G-RAG: Knowledge Expansion in Material Science

Anonymous Author(s)
Affiliation
Address
email

Abstract

In the field of Material Science, effective information retrieval systems are essential for facilitating research. Traditional Retrieval-Augmented Generation (RAG) approaches in Large Language Models (LLMs) often encounter challenges such as outdated information, hallucinations, limited interpretability due to context constraints, and inaccurate retrieval. To address these issues, Graph RAG integrates graph databases to enhance the retrieval process. Our proposed method processes Material Science documents by extracting key entities (referred to as MatIDs) from sentences, which are then utilized to query external Wikipedia knowledge bases (KBs) for additional relevant information. We implement an agent-based parsing technique to achieve a more detailed representation of the documents. Our improved version of Graph RAG called G-RAG further leverages a graph database to capture relationships between these entities, improving both retrieval accuracy and contextual understanding. This enhanced approach demonstrates significant improvements in performance for domains that require precise information retrieval, such as Material Science.

1 Introduction

LLMs exhibit impressive capabilities but encounter challenges such as hallucinations, outdated information, and untraceable, opaque reasoning. The RAG approach addresses these issues by combining the strengths of LLMs with the vast, continuously updated resources of external databases [1]. Graph-enhanced RAG methods build on this by leveraging rich semantic interconnections and relational data, enabling more precise entity linking, enhanced semantic context, and improved knowledge extraction [2, 3]. Additionally, researchers have introduced innovative graph-based context adaptation techniques that refine word embeddings to better capture semantic relationships, consistently outperforming traditional methods in various Natural Language Processing (NLP) tasks [4, 5]. Graph-based RAG provides a more nuanced and accurate representation of complex domains, enabling LLMs to generate responses with enhanced factual precision and contextual relevance [6, 7]. This capability is especially valuable for domain-specific applications in fields such as material science and biomedicine, where accurate and detailed information is crucial [8, 9, 10]. Serving as a domain-specific knowledge server, the Semantic Context Enhancer extracts and delivers detailed descriptions of relevant concepts and entities, including their interrelationships, thereby equipping the LLM with a deeper semantic understanding [10]. Additionally, leveraging graph structures to improve knowledge retrieval and response generation, as exemplified by methods like AriGraph, has shown significant enhancements in decision-making and planning capabilities [11]. This study explores the improvement of information retrieval and knowledge generation in complex, specialized domains through the integration of the G-RAG pipeline, addressing limitations of existing approaches and advancing performance in targeted fields.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

2 Methodology

The retrieval process of Naive RAG includes a diverse range of MatIDs, which ensures variety but can also introduce less relevant information. This issue can be mitigated through prompt engineering in the RAG configuration, allowing the LLM to continue generating accurate responses [12]. However, there are two main limitations to this approach. First, LLMs have a fixed context window, which restricts the number of tokens they can process simultaneously. This limitation hinders the model's ability to manage large volumes of retrieved data effectively [13], especially when the dataset is extensive and varied. Despite advancements like Google's Gemini, which uses a caching system to handle extended contexts, the fixed context window of LLMs remains a significant constraint [14, 15]. Although providing the model with more relevant information might seem beneficial, increasing the context length does not necessarily improve the accuracy of information retrieval or response generation [16]. This problem becomes even more pronounced when the retrieved context includes a mix of diverse but only marginally relevant data, potentially diluting the focus on the critical entities or concepts needed for an accurate response [17]. This is where Graph RAG proves to be valuable, as it enhances the retrieval process by focusing on the most relevant information.

2.1 Graph RAG vs G-RAG

Graph RAG effectively merges the strengths of retrieval-based and generative methods to enhance LLMs' capability to generate accurate, relevant, and contextually enriched responses [18]. While supplying an LLM with text chunks from extensive documents may result in issues with context, factual precision, and language coherence, Graph RAG addresses these limitations by utilizing a knowledge graph as a source of structured, factual information [19]. The knowledge graph provides detailed entity information, including attributes and relationships, allowing the LLM to gain a deeper understanding and produce more informed, precise responses. In our G-RAG system, entity linking is a fundamental component, enabling the extraction of specific entities (key terms or concepts) from the text using an entity extractor like a Span Parser. These identified entities are then used to query an external retriever, which fetches relevant MatIDs and their corresponding information from a Wikipedia knowledge base [20]. This targeted retrieval process ensures that the selected MatIDs are highly relevant and accurate, thereby preserving the integrity of the constructed knowledge graph [21]. Following this, an LLM formulates a query that is sent to the graph database. The graph database retrieves relevant information, which is processed by the LLM to generate a final, comprehensive response. The complete architecture of our G-RAG system is illustrated in Figure 1.

