Enhancing Generative AI Output on Rare Diseases: Development of a Database for Retrieval-Augmented Generation System

Authors: Teppei Okazaki
Toyofumi Fujiwara
Atsuko Yamaguchi

Tokyo City University

Contents

1. Introduction

2. Background
   - What is Rare Diseases?
   - Problems and Solution

3. Methods
   - RAG System
   - Annotation
   - Database
   - Database Query

4. Computational Experiments
   - Evaluation Methods

4.1 Results

4.2 Discission

5. Conclusion

6. Future Work

Contents

1. Introduction

2. Background
   - What is Rare Diseases?
   - Problems and Solution

3. Methods
   - RAG System
   - Annotation
   - Database
   - Database Query

4. Computational Experiments
   - Evaluation Methods

4.1 Results

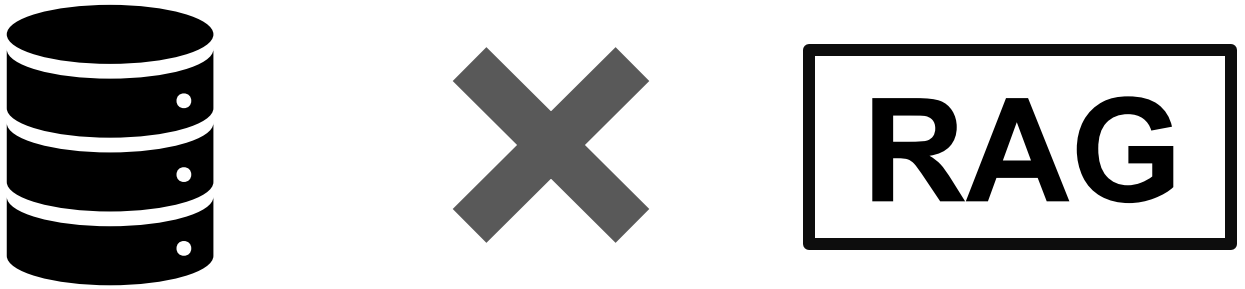
4.2 Discission

5. Conclusion

6. Future Work

Introduction

What we achieved in this study



A rare disease database was developed to implement RAG system that outperformed standard GPT-4o in multiple evaluations.

Backgrounds

What is Rare Diseases?

