SMELLM: A Paradigm to Integrating Domain Knowledge into LLMs via the Retrieval Augmented Generation

Anonymous ACL submission

Abstract

The utilization of large language models (LLMs) offers promising opportunities to expedite scientific discovery. However, deploying LLMs to answer scientific questions within specific interdisciplinary research domains, such as single-molecule electronics, poses various challenges that arise from the uniqueness of domain-specific data, the complexity of domain knowledge, and the uniqueness of domain objectives. To address this gap, we propose a paradigm for integrating domain knowledge from single-molecule electronics into LLMs using the retrieval-augmented generation (RAG) framework, named SMELLM. Evaluation results demonstrate that SMELLM achieves a higher SciBERT score than GPT and ChatGPT, with SMELLM-4.0 notably achieving a SciBERT score of 0.731 and a Faithfulness score of 0.916. The responses generated by SMELLM are firmly grounded in domain-specific facts, indicating significant enhancements in LLM capabilities for domain-specific natural language understanding tasks. Furthermore, SMELLM is adaptable for enhancing and evaluating proficiency in LLM across other scientific domains with low computing resource consumption.

1 Introduction

In the evolving landscape of artificial intelligence, the emergence of Large Language Models (LLMs) has profoundly impacted scientific research and analysis methodologies. These LLMs, capable of zero-shot and few-shot learning, have proven to be highly effective across a range of scientific activities, including literature analysis(Ray, 2023; Schmidt and Meir, 2023), unstructured data preprocessing(Zheng et al., 2023b), hypothesis generation(Zheng et al., 2023a), automated design of experiments(Bran et al., 2023; O'Donoghue et al., 2023), and data interpretation(Liu et al., 2023; Creswell et al., 2022). The adaptability of LLMs renders them invaluable assets for simplifying complex data analysis, extracting information across numerous disciplines, and enhancing decision-making processes. Their proficiency in understanding and generating text that resembles human communication has pioneered new avenues in knowledge discovery, driving significant progress in various scientific areas. Models such as ChatGPT(Brown et al., 2020), LLama(Touvron et al., 2023), Gemini(Google, 2023), and Claude(Models, Year) have been instrumental in enhancing the efficiency of scientific inquiry, promoting the accumulation of cross-disciplinary knowledge, and encouraging interdisciplinary collaboration. However, incorporating domain knowledge for LLMs presents significant challenges, including the accurate generation and comprehension of domain-specific content, the limitations of computing resources, and the lack of standardized evaluation methods(Koo et al., 2023; Chang et al., 2023; Ling et al., 2023; Guo et al., 2023). This gap poses difficulties for non-AI professionals, complicating the identification of current limitations, obstacles, unresolved issues, and potential areas for future research.

In this study, we concentrate on single molecular electronics (SME), an active and multidisciplinary field that intersects chemistry, physics, biology, materials science, and engineering. We introduce a pipeline for creating a high-precision dataset of SME literature that can be applied to other complex and highly specialized fields of scientific research, particularly those lacking publicly available data and demanding real-time updates. Based on the SME dataset, domain knowledge is integrated into the generation process of LLMs via the RAG framework, adopting a low-cost and computationally efficient approach known as SMELLM. The SMELLM architecture, supported by the Autogen(Wu et al., 2023) framework, is shown in Figure 1. Autogen provides a structured sequence of activities and interactions within an agent-based setting. Controlled by the ConversableAgent, the

UserAgent, and the AssistantAgent collaboratively finalize tasks through internal contextual conversations. The AssistantAgent discerns the necessary actions or responses to specific queries or situations by evaluating the current state and devising effective solutions with LLMs. Subsequently, the UserAgent utilizes tools to execute feedback. This iterative process continues until a termination condition is satisfied.