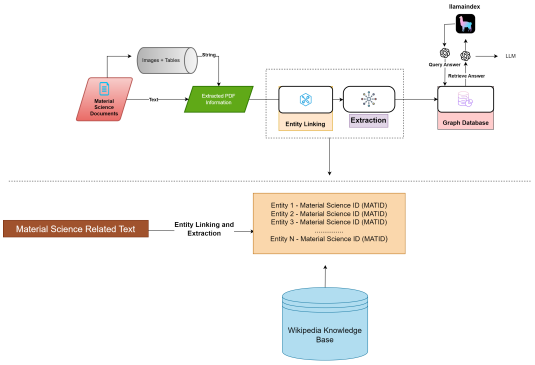

Figure 1: Architecture of G-RAG System

2.2 PDF Parsing

We parse PDFs by categorizing their content into text, figures, and tables. For figure extraction, we employ the Phi-3.5 Vision Instruct model, specifically tailored to identify material science-related images using a vision agent system. We utilize Microsoft's Table Transformer in the tabular data extraction process. Furthermore, we apply a smart chunking technique to enhance the precision of data segmentation [22]. Accurate parsing is essential for subsequent tasks such as Entity Linking, Relation Extraction, and Graph Retrieval Augmented Generation, as it ensures the accuracy and

relevance of the answers retrieved from the database. Appendix A.1 provides a detailed overview of our document parsing process.

2.3 Entity Linking and Relation Extraction

Entity Linking (EL) refers to the process of mapping ambiguous mentions in a text to specific, identifiable named entities within a knowledge base [23]. It involves recognizing all potential entities mentioned in the given input and accurately associating them with corresponding entries in a reference knowledge base, such as Wikipedia. Relation Extraction (RE) refers to the process of identifying and classifying semantic relationships between entities mentioned within a given text. This task involves mapping the detected entities to specific relation categories defined in a reference knowledge base, such as Wikipedia. The entity linking and relation extraction process is depicted in Appendix A.2.

2.4 Span Parser

The Span Parser module functions as our G-RAG system's initial information retrieval component, employing an approach inspired by the Retrieval Process [24]. This module operates on the principle of semantic similarity between the current knowledge base (KB) and a comprehensive collection of textual passages (Wikipedia Database) representing entities and relations. At its core, the Span Parsing module utilizes an encoder to generate dense vector representations of both the knowledge base (KB) q and each passage p in the additional knowledge base collection. These representations, denoted as $E(q)$ and $E(p)$ respectively, are high-dimensional embedding that capture the semantic content of the text. The module computes a similarity score between the current Knowledge Base and additional Knowledge Base (Wikipedia data) using a dot product operation, yielding the most relevant relations with respect to the extended knowledge base q:

$$\text{sim}(q, p) = E(q)^\top \cdot E(p)$$

This score quantifies the relevance of each passage of the additional knowledge base to the given current KB passage's sentence, enabling the module to rank and retrieve the most pertinent information.

2.5 Passage Processor

The Passage Processor (PP) component in our G-RAG system employs a unified approach to process the existing knowledge base and retrieved passages. Given a current Knowledge Base (KB) Q and a set of N retrieved passages $\{P_1, \ldots, P_n\}$, the Passage Processor constructs chunks of current KB. In each chunk, we utilize each input sequence $S = [Q; \tau_0; P_1; \tau_1; \ldots; P_n; \tau_n]$, where τ_i are delimiter tokens. This sequence is encoded using a Transformer model T, producing contextual embedding $E = T(S)$. The Passage Processor subsequently identifies relevant spans within Q through a two-stage process [24]. Initially, it computes start probabilities $P^s(q_i)$ for each token q_i in Q using a learned function $f^s(E)$. Subsequently, for each potential start position s, it calculates end probabilities $P^e(q_j \mid s)$ for tokens q_j (where $j \geq s$) using another learned function $f^e(E, s)$. This formulation enables the prediction of overlapping spans, enhancing the model's capability to handle complex queries. During the process, spans (s, e) are predicted if $P^s(q_s) > \theta_s$ and $P^e(q_e \mid s) > \theta_e$, where θ_s and θ_e are predefined thresholds. This design enables the Passage Processor to process the entire knowledge base chunk by chunk efficiently, identifying relevant text spans for downstream tasks such as entity linking and relation extraction.

3 Experimental Settings

Our dataset consists of ten carefully designed handwritten queries, aimed at evaluating and differentiating the capabilities of various RAG systems. Sample queries from this dataset are presented in Appendix A.3. To evaluate the performance of RAG systems, we employ various metrics, including correctness, faithfulness, context, and answer relevancy scores. Correctness assesses the accuracy of the generated response, while faithfulness evaluates the factual accuracy based on the retrieved documents. Finally, the context and answer relevancy score measures how well the response aligns with the given query. A detailed description of these evaluation metrics is provided in Appendix A.4. For entity linking and relation extraction, we use the relik-entity-linking-large model [25], while the jina-embeddings-v2-base-en model [26], with a sequence length of 8192, is employed for

embeddings. Additionally, we utilize LLama 3.1 8B and LLama 3.1 70B as large language models, both of which produce comparable results.

4 Results and Discussion

This section presents all of our experimental results. We conducted the computational tasks using the NVIDIA Tesla A100 Ampere 40 GB GPU. The performance of the Naive RAG, Graph RAG, and the G-RAG system was evaluated using our dataset. Appendix A.5 provides example queries and the corresponding responses from the RAG systems, evaluated across different metrics. The experimental results are summarized in Table 1.