DEFINITION

US

Diseases Affects less than **200,000** people

EU

Diseases affects less than **1 in 2000** people

TYPES

There are approximately

7,000 rare diseases exist

EFFECT

350 million

people are affected worldwide

Backgrounds

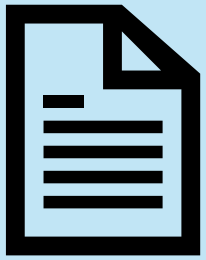
Small number of patients

Insufficient knowledge on Rare diseases

Difficult to obtain accurate information

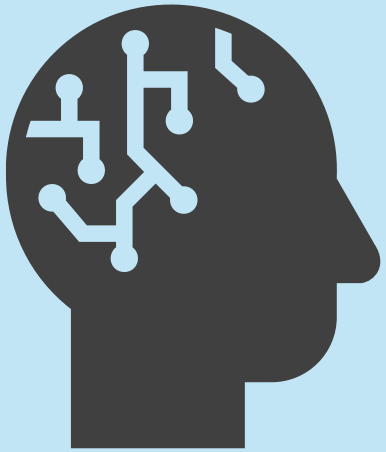
Backgrounds

Research Paper

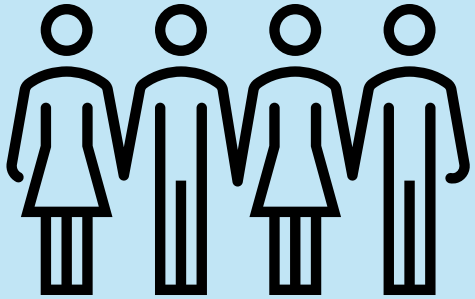


Hard to understand without specialized knowledge

LLM



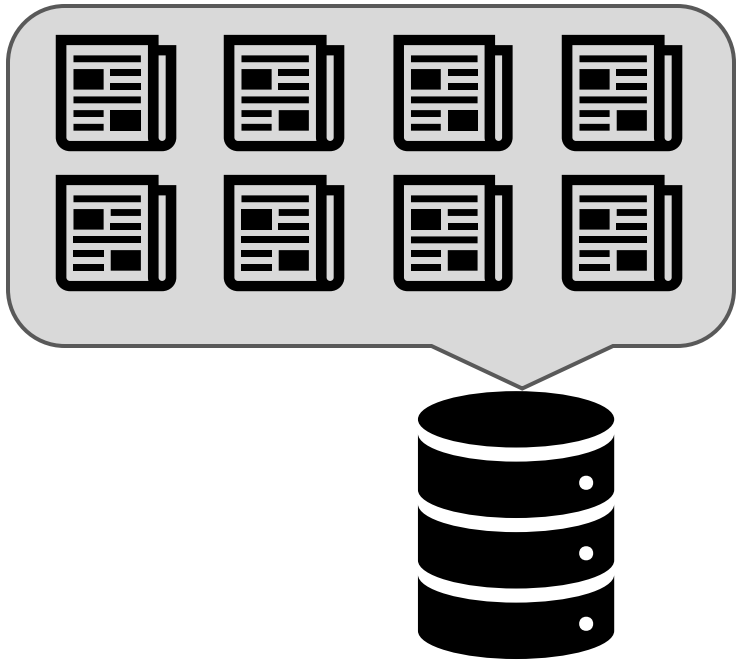
General People



Understandable even for people without specialized knowledge

Methods

What we develop in this study



Database with Medical Literature Published in 2023

×

OpenAI

Contents

1. Introduction

2. Background
   - What is Rare Diseases?
   - Problems and Solution

3. Methods
   - RAG System
   - Annotation
   - Database
   - Database Query

4. Computational Experiments
   - Evaluation Methods

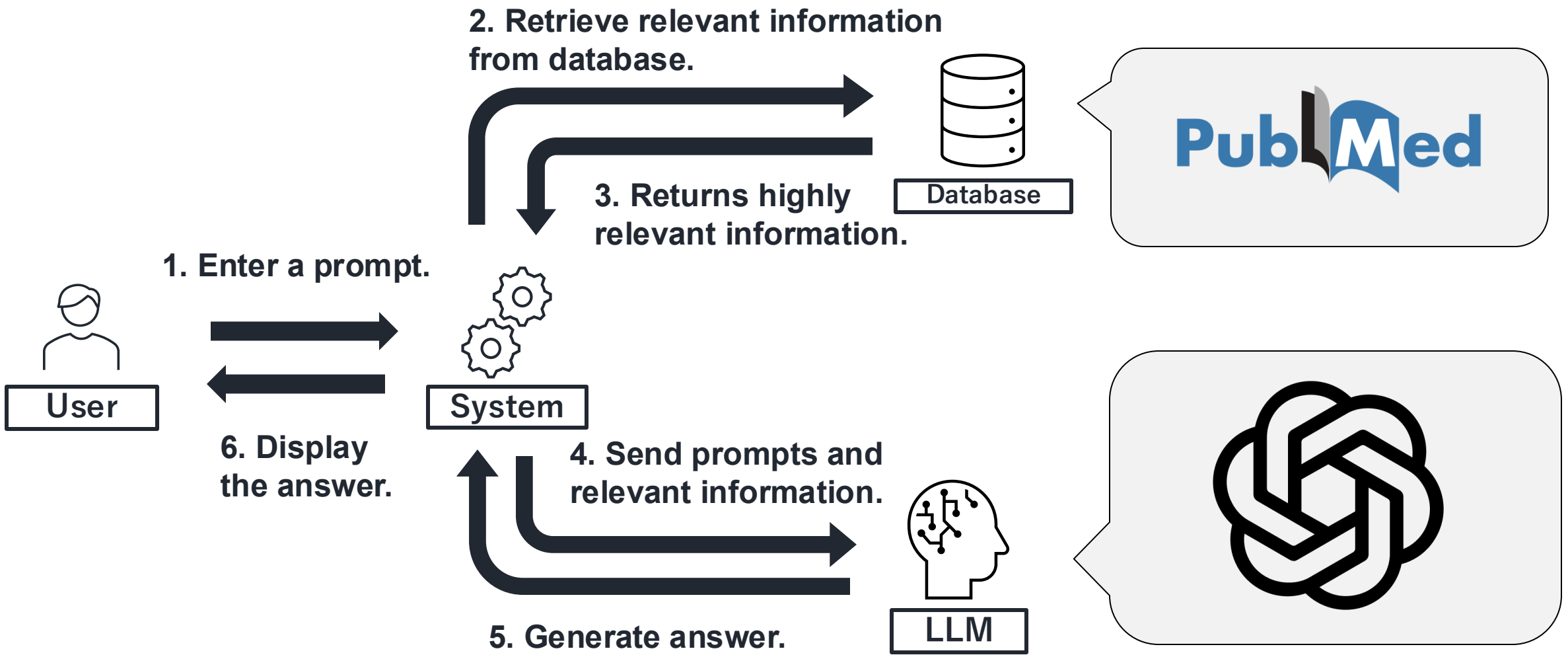
4.1 Results

4.2 Discission

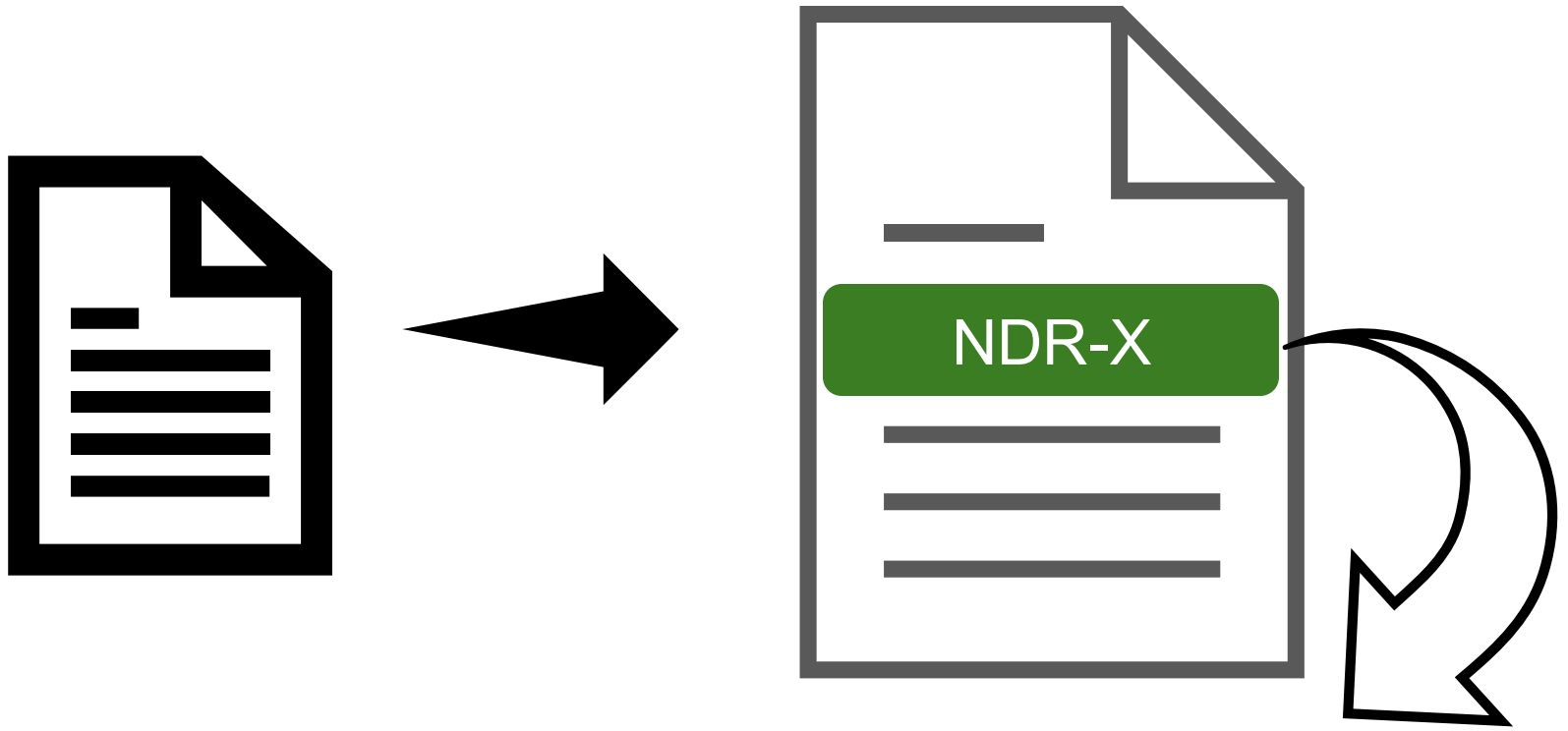
5. Conclusion

6. Future Work

Methods – RAG System

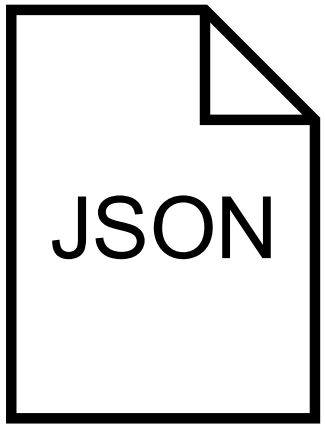


Methods - Annotation



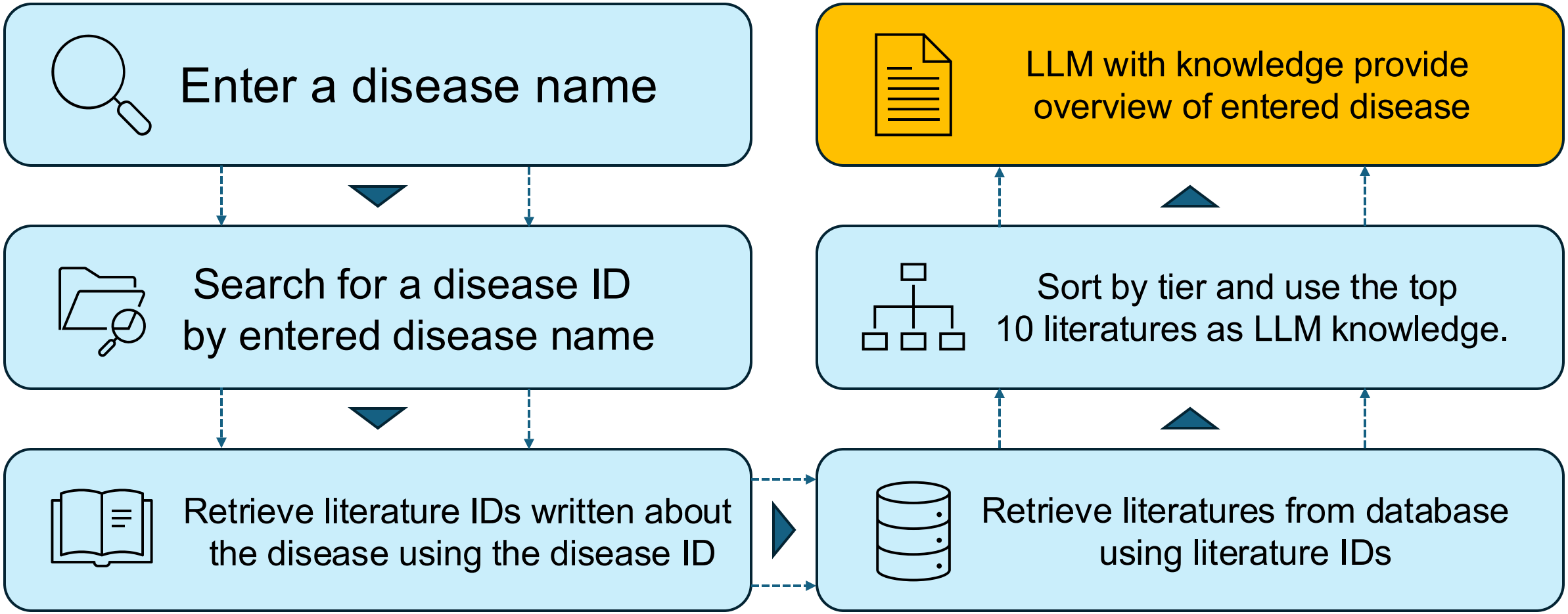
Disease: NDR-X (Non-diagnosable rare disease X)