In the development of the SMELLM framework, we first establish a citation graph database named Graph_DB by employing Neo4j and integrating relevant literature metadata. Simultaneously, Pinecone_Vector_DB, a cloud-based database constructed on Pinecone, stores the original text post-segmentation along with their corresponding vectors. Furthermore, Keyword_Analysis, utilizing Graph_DB, facilitates keyword and subtopic analyses. External search resources such as Google, Google Scholar, and Wikipedia, in addition to tools like PubChem and ChemCrow, are utilized. Additionally, responses from these tools are systematically managed through an importance-ranking method, optimizing text length, and refining database queries with domain-specific keywords for both input and response sequences. Empowered by Autogen and facilitated by the OpenAI-powered API, SMELLM enables the systematic exploration of inquiries relevant to SME.

An example of SMELLM is presented in Figure 2. Our research showcases the implementation of an LLM retrieval enhancement framework within the SME domain, offering insights and guidelines for other fields to develop LLMs customized to their professional domains effectively and accurately.

In summary, we make the following contributions:

1. We devise a pipeline aimed at generating a domain knowledge dataset of high accuracy and its subsequent databases. We provide an explanation on how domain specialists can actively participate in the development of bespoke knowledge bases for the domain specification of LLMs.

2. We propose an automated approach, devoid of the need for experts, capable of generating Question-Answering (QA) test datasets. This methodology bears relevance to any highly specialized research field that does not possess a publicly available test dataset.

3. We illustrate the transformation of databases into beneficial resources for agents in order to optimize the information retrieval process within the RAG framework. A critical aspect to note here is that SMELLM does not necessitate the use of high-performance computing resources, thus making it an appropriate choice for smaller organizations and individuals.

2 Related Work

Hallucination(Zhang et al., 2023) in LLMs pertains to responses that are either factually inaccurate, irrelevant, or nonsensical. LLMs tend to produce less precise and relevant responses when dealing with domain knowledge that is under-represented in its training data or scenarios necessitating real-time information. Various strategies, including human feedback, relevance and accuracy inspections(Cao et al., 2020), and task-specific fine-tuning(Xia et al., 2024), have been utilized to mitigate hallucination and bolster the quality of responses. A recent study by Ling et al. (2023) categorized strategies for domain specialization in LLMs into three divisions: model fine-tuning, prompt crafting, and external augmentation. These divisions correspond to different assumptions about levels of accessibility.

Fine-tuning(Ovadia et al., 2023; Cheon and Ahn, 2022), while effective for tailoring LLMs for specified tasks and enhancing their performance on particular datasets, comes with inherent limitations related to computational resource usage and intricate technicalities. Even if a research organization secures the requisite hardware, the effectiveness is substantially reduced by the nuanced aspects of research domains, scarcity of specialized training data, and the cross-disciplinary complexities involved.

In-context Learning (ICL) facilitates learning from a handful of contextually-presented examples, with its effectiveness being contingent upon the quality and relevance of the provided examples. Chain-of-thought (COT)(Wei et al., 2023) endeavors to augment the reasoning capabilities of LLMs by integrating intermediate reasoning steps directly into the prompts. The efficiency of COT primarily resides in its step-by-step approach, proving notably successful in reasoning tasks, especially when combined with few-shot prompting. Furthermore, ExpertPrompting(Xu et al., 2023) employs ICL to autonomously generate thorough and indi-