Table 1: Experimental Results

Pipeline	Score	No. of queries	Mean	Standard Deviation
Naive RAG	Correctness	10	2.43	1.51
	Faithfulness		0.70	0.48
	Relevancy		**0.39**	0.28
Graph RAG	Correctness	10	3.30	2.00
	Faithfulness		**0.90**	**0.32**
	Relevancy		0.18	**0.26**
G-RAG	Correctness	10	**3.90**	**1.10**
	Faithfulness		**0.90**	**0.32**
	Relevancy		0.34	0.32

The comparative analysis of three RAG pipelines - Vector/Naive RAG, G-RAG, and Graph RAG - showed interesting patterns in their performance across three critical dimensions. A one-way Analysis of Variance (ANOVA) as described in Appendix A.4.5 was performed, examining correctness $F(2, 24) = 2.39$, $p = 0.113$, faithfulness $F(2, 27) = 1.04$, $p = 0.368$, and context and answer relevancy $F(2, 27) = 1.04, p = 0.368$. While no statistically significant differences were found at the standard significance level ($\alpha = 0.05$), the descriptive statistics highlighted meaningful variations in performance. Specifically, Vector/Naive RAG outperformed the others in terms of context relevancy, with a mean score of 0.3875. This was followed by G-RAG (mean score of 0.3375), while Graph RAG exhibited the lowest mean score of 0.1750. The substantial standard deviations observed across all pipelines, ranging from 0.2630 to 0.3162, suggest notable performance variability depending on the query. This variability highlights the challenge of consistency in RAG systems. The superior performance of G-RAG over the basic Graph RAG can be attributed to the inclusion of a material science knowledge base, emphasizing the critical role of domain-specific knowledge in enhancing model accuracy. The superior context relevancy performance of the traditional Vector/Naive RAG challenges the assumption that graph-based approaches inherently provide better retrieval capabilities. G-RAG has proven to be a well-rounded solution, effectively balancing the metrics of correctness, relevancy, and faithfulness. The significant drop in relevancy scores for Graph RAG highlights the critical role of entity linking in G-RAG's design. This suggests that the effectiveness of knowledge integration mechanisms, including entity linking, plays a substantial role in improving retrieval performance. These findings indicate that while graph-based approaches show promise, their success heavily depends on the quality of knowledge integration and the sophistication of the entity-linking.

5 Conclusion and Future Work

Our findings indicate that integrating graph-based techniques and ensuring robust entity linking with external databases can significantly enhance the performance of the Graph RAG pipeline, particularly in terms of response relevance and accuracy. This approach also mitigates the challenge of maintaining relevance observed in standard Graph RAG implementations. Future work could include developing a larger knowledge base tailored to material science as an extended information source, as well as creating a material science-specific entity linking model. Additionally, establishing a comprehensive evaluation metric for Graph RAG would provide deeper insights into the process and its effectiveness.

References

[1] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[2] Sanat Sharma, David Seunghyun Yoon, Franck Dernoncourt, Dewang Sultania, Karishma Bagga, Mengjiao Zhang, Trung Bui, and Varun Kotte. Retrieval augmented generation for domain-specific question answering. *arXiv preprint arXiv:2404.14760*, 2024.

[3] Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, H. Qu, and Jian Guo. Think-on-graph 2.0: Deep and interpretable large language model reasoning with knowledge graph-guided retrieval. *arXiv preprint arXiv:2407.10805*, 2024.

[4] Tanvi Sandhu and Ziad Kobti. Exploration of word embeddings with graph-based context adaptation for enhanced word vectors. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*, 2024.

[5] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024.

[6] What Is Graph RAG? — ontotext.com. https://www.ontotext.com/knowledgehub/fundamentals/what-is-graph-rag/. [Accessed 02-09-2024].

[7] Chansol Park, Hayoung Lee, and Ok-Ran Jeong. Leveraging medical knowledge graphs and large language models for enhanced mental disorder information extraction. *Future Internet*, 2024.

[8] Markus J. Buehler. Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design. *ACS Engineering Au*, 2024.

[9] Julien Delile, Srayanta Mukherjee, A. V. Pamel, and Leonid Zhukov. Graph-based retriever captures the long tail of biomedical knowledge. *arXiv preprint arXiv:2402.12352*, 2024.

[10] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges, 2023.

[11] Petr Anokhin, Nikita Semenov, A. N. Sorokin, Dmitry Evseev, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *arXiv preprint arXiv:2407.04363*, 2024.

[12] Thomas Merth, Qichen Fu, Mohammad Rastegari, and Mahyar Najibi. Superposition prompting: Improving and accelerating retrieval-augmented generation, 2024.

[13] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023.

[14] Cunchen Hu, Heyang Huang, Junhao Hu, Jiang Xu, Xusheng Chen, Tao Xie, Chenxi Wang, Sa Wang, Yungang Bao, Ninghui Sun, and Yizhou Shan. Memserve: Context caching for disaggregated llm serving with elastic memory pool, 2024.

[15] Xiaohua Wang, Zhenhua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Li Wang, Shizheng Li, Qian Qi, et al. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*, 2024.

[16] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

[17] Anand Subramanian. Building a Biomedical Entity Linker with LLMs — towardsdatascience.com. https://towardsdatascience.com/building-a-biomedical-entity-linker-with-llms-d385cb85c15a. [Accessed 02-09-2024].

[18] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.

[19] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[20] Wikipedia Knowledge Base. relik-ie/relik-reader-deberta-v3-large-re-wikipedia. [Accessed 04-09-2024].

[21] Entity Linking and Relationship Extraction With Relik in LlamaIndex — neo4j.com. https://neo4j.com/developer-blog/entity-linking-relationship-extraction-relik-llamaindex/. [Accessed 02-09-2024].