Methods - Database




```
{
"bibliographic information":{
        "pmid": "12345678",
        "tittle": "Case Report of NDR-X",
        "abstract": "This paper discuss a rare … ",
        "mesh": [
                "id": "D000000",
                "label": "Case Reports"
                ],
                [
                "id": "D000001",
                "label": "Rare Diseases"
                ]
        },
"annotation information":[
        "mondo_id": "mondo:00000001",
        "name": "NDR-X (Non-diagnosable rare disease X)",
        "subset": "rare"
        ]
}
```

Methods - Database Query



Contents

1. Introduction

2. Background
   - What is Rare Diseases?
   - Problems and Solution

3. Methods
   - RAG System
   - Annotation
   - Database
   - Database Query

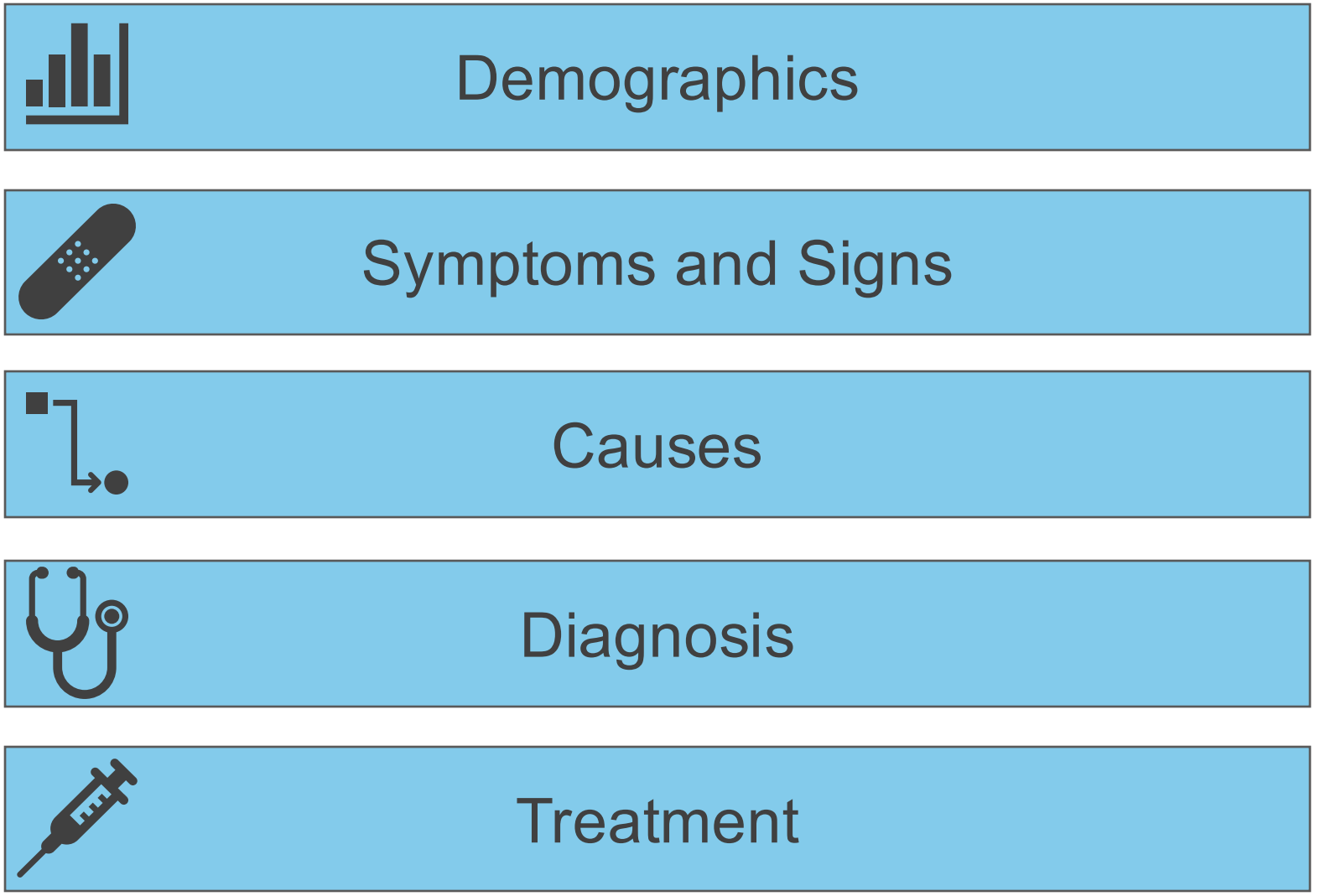
4. Computational Experiments
   - Evaluation Methods