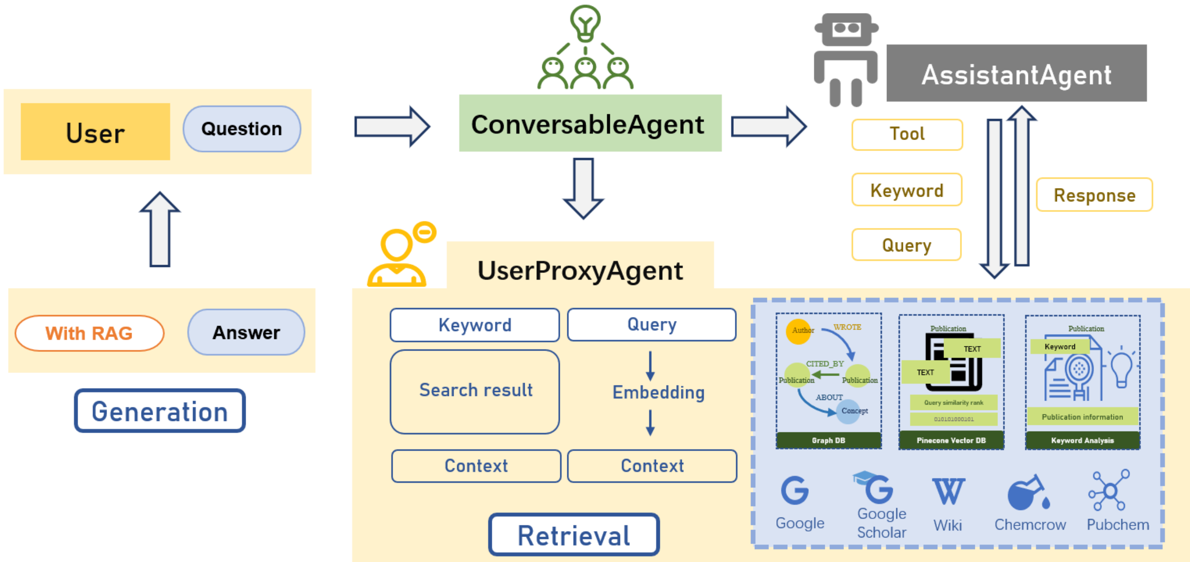

Figure 1: **Overview of SMELLM** The AssistantAgent is responsible for determining the selection of tools and the necessary parameters. The UserAgent is entrusted with executing the designated tools and providing responses within the established context. The ConversableAgent orchestrates the facilitation of communication and action planning.

vidualized descriptions of the expert identity for each specific instruction.

Incorporating retrieval augmented generation (RAG)(Gao et al., 2024; Lewis et al., 2021) into LLMs ensures access to the most recent and reliable factual information by cross-referencing the model's responses with original content, fostering trust in the information's accuracy. Significant examples include LangChain(Chase, 2022), Autogen(Wu et al., 2023), LlamaIndex(Liu, 2022), and MetaGPT(Hong et al., 2023). The inclusion of LLMs into these frameworks permits the creation of customizable, conversational agents designed for seamless collaboration with humans, tools, and other AI agents. This collective approach optimizes problem-solving strategies, increasing automation and efficiency, with LLMs serving diverse and essential roles in practical applications within scientific inquiry.

Several strategies have been implemented to incorporate domain knowledge into existing LLMs, primarily through the use of domain-specific contextual databases, existing APIs, and pre-trained models for handling collaborative domain tasks. Chemcrow(Bran et al., 2023) employed the Langchain framework to integrate 17 expert-designed tools, facilitating the resolution of chemistry-related tasks, particularly in molecule synthesis planning. Shen et al. (2023) devised a Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) framework tailored for biomedical tasks. Shen et al. (2023) introduced an LLM-powered agent leveraging various pre-trained models within machine learning communities to address AI tasks effectively. Yager (2023) provided LLMs with domain-specific contextual knowledge retrieval in physical science content and images. Balaguer et al. (2024) introduced a pipeline for fine-tuning LLMs to provide location-specific insights to farmers. However, these applications generally focus on open domains that require extensive literature, such as physics, chemistry, and biology, or demand fine-tuning to accommodate specific tasks, often allowing for lenient precision requirements for data retrieval.

3 Tasks Generation and Evaluation

Evaluating the performance of RAG frameworks typically involves curated datasets and predefined tasks customized for specific domain applications. In open-domain tasks, utilizing existing datasets enables the adoption of diverse strategies for developing or enhancing QA test datasets. These strategies include domain expert-crafted datasets(Otegi et al., 2022), transfer learning(Yue et al., 2022; Cheon and Ahn, 2022), template-based generation(Fabbri et al., 2020), and data augmentation(Song and Zhao, 2016). In domains of specificity and interdisciplinary scientific research, there is a notable shortage of test datasets for Natural Language Un-