[22] Darren Oberst, MacOS, Jeff Turnham, Jessica Berliner, Will Taner, Prashant Rajesh Iyer, NYDocutest, Adelina Jiang, Aryan Chauhan, Christopher Harrison, SNEHA KUMARI, Virendra Singh, simonrisman, shneeba, Vedant Sudhir Patil, Rohan Sharma, Raghav Dixit, Rahul Khandait, Uğur Çekmez, Ayushri, adithyasudhan, Chelsea, Ujjwal Jha, momodingaling, osi1880vr, Kamalakar Satapathi, philipkd, Rajesh Adhikari, Saurav Kumar Mahato, and Seva Skvortsov. llmware-ai/llmware, 9 2024.

[23] Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. Named entity recognition for entity linking: What works and what's next. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen tau Yih, editors, *EMNLP (Findings)*, pages 2584–2596. Association for Computational Linguistics, 2021.

[24] Riccardo Orlando, Pere-Lluis Huguet-Cabot, Edoardo Barba, and Roberto Navigli. Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget, 2024.

[25] Relik Entity Linking Model. https://huggingface.co/sapienzanlp/relik-entity-linking-large. [Accessed 02-09-2024].

[26] Jina Embeddings Model. https://huggingface.co/jinaai/jina-embeddings-v2-base-en. [Accessed 02-09-2024].

A Appendix

A.1 Documents Parsing Method

This section illustrates our document parsing pipeline, as shown in Figures 2 and 3. Efficient document parsing is crucial for enabling RAG systems to generate responses with high factual accuracy and precision.

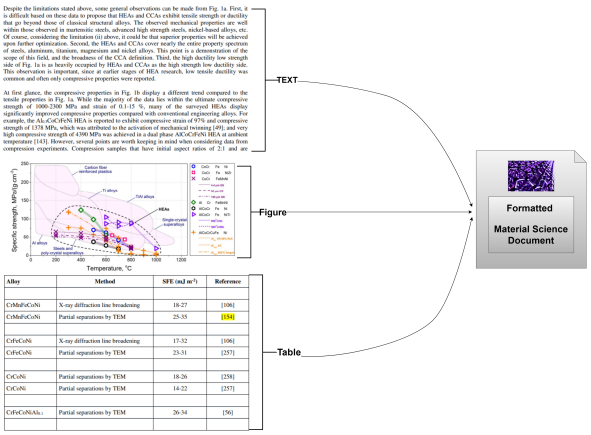

Figure 2: Document Parsing

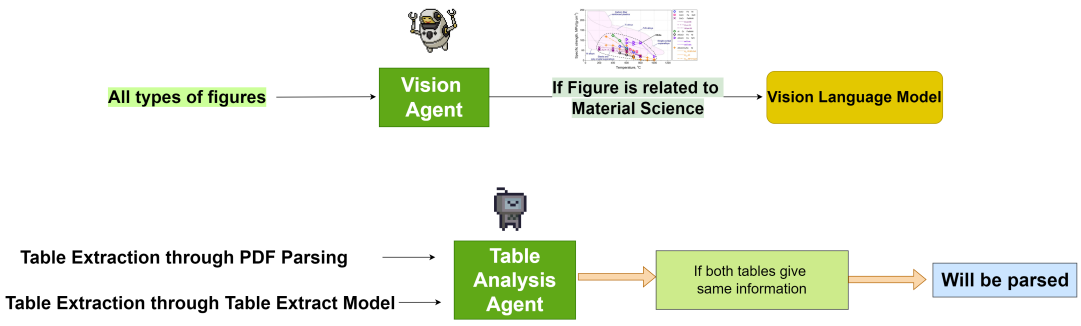

Figure 3: Validity Check by Agent System

A.2 Entity Linking and Relation Extraction

In this section, we provide a visual representation of the entity linking and relation extraction process, as depicted in Figures 4, 5, 6, and 7. These processes are essential components of our G-RAG system.

Coreference Resolution: Coreference resolution, mentioned in Figure 4 involves identifying different expressions in a text that refer to the same entity. This process is crucial for understanding the relationships between various mentions of an entity within a given context.