4.1 Results
4.2 Discussion

5. Conclusion

6. Future Work

Computational Experiments



Computational Experiments

Evaluation Method

GPT-4o ×

Reference Set

NORD®
National Organization for Rare Disorders

Evaluation Objects

Standard GPT-4o

VS

GPT-4o with RAG

in 38 Rare Diseases

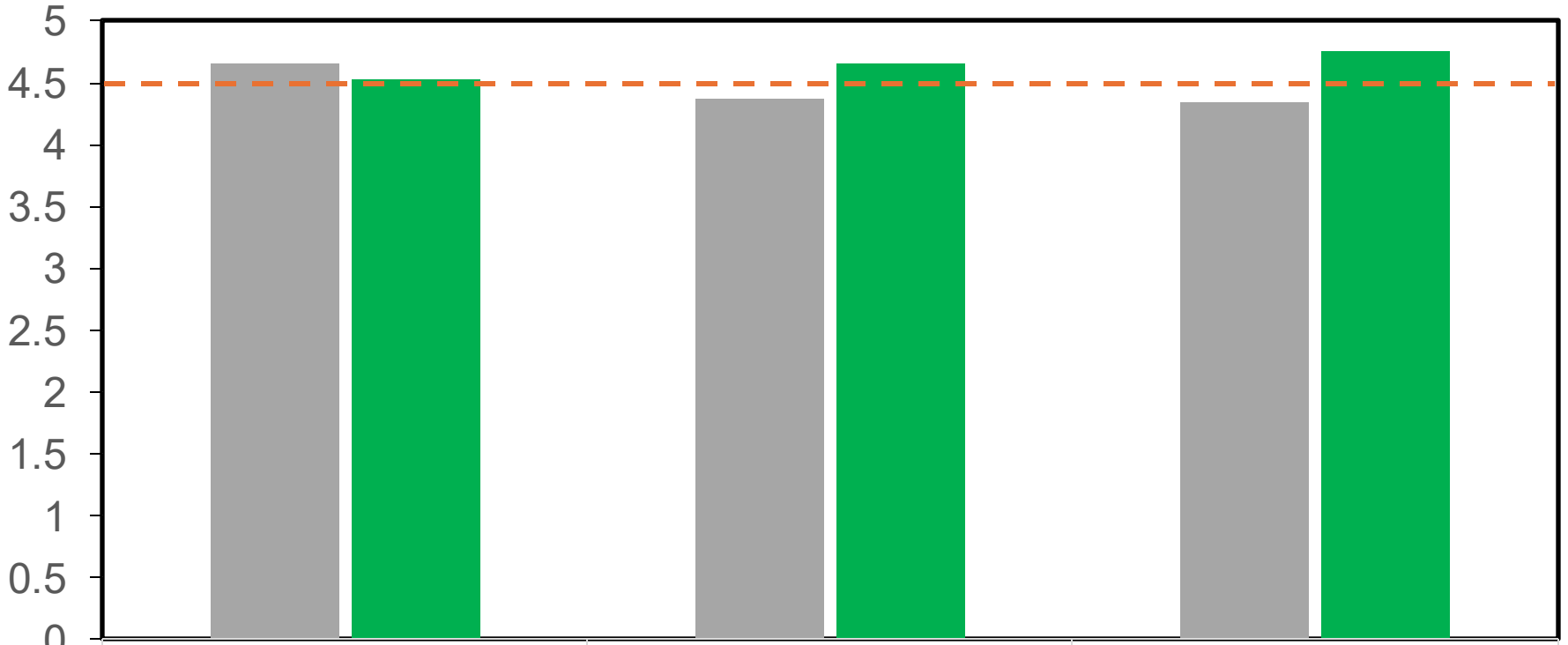
Computational Experiments

Evaluation Criteria | 5-Point Scale

- Correctness
- Comprehensiveness
- Currentness

Results

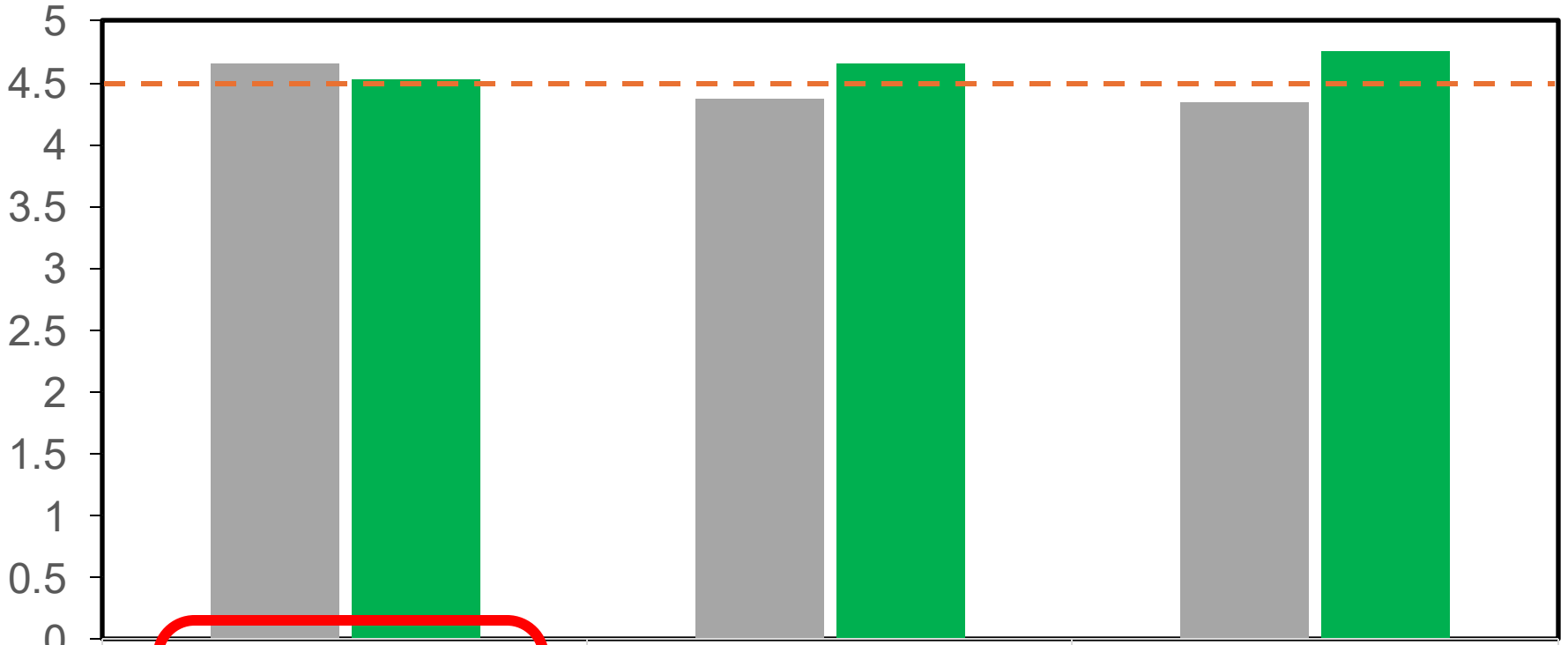
Fig. 1: Comparison of average scores for each criterion



	Correctness	Comprehensiveness	Currentness
Standard GPT-4o	4.6579	4.3684	4.3421
RAG-implemented GPT-4o	4.5263	**4.6579**	**4.7632**

Discussion

Fig. 1: Comparison of average scores for each criterion

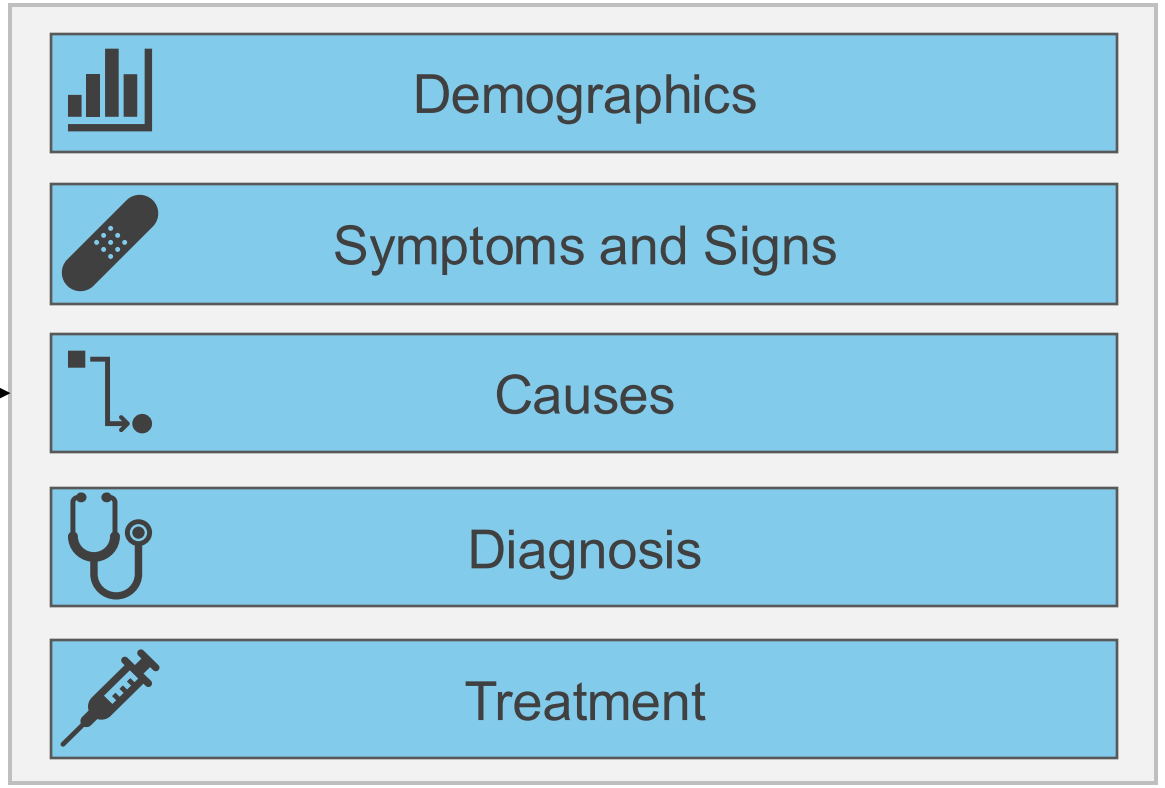


	Correctness	Comprehensiveness	Currentness
Standard GPT-4o	4.6579	4.3684	4.3421
RAG-implemented GPT-4o	4.5263	4.6579	4.7632

