ClinicalRAG: Automating Pharmaceutical Label Quality Control with Hierarchical RAG and Large Language Models

Qiaohui Zhou[*†1,2], **Zhongliang Zhou**[*1‡], **Michael Johnson**[1], **Michelle Ngo**[1], **Federico Ferrari**[1], **Junshui Ma**[1]

[1]Merck & Co., Inc., Biometrics Research, Rahway, NJ, USA,
[2]Georgetown University
{zhongliang.zhou, michael.johnson6, michelle.ngo, federico.ferrari, junshui_ma}@merck.com, qz111@georgetown.edu

Abstract

Every pharmaceutical product must be accompanied by a comprehensive label that delineates its indications, usage, dosages, and side effects, essential for safe medication practices. Traditionally, creating drug labels is labor-intensive and dependent on manual quality checks. Recent advancements in Large Language Models (LLMs) offer a promising avenue to streamline this process. In this paper we introduce **ClinicalRAG**, an automated labeling quality control pipeline that integrates LLM with hierarchical Retrieval Augmented Generation that allows to cross-check every statement in the drug label document. ClinicalRAG enhances the reliability of automated drug labeling by systematically reducing hallucination risks, achieving an accuracy of 96.1% in internal validation. With user-friendly interface, our pipeline aims to support pharmaceutical company in drug approval and expedite patients' access to new treatments.

Introduction

The development of new medicines involves several stages before reaching the market (Deore et al. 2019). One critical aspect of this approval process is developing the drug's label (Clarridge, Chin, and Stone 2022). The label provides detailed information on the drug's intended use, dosage instructions, potential side effects, etc. Creating accurate and comprehensive labels is complex and time-consuming, traditionally requiring extensive efforts from domain experts to find tables, listings and figures that are generated within the clinical trial analysis pipeline supporting each statement and value in the label. While accurate, this process can take over several months, posing a potential risk of delaying market entry and patients' access to critical treatments. Recently, large language models (Achiam et al. 2023; Touvron et al. 2023; Enis and Hopkins 2024) have demonstrated significant advancements in natural language understanding and reasoning capabilities (Zhou et al. 2024; Tan et al. 2023). On a smaller scale, AI systems utilizing these models have begun to emerge for tasks such as legal document writing. A major challenge with current large language models is the hallucination problem, where the models generate responses with factual inaccuracies (Yao et al. 2024; Hadi et al. 2023). This issue significantly undermines their reliability in critical applications. One promising solution is retrieval-augmented generation (RAG), which grounds LLM outputs in specific content (Ding et al. 2024). A basic RAG system typically splits data into chunks, translates each chunk into embeddings using an embedding model, and compares an input query to these embeddings to retrieve the most relevant chunks for final answer generation. Due to the heterogeneity of different document types, this simple approach may not effectively retrieve the relevant information needed for accurate answers.

We introduce a hierarchical RAG system to automate label quality control by constructing a hierarchical tree of clinical trial data, where information tables serve as nodes. Using prompt engineering and few-shot learning, paragraphs are broken down into statements, and the system searches the tree for matching tables. Validation is done with a self-consistency LLM approach, achieving 96.1% accuracy in internal case study. ClinicalRAG aims to reduce the time and resources needed for drug approval, accelerating treatment availability. This technology can also be applied to other R&D tasks, such as automated review of critical filing documents and reports.

ClinicalRAG Framework

ClinicalRAG is inspired by expert-driven drug label quality control (QC), where experts extract quantitative information, identify relevant tables, and verify data against label documents. While accurate, this process is slow and expertise-dependent. To streamline it, we restructured the QC process into four modules focused on table identification and verification.

Hierarchical Database Construction

The first step in our pipeline is constructing a database. Clinical studies are registered using RTF tables, which we convert into a unified markdown format using programming and OCR to standardize titles, headers, and data. Identifying the correct table can be challenging due to similar formats across studies. To overcome this, we use an internal file

[*]These authors contributed equally.
[†]This work was conducted during internship at Merck Biometrics Research
[‡]Corresponding Author