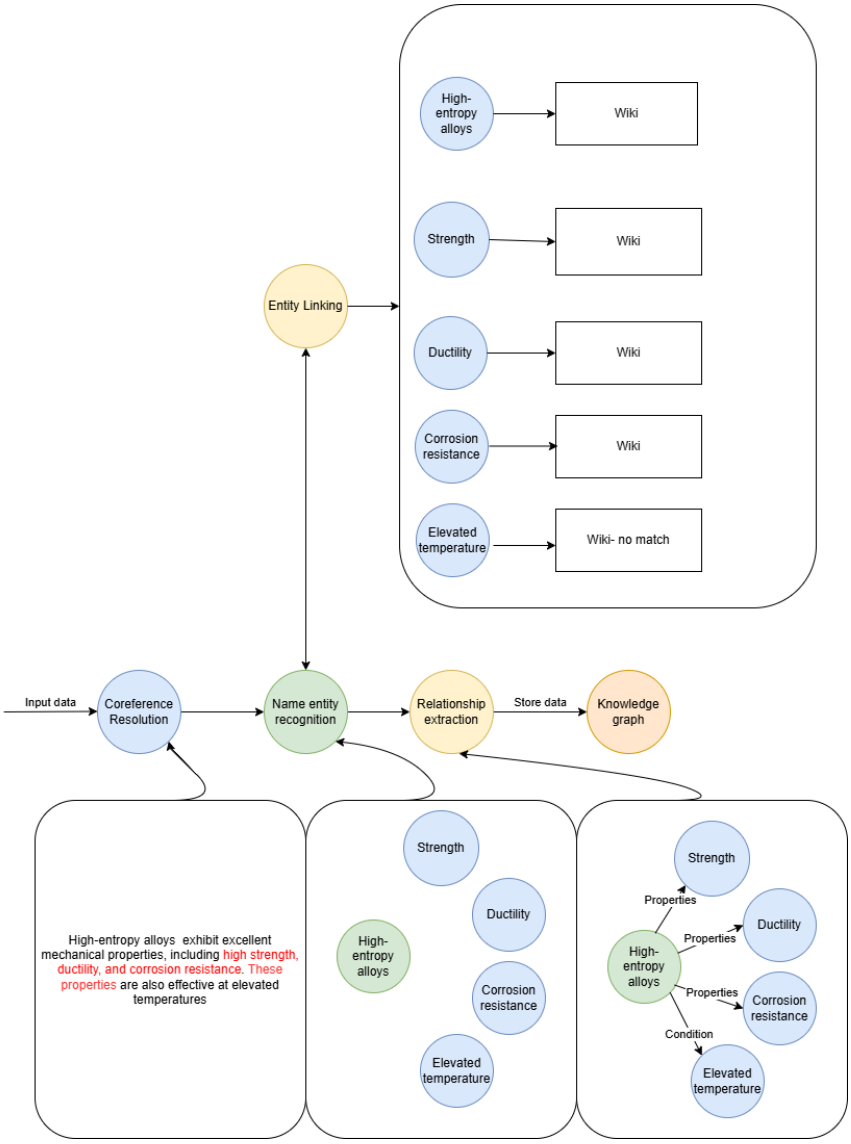

Figure 4: Entity Linking and Relation Extraction

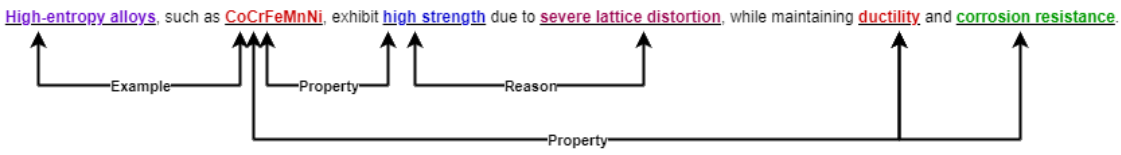

Figure 5: Entity Linking

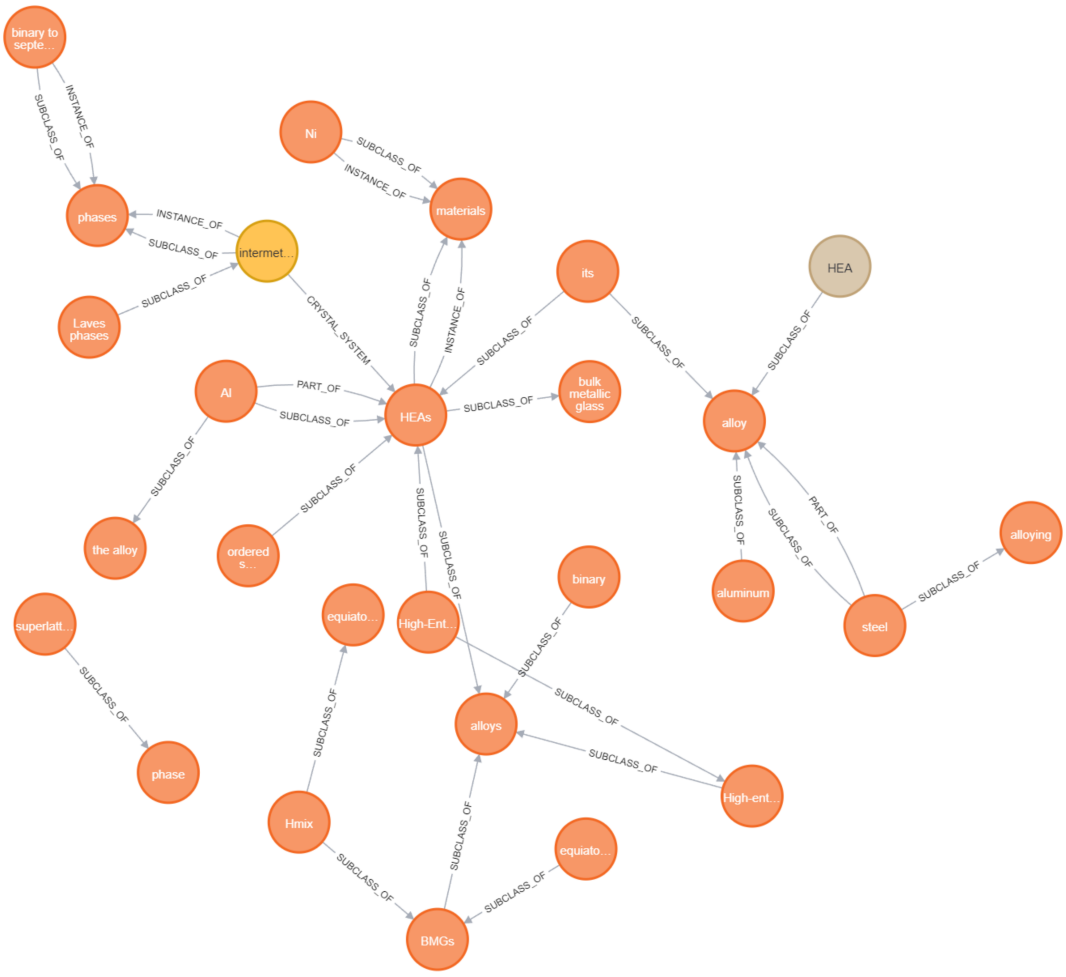

Figure 6: Relationship among Various High-entropy alloy Components

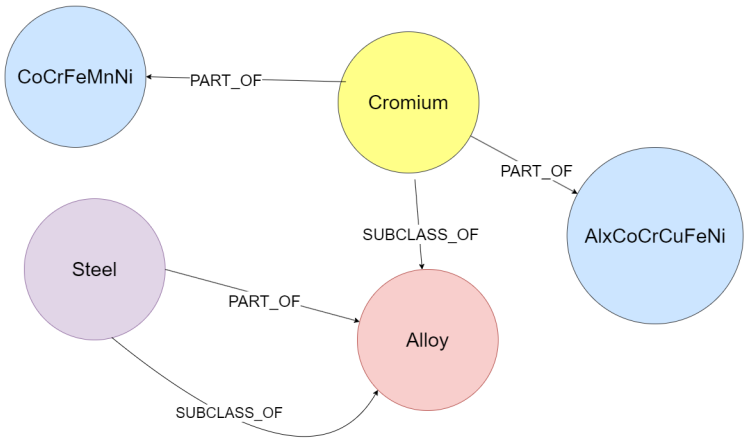

Figure 7: Another Relationship among Various High-entropy alloy Components

A.3 Examples from Our Dataset

In this section, we present sample queries from our dataset in Table 2, covering a range from simple to more complex queries.

Table 2: Example Queries

Query	**What is the yield strength of the CrMnFeCoNi alloy at 600 K, 700 K with 4 µm grain size?**
Ground Truth	The yield strength of the CrMnFeCoNi alloy at 600 K is 290 MPa, and at 700 K, it is 285 MPa.
Query	**What is the CRSS of CrMnFeCoNi at the tension in room temperature?**
Ground Truth	The Critical Resolved Shear Stress (CRSS) of the CrMnFeCoNi alloy has been measured at 53 MPa at room temperature and 175 MPa at 77 K.
Query	**What is the stacking fault energy of CrCoNi?**
Ground Truth	The stacking fault energy of CrCoNi is $18 - 26\,\mathrm{mJ/m^2}$.
Query	**At room temperature, what is the Hall-Petch slope of the cantor alloy?**
Ground Truth	At room temperature, the Hall-Petch slope of the cantor alloy was determined to be $494\,\mathrm{MPa}\,\mu\mathrm{m}^{-1/2}$.
Query	**What is the stacking fault energy of the cantor alloy?**
Ground Truth	The stacking fault energy of the cantor alloy was estimated to be $\sim 30\,\mathrm{mJ\,m^{-2}}$.
Query	**What is the yield strength and ultimate tensile strength of TiZrNbHfTa after 1000°C annealing?**
Ground Truth	After $1000\,^\circ\mathrm{C}$, the yield strength will be $1145\,\mathrm{MPa}$, and the ultimate tensile strength will be $1262\,\mathrm{MPa}$.