Discussion

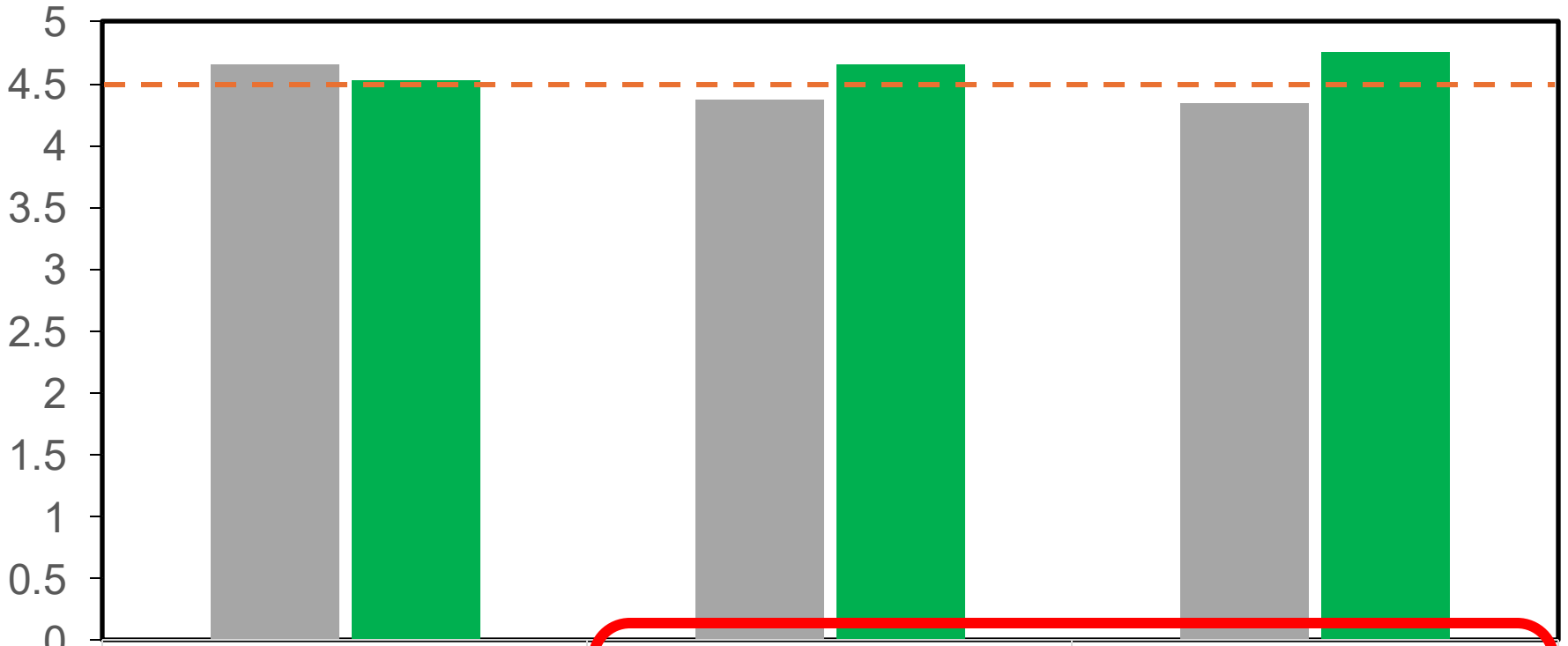
Why did Correctness score lower?

Some of generated overview by RAG-implemented LLM did not follow instructed format