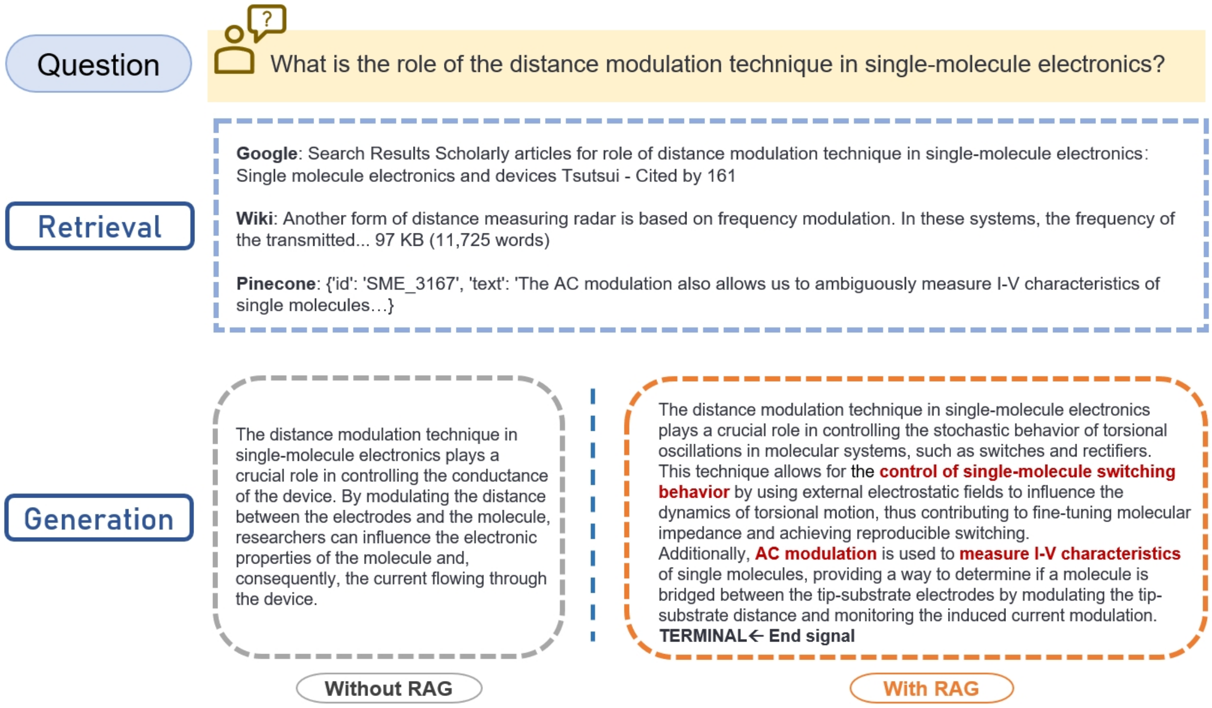

Figure 2: **An Example of SMELLM** In comparison to results without RAG integration, RAG generates responses containing relevant information that aligns with retrieval contexts.

derstanding tasks. Creating and evaluating test datasets manually persist as common practices, despite the significant time and effort required.

3.1 Tasks Generation

In this study, we utilize ChatGPT, augmented by review papers related to SME, to produce a specialized QA test dataset. Our methodology entails employing specific prompts to guide the format and content of questions, thus ensuring precision and reliability. Subsequently, we eliminated questions featuring ambiguous references or about particular experimental configurations. The QA tasks of SME comprises 50 meticulously curated questions. This methodology showcases adaptability across diverse research domains, providing an efficient approach to generating and refining domain-specific QA systems.

3.2 Evaluation

QA tasks often demand nuanced understanding of context, requiring retrieved information to closely align with the query's intent. The retriever component relies heavily on the efficacy of tools chosen for information retrieval and the precision of generated queries. The answer generation component plays a pivotal role in transforming retrieved information into coherent and accurate responses to user queries. Thus, the comprehensive evaluation of RAG frameworks includes two key components: Generation and Retrieval(Es et al., 2023; Chen et al., 2023). In this study, SMELLM is evaluated from the three aspects:

1. Accurate Tool Selection and Keyword or Cypher Generation;

2. Contextual Relevance and Accuracy of Retrieved Results;

3. Precision and Accuracy of Generated Answers.

To facilitate SMELLM in generating correct and complete matches for the appropriate node property names consistently in Graph_DB, we implemented a Cypher correction algorithm using similarity-based replacement. For Keyword_analysis, domain-specific keywords were utilized to refine search queries.