Query	**What is the CRSS of CrFeCoNiAl0.3 in compression at room temperature?**
Ground Truth	CRSS of CrFeCoNiAl0.3 in compression at room temperature is 54 MPa.

A.4 LLM RAG Evaluation Metrics

This section provides detailed descriptions of the various evaluation metrics used for RAG systems.

A.4.1 Correctness

Given a query q, a generated answer g, and an optional reference answer r, the `CorrectnessEvaluator` computes a score s using an LLM. This score is then compared against a threshold T to determine whether the generated answer is correct or passing.

$$\begin{aligned}
\text{Prompt} &\quad \text{Constructed from } q, g, r \\
E(g, q, r) &\quad \text{LLM Response to Prompt} \\
(s, \text{reasoning}) &\quad \text{parser_function}(E(g, q, r)) \\
\text{passing} &\quad s \geq T \\
\text{EvaluationResult} &\quad \{q, g, \text{passing}, s, \text{reasoning}\}
\end{aligned}$$

A.4.2 Faithfulness Evaluation

Given a query q, a response r, and a set of context documents C, the `FaithfulnessEvaluator` performs the following steps:

$$\begin{aligned}
\text{Context Documents} &\quad \text{Transform } C \text{ into Document objects} \\
\text{Index} &\quad \text{Create SummaryIndex from Document objects} \\
\text{Query Engine} &\quad \text{Create query engine using LLM, eval_template, and refine_template} \\
\text{Evaluation} &\quad \text{Perform a query on the response using the query engine} \\
\text{Raw Response} &\quad \text{Obtain raw_response_txt from the query engine} \\
\text{Passing} &\quad \begin{cases} \text{True} & \text{if yes is found in raw_response_txt} \\ \text{False} & \text{otherwise} \end{cases} \\
\text{Score} &\quad \begin{cases} 1.0 & \text{if passing is True} \\ 0.0 & \text{otherwise} \end{cases} \\
\text{Feedback} &\quad \text{raw_response_txt}
\end{aligned}$$