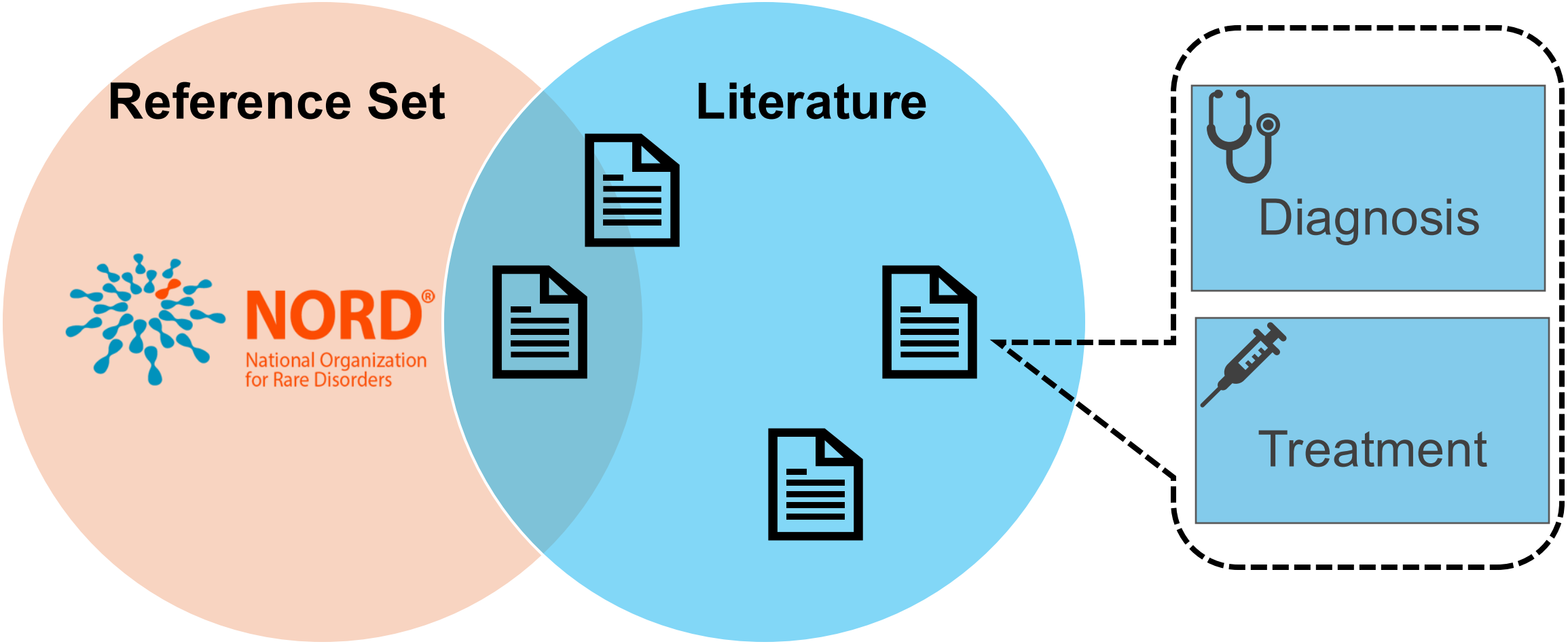
Discussion

Fig. 1: Comparison of average scores for each criterion



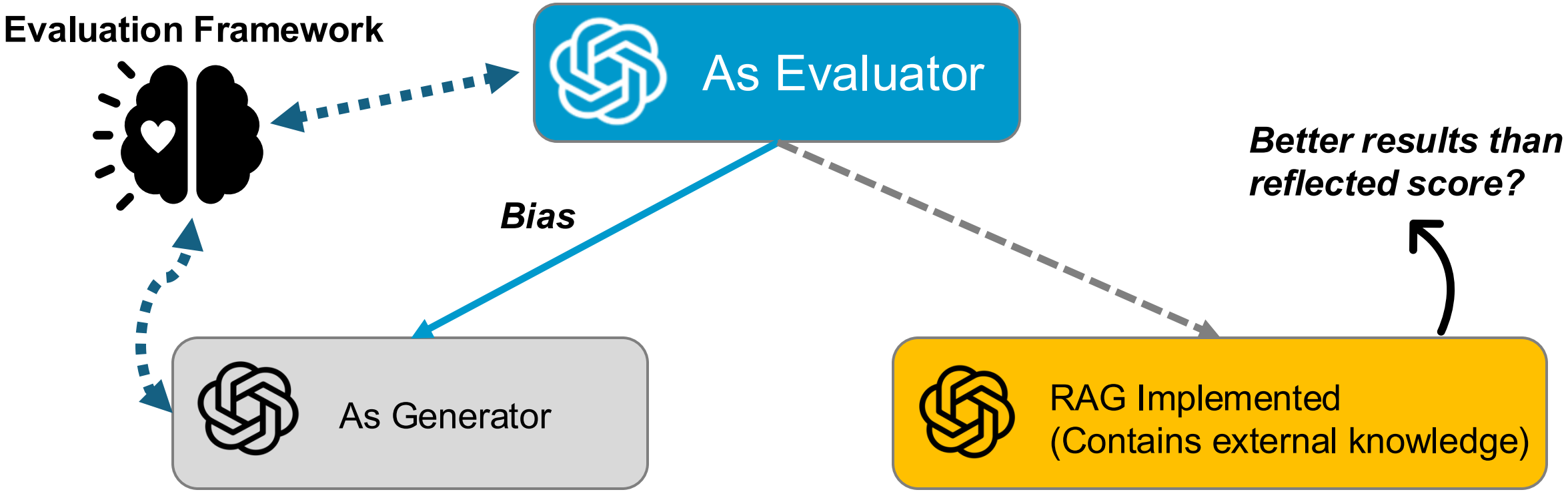
	Correctness	Comprehensiveness	Currentness
Standard GPT-4o	4.6579	4.3684	4.3421
RAG-implemented GPT-4o	4.5263	**4.6579**	**4.7632**

Discussion

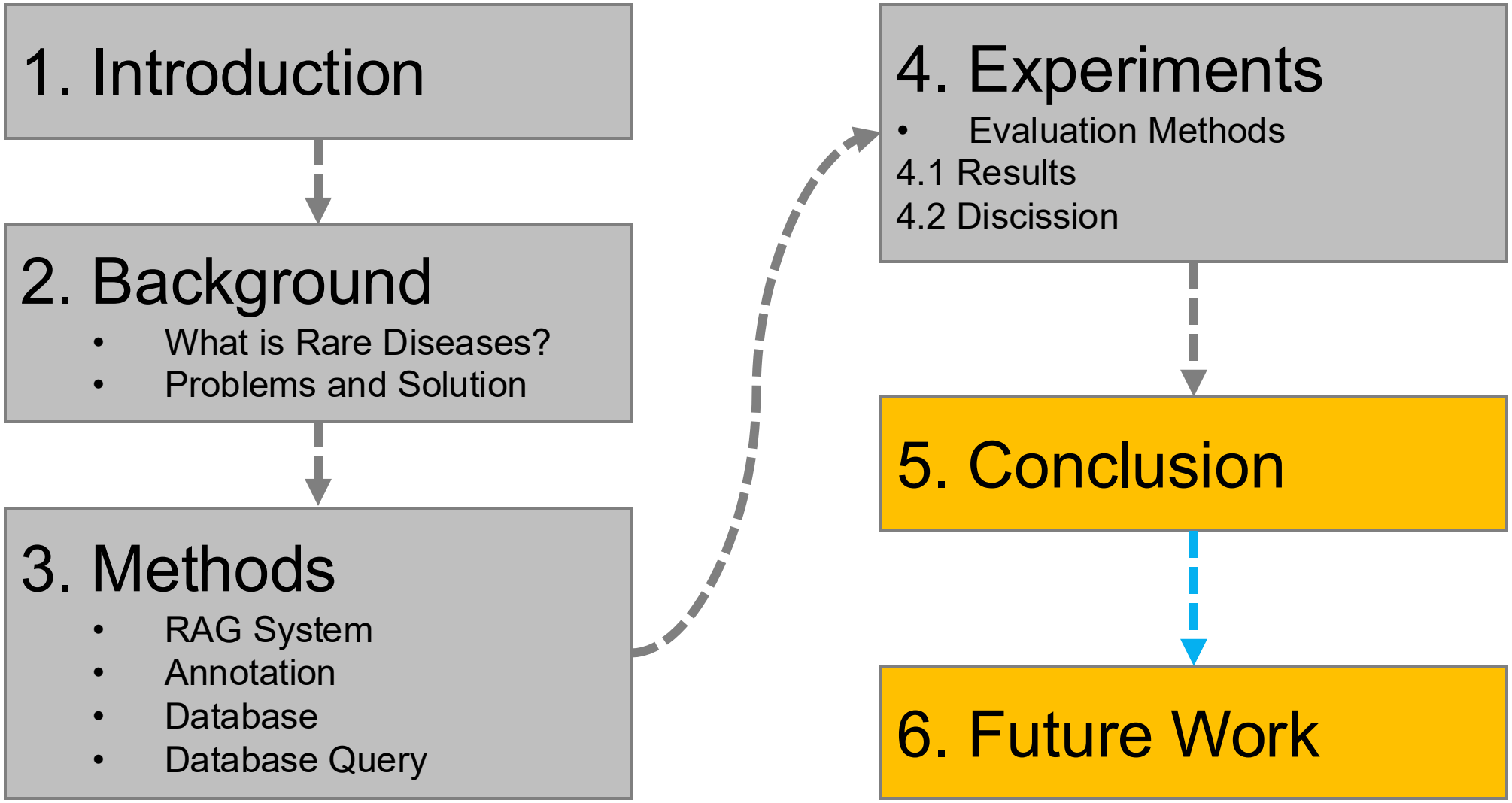

Reference Set
NORD®
National Organization
for Rare Disorders
Literature
Diagnosis
Treatment

Discussion

Sharing the same framework cause a few bias toward Standard GPT-4o

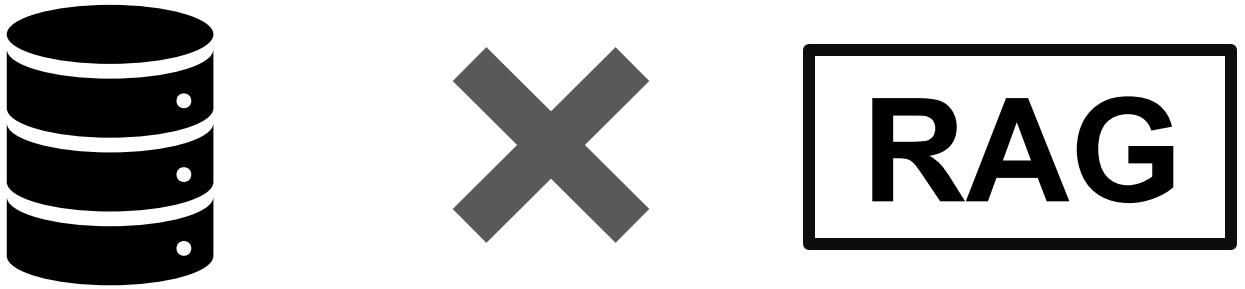


Contents


1. Introduction
2. Background
• What is Rare Diseases?
• Problems and Solution
3. Methods
• RAG System
• Annotation
• Database
• Database Query
4. Experiments
• Evaluation Methods
4.1 Results
4.2 Discission
5. Conclusion
6. Future Work

Conclusion

What we achieved in this study



A rare disease database was developed to implement RAG system that outperformed standard GPT-4o in multiple evaluations.