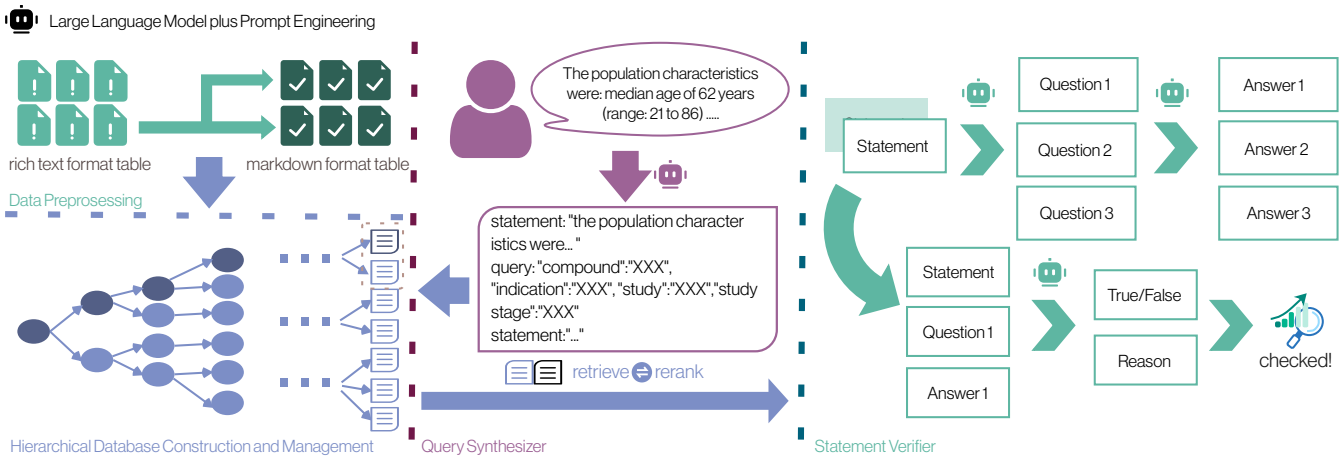

Figure 1: The architecture of the ClinicalRAG pipeline.

hierarchy that organizes information by compound, indications, and protocols, with leaf nodes representing different tables.

Query Synthesizer

ClinicalRAG breaks paragraphs into discrete statements with queries containing hierarchical information for accurate table retrieval in predefined hierarchical database. In this way, it efficiently pairs each statement in the label document with the data from the corresponding table. Using few-shot learning and prompt-based techniques, we guide LLMs to analyze text, retain context, and identify references like trial numbers or indications. This approach improves the system's ability to navigate complex data and verify information effectively.

Hierarchical Database Retriever

After query synthesis, we use a depth-first search (DFS) to locate candidate tables in our hierarchical database, reducing the search space compared to conventional RAG systems. Since multiple tables may be retrieved, we apply an LLM-based reranking to select the top-k most relevant results, ensuring high accuracy in information verification.

Self-Consistency Verifier

To enhance precision and traceability, we convert multi-figure statement into questions and use LLM-based self-consistency TableQA modules for verification. Each figure is verified against retrieved tables in the last step, with validation occurring if the top table confirms it. A color-coded system is implemented in the user interface to highlight areas for human review: yellow for information verified by non-top tables, red for unverified data and no color as successfully verified.

User Interface

We developed a web-based interface for ClinicalRAG. Users can paste paragraphs for QC, select compounds, indications, and protocols, and view corresponding tables with verified results. This transparent process enhances trust and usability by clearly showing how decisions are made.

Model Name	Top-1			Top-3		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
GPT4	0.97	0.33	0.59	1.00	0.30	0.65
Vanilla RAG	0.47	0.97	0.72	0.79	0.94	0.87
ClinicalRAG	0.84	1.00	0.92	0.96	0.94	0.95

Table 1: Comparison of performance using Top-1 and Top-3 most relevant retrieved tables.

Dataset and Case Study

To evaluate the accuracy of our framework, we leveraged MSD's official product KEYTRUDA® label document[1]. We identified 68 quantitative results, annotated by expert statisticians, that required verification. Additionally, we generate the same number of synthetic false data using GPT-4.

We evaluated ClinicalRAG against a Naïve RAG framework and LLM solution (labeled GPT4 in Table 1). The Naïve RAG reranked all tables based on similarity scores between embeddings of tables and statements and use the retrieved tables to generate answers. For the LLM solution, we attached all tables to the context window and asked the model to verify the statement. We evaluated model performance using sensitivity, specificity, and accuracy. As shown in Table 1, ClinicalRAG achieved 92% accuracy and 100% specificity with just one retrieved table, significantly outperforming Naïve RAG and LLM solutions.

Conclusion

We introduce ClinicalRAG, a new pipeline that automates drug labeling quality control with high efficiency and accuracy. Its easy-to-use web interface integrates smoothly into existing workflows, improving transparency and efficiency. This system lessens the workload on human experts and can potentially expedite the drug approval process.

[1]https://www.accessdata.fda.gov/drugsatfda_docs/label/2022/125514s125lbl.pdf

Acknowledgements

We would like to thank Lina Yin and Sabrina Wan from Late Development Statistics, Merck & Co., Inc., Rahway, NJ, USA, for providing the label data and original RTF tables.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Clarridge, K. E.; Chin, S. J.; and Stone, K. D. 2022. Overview of FDA drug approval and labeling. *The Journal of Allergy and Clinical Immunology: In Practice*, 10(12): 3051–3056.

Deore, A. B.; Dhumane, J. R.; Wagh, R.; and Sonawane, R. 2019. The stages of drug discovery and development process. *Asian Journal of Pharmaceutical Research and Development*, 7(6): 62–67.

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

Enis, M.; and Hopkins, M. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. *arXiv preprint arXiv:2404.13813*.

Hadi, M. U.; Qureshi, R.; Shah, A.; Irfan, M.; Zafar, A.; Shaikh, M. B.; Akhtar, N.; Wu, J.; Mirjalili, S.; et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Tan, C.; Cao, Q.; Li, Y.; Zhang, J.; Yang, X.; Zhao, H.; Wu, Z.; Liu, Z.; Yang, H.; Wu, N.; et al. 2023. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. *arXiv preprint arXiv:2312.17016*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zhou, Z.; Zhang, J.; Guan, Z.; Hu, M.; Lao, N.; Mu, L.; Li, S.; and Mai, G. 2024. Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2749–2754.