The evaluation result is given by:

$$\text{EvaluationResult} = \{q, r, C, \text{passing}, \text{score}, \text{feedback}\}$$

A.4.3 Answer Relevancy

Let q be the query, r the response, and $\{c_1, c_2, \ldots, c_n\}$ the contexts. *Define the following:*

$$\text{Documents} = \{d_i \mid d_i = \text{Document}(\text{text} = c_i) \text{ for } i = 1, 2, \ldots, n\}$$

$$\text{Index} = \text{SummaryIndex}(\text{Documents})$$

$$\text{query\_response} = \text{Question: } q \text{ Response: } r$$

Evaluate the query-response pair with:

$$\text{response\_obj} = \text{QueryEngine}(\text{Index}).\text{aquery}(\text{query\_response})$$

Let:

$$\text{raw\_response\_txt} = \text{str}(\text{response\_obj})$$

Then:

$$\text{passing} = \begin{cases} \text{True} & \text{if “yes” is in raw\_response\_txt.lower()} \\ \text{False} & \text{otherwise} \end{cases}$$

$$\text{score} = \begin{cases} 1.0 & \text{if passing} \\ 0.0 & \text{otherwise} \end{cases}$$

The output is:

$$\text{EvaluationResult} = \{q, r, \text{passing}, \text{score}, \text{feedback} = \text{raw\_response\_txt}, \text{contexts} = \{c_1, \ldots, c_n\}\}$$

A.4.4 Context Relevancy

Let q be the query, $\{c_1, c_2, \ldots, c_n\}$ be the contexts. Define:

$$\text{Documents} = \{d_i \mid d_i = \text{Document}(\text{text} = c_i)\}$$

$$\text{Index} = \text{SummaryIndex}(\text{Documents})$$

Evaluate the query q *using*:

$$\text{query\_engine} = \text{Index.as\_query\_engine}(\text{llm}, \text{eval\_template}, \text{refine\_template})$$

$$\text{response\_obj} = \text{query\_engine.aquery}(q)$$

Let:

$$\text{raw\_response\_txt} = \text{str}(\text{response\_obj})$$

Parse the result:

$$\text{score}, \text{reasoning} = \text{parser\_function}(\text{raw\_response\_txt})$$

Score threshold:

$$\text{score\_threshold} = 4.0$$

Calculate:

$$\text{score} = \frac{\text{score}}{\text{score\_threshold}}$$

Return:

$$\text{EvaluationResult} = \{q, \{c_1, \ldots, c_n\}, \text{score}, \text{feedback} = \text{raw\_response\_txt}, \text{invalid\_result}, \text{invalid\_reason}\}$$

A.4.5 Analysis of Variance (ANOVA)

ANOVA is a fundamental statistical method used to compare means across multiple groups to determine if there are statistically significant differences between them. This study utilizes a one-way ANOVA, which examines the effect of a single independent variable - in this case, the type of RAG pipeline - on a dependent variable (performance metrics). The mean score reflects the average performance of each method across all 10 queries, offering an overall assessment of its effectiveness for the given metrics. A mean score closer to the highest possible value suggests that the method consistently delivers superior results, indicating strong performance across various queries. Conversely, a lower mean score points to weaker overall performance, highlighting areas where the method may be less effective. Essentially, the mean score serves as a summary indicator of each method's typical efficacy, providing a clear comparison of their relative strengths in achieving the desired outcomes.

The F-statistic in ANOVA quantifies the ratio of variance between groups to variance within groups, with larger F-values indicating greater differences among the groups. The degrees of freedom (df) are denoted as F(2, 24) for correctness and F(2, 27) for faithfulness and relevancy, indicating the number of independent values that can vary in the calculation. Here, the first value (2) represents the degrees of freedom for the groups (number of groups minus 1), and the second value (24/27) represents the degrees of freedom for the error term (total observations minus the number of groups). The p-value assesses the probability of observing such differences between groups. Typically, a p-value below the significance level ($\alpha = 0.05$) suggests statistically significant differences between the groups.

A.5 Examples of Different Metrics

This section presents three example queries along with their corresponding responses from the RAG systems. Providing optimal responses to these queries requires effective information retrieval from text, figures, and tables. Additionally, we illustrate the application of various RAG evaluation metrics through these examples. Table 3 compares the performance using high-entropy alloy documents, demonstrating that G-RAG delivers contextually more relevant responses. The query in Table 4 necessitates accurate information extraction from Figure 8. The responses show that G-RAG with the parsing method outperforms other RAG systems, providing results that are closely aligned with the ground truth values. Finally, the query in Table 5 requires accurate value extraction from the table presented in Figure 9. In this case, G-RAG with the parsing method also outperforms other RAG systems. These results demonstrate that the integration of an efficient parsing method significantly enhances the performance of the G-RAG system.