Future Work

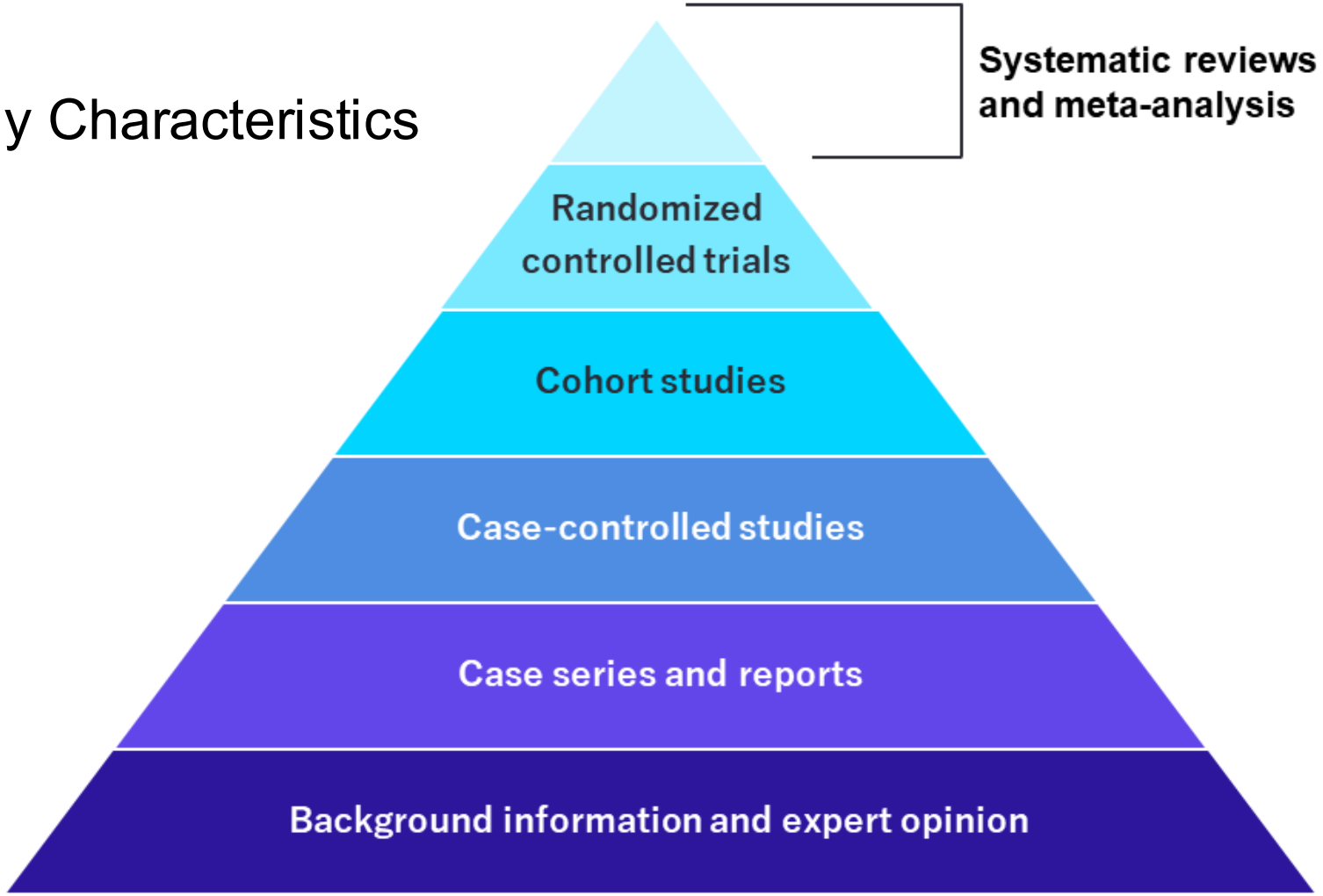
- **Database Usage** - We'll investigate additional ways to leverage and **expand the use of database** we developed.
- **Limited Input** - Currently the input is **the name of the disease**, so we would like to be able to identify it **from symptoms and other information** in the future.
- **Interfaces** - Our goal is to make it **user-friendly for patients, their families and doctors**, to do so, development of an application such as GUI is necessary.
- **Data Limitation** - Currently, the medical literature in the database are limited to those published in 2023. To ensure the comprehensiveness of the information, we aim to **include all literature indexed in PubMed**.

References

[1] Stephen C. Groft, Manuel Posada, and Domenica Taruscio. Progress, challenges and global approaches to rare diseases. Acta Paediatrica, 110(10):2711–2716, Oct 2021. © 2021 Foundation Acta Paediatrica. Published by John Wiley & Sons Ltd.
[2] Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. The Lancet, 371(9629):2039–2041, Jun 2008.
[3] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343, 2023.
[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K ¨ uttler, Mike Lewis, Wen-tau Yih, Tim Rockt ¨ aschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474, 2020.
[5] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.
[6] https://www.rarediseaseday.org/

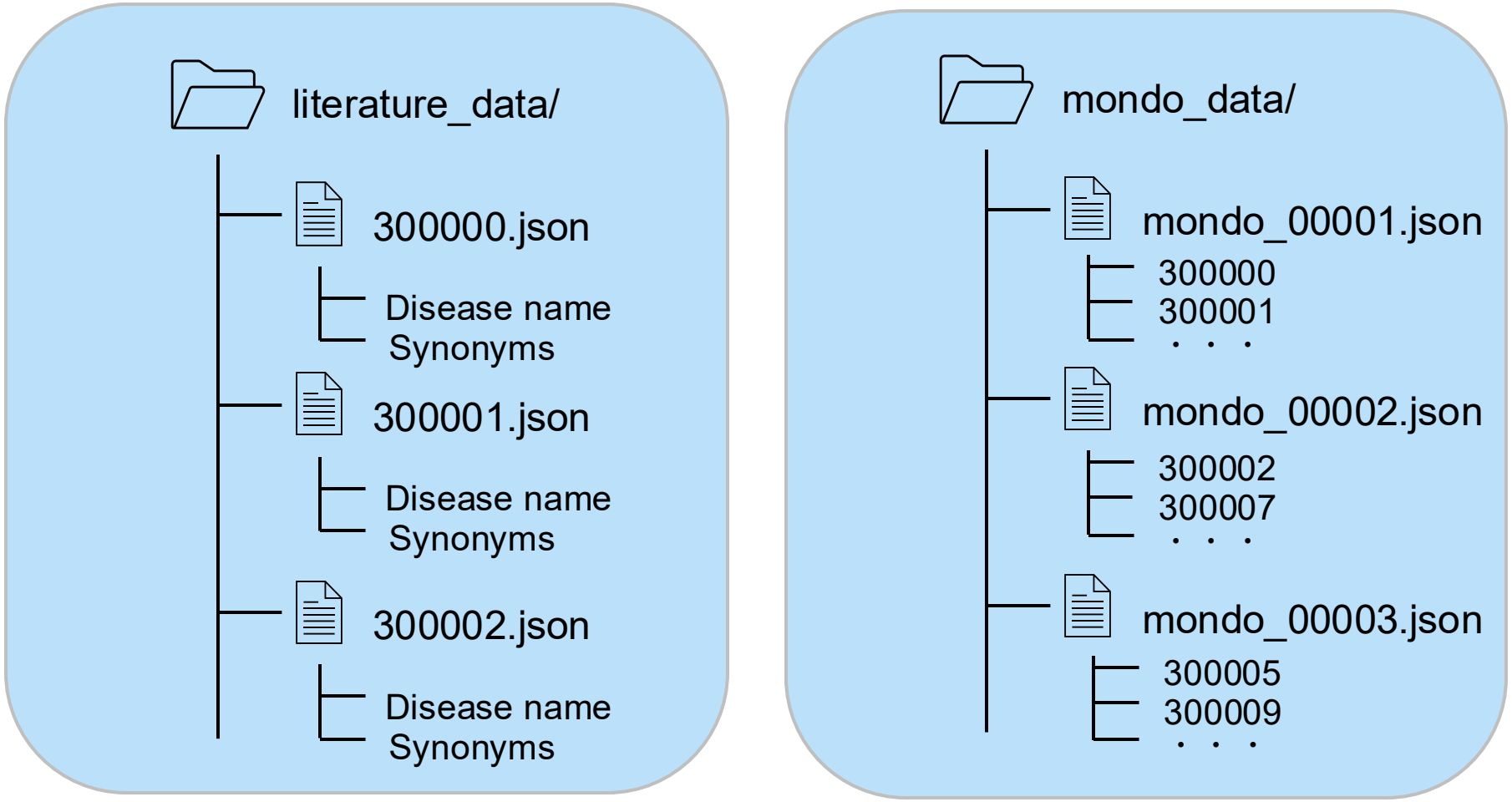
For Questions Answering

Study Characteristics



Brian Sullivan Moira Tannenbaum, Stacy Sebastian. Level of Evidence. https: //openmd.com/guide/levels-of-evidence.

