Trustworthy Large Language Models, ????

[31] S. Zeng, J. Zhang, P. He, J. Ren, T. Zheng, H. Lu, H. Xu, H. Liu, Y. Xing, J. Tang, Mitigating the privacy issues in retrieval-augmented generation (rag) via pure synthetic data, 2025. URL: https://arxiv.org/abs/2406.14773. arXiv:2406.14773.

[32] H. Gao, Y. Zhang, Vrsd: Rethinking similarity and diversity for retrieval in large language models, 2024. URL: https://arxiv.org/abs/2407.04573. arXiv:2407.04573.