Table 3: Comparison using **High-entropy alloy** Documents **(Context Relevancy)**

Query	**What is Chromium?**
G-RAG	Chromium is often a significant component in High-Entropy alloys (HEAs), particularly in the CoCrFeNi alloy, where it enhances corrosion resistance in both NaCl and H_2SO_4 solutions. However, when Chromium is replaced with Manganese in the Co-Cr-Fe-Ni alloy, the corrosion resistance in H_2SO_4 solution decreases, becoming inferior to 304 SS. The addition of certain elements like Al, Cu, B, and Mo can alter the corrosion resistance of various HEAs, often due to the formation of new phases or the impact on the existing phases' composition. The relationship between the degree of substitution (δ) and the enthalpy of mixing (H_{mix}) for intermetallic compounds is an area of interest, potentially implying that as composition increases, the enthalpy of mixing decreases, which could have implications for the thermodynamics of alloys and intermetallic compound formation.
Naive RAG	Chromium is a chemical element with the symbol Cr and atomic number 24. It is a hard, silver-white, and highly corrosion-resistant metal that is used in a variety of applications, including stainless steel, chrome plating, and catalytic converters in vehicles.
Graph RAG	Chromium is a chemical element with the symbol Cr and atomic number 24. It is a steely-grey, lustrous, hard, and brittle transition metal. Chromium is also the main additive in stainless steel, to which it adds anti-corrosive properties. It is highly valued as a metal that is able to be highly polished while resisting tarnishing. Chromium is also used in various industrial applications, including the production of stainless steel, and chrome plating, and as a corrosion-resistant coating.

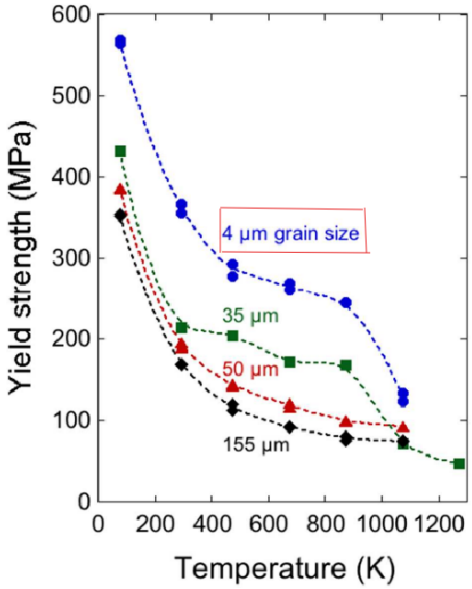

Figure 8: Information Extraction from Graph Images

Table 4: Comparison of Correct Information Extraction from Graph

Query	**What is the yield strength of the CrMnFeCoNi alloy at 600 K, 700 K with 4 μm grain size?**
Ground Truth	The yield strength of the CrMnFeCoNi alloy at 600 K is 290 MPa, and at 700 K, it is 285 MPa.
(G-RAG + **Parsing Method**)	The yield strength of CrMnFeCoNi alloy at 600 K is 310 MPa, and at 700 K is around 300 MPa.
Naive RAG	Unfortunately, the information provided does not contain any data about the yield strength of the CrMnFeCoNi alloy at 600 K, 700 K, or with a 4 μm grain size.
G-RAG	The yield strength of CrMnFeCoNi alloy at 600 K is 450 MPa, and at 700 K is 400 MPa.

Alloy	Loading	Temperature	CRSS (MPa)	Reference
CrMnFeCoNi	Tension	RT	53 - 60	[228]
CrMnFeCoNi	Tension	RT	63	[229]
CrMnFeCoNi	Tension	RT	70	[224]
CrMnFeCoNi	Tension	RT	78 - 82	[225]
CrMnFeCoNi	Tension	RT	82	[230]
CrMnFeCoNi	Compression[1]	RT	33 - 43	[154]
CrMnFeCoNi	Tension	77 K	135	[229]
CrMnFeCoNi	Tension	77 K	155	[230]
CrMnFeCoNi	Tension	77 K	145-172	[228]
CrMnFeCoNi	Tension	77 K	175	[224]
CrMnFeCoNi	Tension	203 K	103	[230]
CrMnFeCoNi	Tension	243 K	92	[230]

Figure 9: Information Extraction from Table Images

Table 5: Comparison of Correct Information Extraction from Table

Query	**What is the CRSS of CrMnFeCoNi at the tension in room temperature?**
Ground Truth	The Critical Resolved Shear Stress (CRSS) of the CrMnFeCoNi alloy has been measured at 53 MPa at room temperature and 175 MPa at 77 K.
(G-RAG + **Parsing Method**)	The CRSS of the CrMnFeCoNi alloy has been determined to be 53 MPa at room temperature and 175 MPa at 77 K.
Naive RAG	Unfortunately, the provided context does not contain information about the CRSS (Critical Resolved Shear Stress) of CrMnFeCoNi at tension at room temperature.
G-RAG	Sorry, I do not have any information about the CRSS of CrMnFeCoNi at the tension at room temperature.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims made in the abstract and introduction accurately reflet the paper's contirbution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work as well as future improvements that can be made to our work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided the full set of assumptions and a complete proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The authors of this paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have a proper codebase of everything we do to make it as reproducible as possible. But for reproducing it, if anyone wants to create a new knowledge base with the given documents, it is mandatory to have at least 24 GB VRAM and 60 GB System RAM.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have specified all the training and test details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have used Analysis of Variance to compare the means of our metrics system.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have added all types of computer workers and memory to better reproduce our experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute worker CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs and estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurRIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed how it will bring out the positive impact on the research of material science with the help of LLM.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no high risk of using our experiment.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have given proper credit.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets are well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not do our experiments on human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper doesn't have any potential risks incurred by study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.