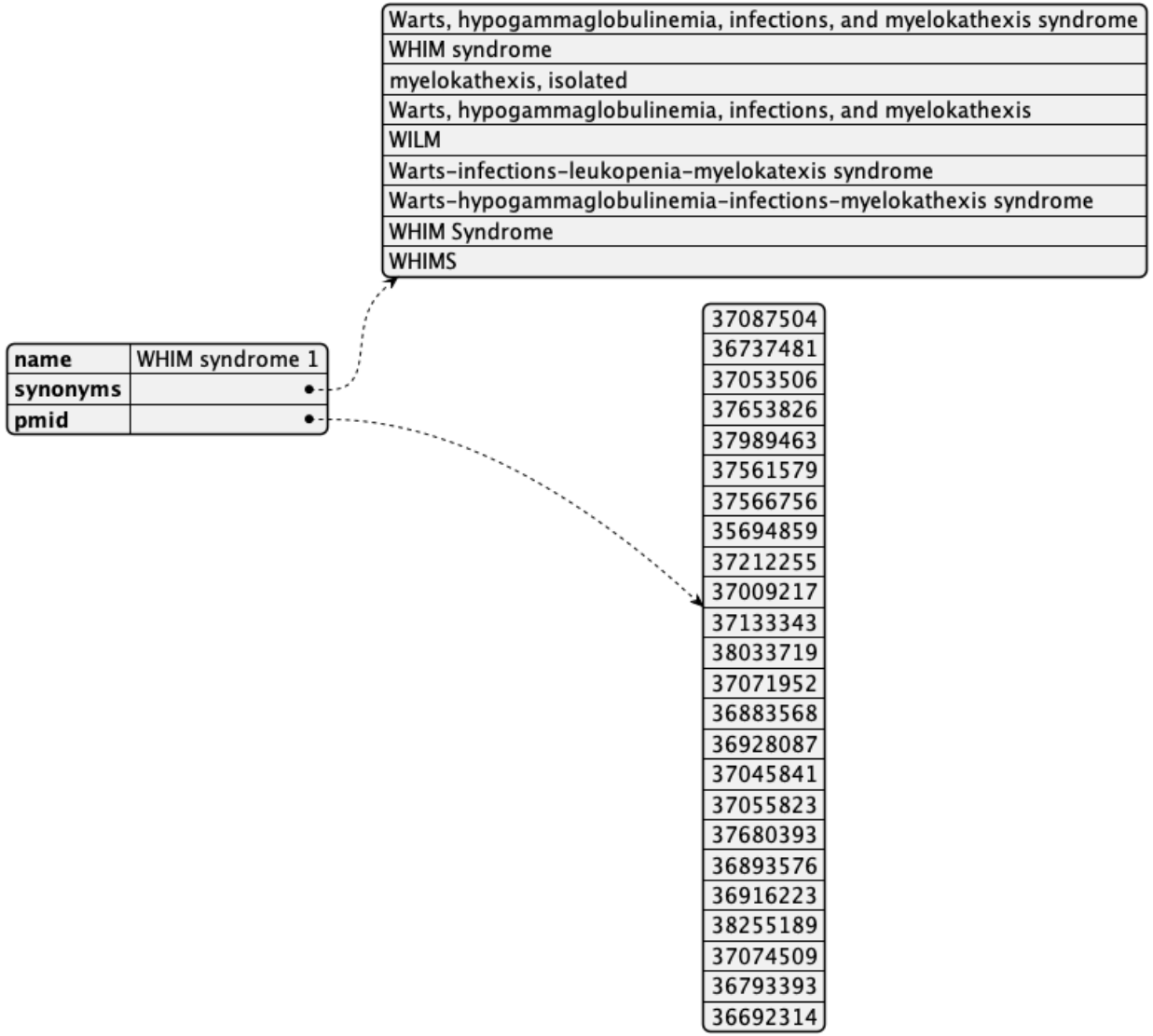
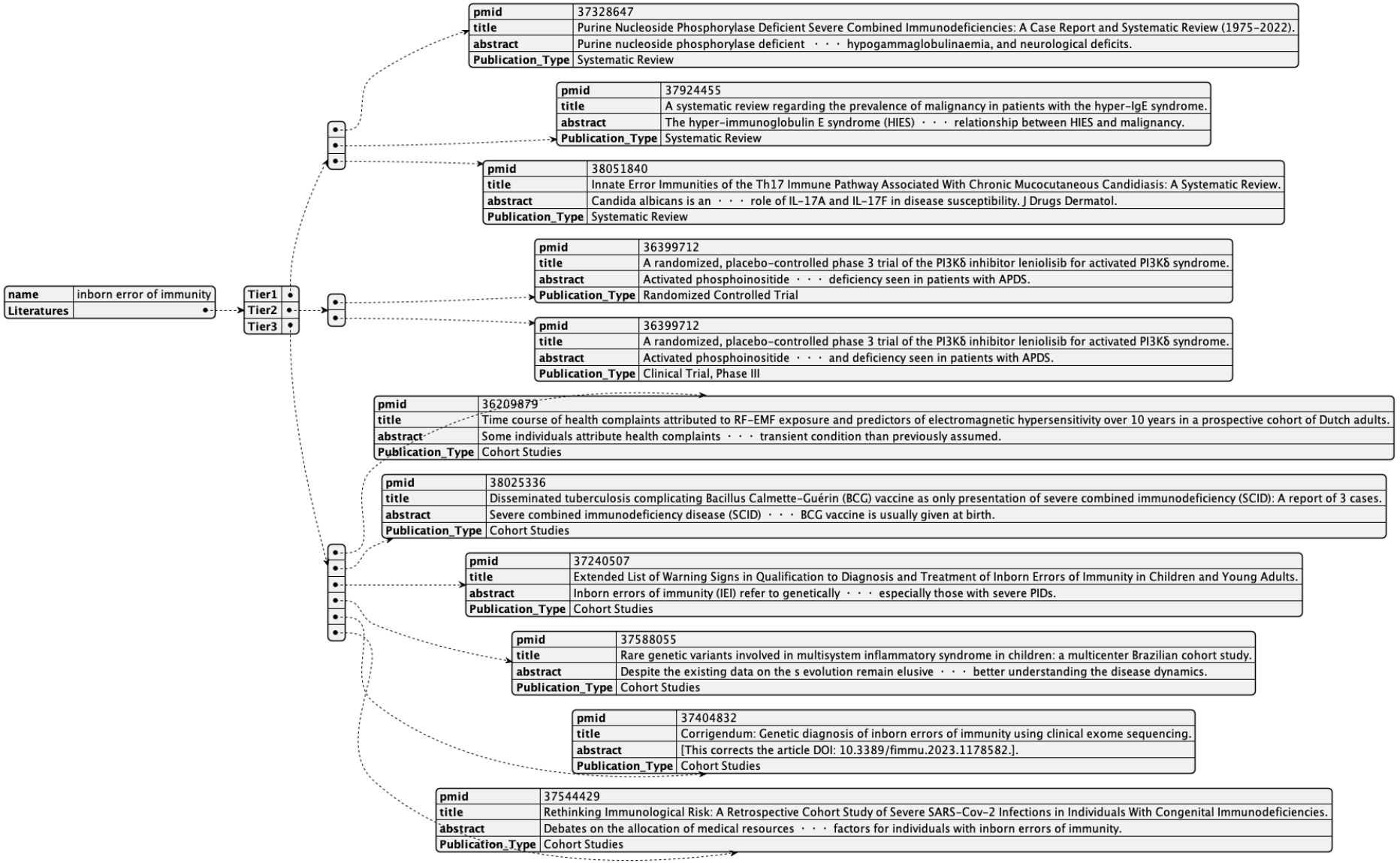
For Questions Answering



For Questions Answering

20 / 38	RAG-implemented model scored higher
7 / 38	Normal GPT-4o scored higher
9 / 11	They scored equally but got full score

For Questions Answering



For Questions Answering