In the generation phase, we applied automatic metrics such as BLEU-1/4(Papineni et al., 2002), ROUGE-L(Lin, 2004), and sentence similarity based on SciBERT(Beltagy et al., 2019) (specialized for science) to evaluate the quality of generated answers by each model ROUGE-L is used for summarization, and BLEU, which relies on

N-gram co-occurrence, is used for machine translation. Furthermore, we use RAGAS(Es et al., 2023) as an automated evaluation framework that encompasses all tools employed in completing a QA task. Besides the metrics, ChatGPT-4 is deployed for comprehensive automatic evaluation, utilizing a prompt that facilitates the comparison and contrast of responses.

4 Result and discussion

4.1 Retrieval

In the exploration of domain-specific QA capabilities, SMELLM has demonstrated proficiency in identifying critical aspects of queries. In the preliminary experiment, SMELLM-3.5-pre revealed a high context recall but a low context precision (Context Precision = 0.305). This suggests that although the search process retrieved valuable information, it often included extraneous, redundant data, thus impeding the information generation process. To mitigate this problem, we employed compression and importance-ranking techniques to refine the search outcomes, as shown in Figure 3. The application of importance-ranking, with a retention rate set at 0.4, enabled us to improve precision while maintaining a high recall rate. This strategy, utilized during the testing phase of both SMELLM-3.5 and SMELLM-4.0, allowed us to partially overcome the limitations imposed by context length and ensure that only the most relevant and salient information is considered during text generation.

In the retrieval phase, as shown in Figure 4, we underscore the primary and auxiliary tools employed by SMELLM-3.5 and 4.0. The information retrieval process of SMELLM relies heavily on Pinecone_Vector_DB, whose utilization has increased from 38 instances in SMELLM-3.5 to 51 in SMELLM-4.0. This upward trend indicates its growing importance within the framework, especially with the integration of more advanced LLM capabilities. Furthermore, Google Scholar and Wiki contribute significantly to academic and general knowledge, respectively. The marked discrepancy in the frequency of calls to these tools between SMELLM-3.5 and SMELLM-4.0 suggests a notable shift in the framework's reliance on various knowledge sources. This data not only illuminates the intricacies of SMELLM's tool dependencies but also underscores the pivotal role of Pinecone_Vector_DB in advancing the retrieval process for completing domain-specific QA tasks.

The effectiveness of LLMs in QA tasks is heavily influenced by the comprehensiveness and depth of their training data, which in turn impacts their ability to understand various domains and employ appropriate methodologies for generating precise responses. The result of QA tasks has shown that integrating Pinecone_Vector_DB into SMELLM-4.0, along with the SME dataset and GPT-4.0 training data, significantly enhances domain-specific knowledge coverage. However, the training data for the GPT-3.5 model might not fully cover the range of queries, and relying solely on Pinecone_Vector_DB for effective QA tasks completion seems insufficient. Therefore, it is imperative to incorporate supplementary external search tools. Advancements in model training, meticulous data curation, and the integration of additional knowledge sources are crucial for improving the state-of-the-art in delivering accurate and comprehensive responses.

The trade-offs among faithfulness, context recall, and precision of SMELLM-3.5 and SMELLM-4.0 are presented in Table 1. SMELLM-4.0 demonstrates a slightly higher faithfulness value of 0.916 compared to SMELLM-3.5, which stands at 0.865. This suggests that its responses rely more on retrieved information with fewer instances of hallucination. SMELLM-3.5 achieves a context recall value of 0.817, while SMELLM-4.0 performs slightly lower at 0.731. This implies that while SMELLM-4.0 improves in faithfulness, it struggles to recall context from the input, potentially overlooking relevant details due to the singularity of its information sources. The context precision scores indicate the value of importance-ranking, particularly in managing multiple information sources.

LLM	FF	CR	CP
SMELLM-3.5	0.865	**0.817**	**0.387**
SMELLM-4.0	**0.916**	0.731	0.330

Table 1: **Retrieval metrics of SMELLM** Test SMELLM in QA tasks with a retention rate of 0.4.Faithfulness (FF) measures the fidelity of the answer to the context, preventing hallucinations. Context Precision (CP) indicates the relevance of the retrieved context to the question. Context Recall (CR) indicates the ability to retrieve all necessary information for answering the question.

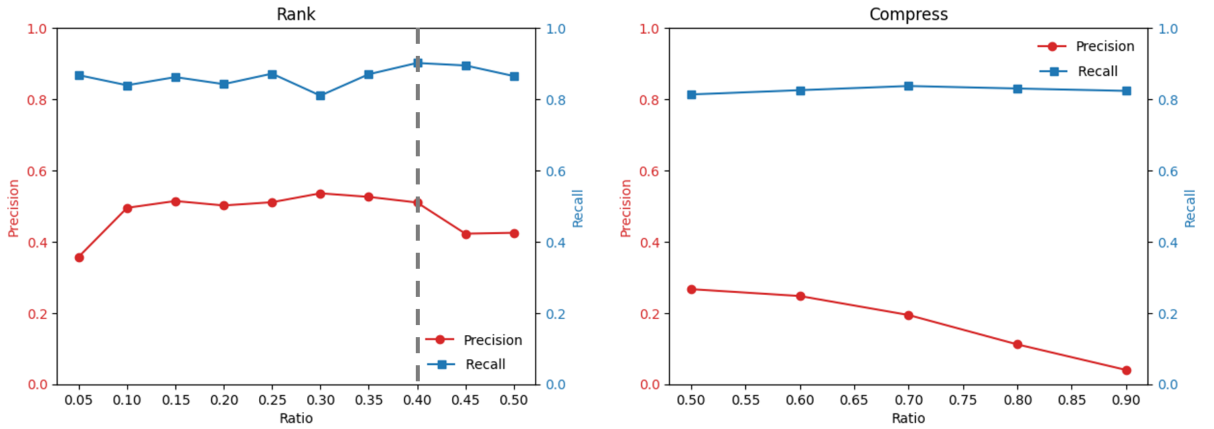

Figure 3: **Compression and importance-ranking in SMELLM-3.5-pre** The red curve represents precision across different thresholds, while the blue curve represents recall.

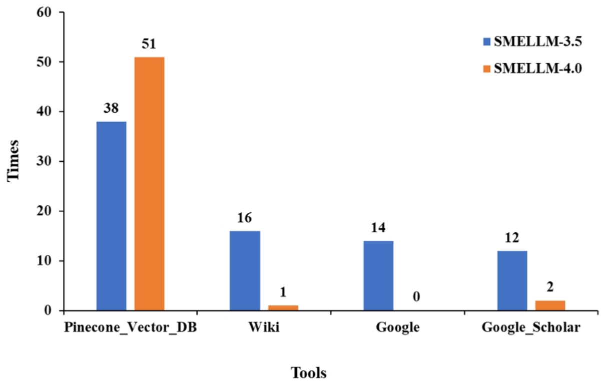

Figure 4: **The tool utilization of SMELLM-3.5 (blue) and SMELLM-4.0 (orange) in the QA task**

4.2 Generation

In the generation phase shown in Table 2, both SMELLM-3.5 and SMELLM-4.0 with higher SCIBERT scores than GPT and ChatGPT, suggesting that the generated answers closely resemble the scientific semantics of the ground truth. Additionally, SMELLM-4.0 achieves the highest BLEU scores, with BLEU-1 at 0.1 and BLEU-4 at 0.018, demonstrating its capacity to accurately capture domain-specific keywords and phrases. Notably, SMELLM-3.5 also attains the highest ROUGE F-score of 0.179, reinforcing its proficiency in maintaining semantic similarity and content overlap with the ground truth.

The responses generated by ChatGPT exhibit user-friendly traits, albeit occasionally being verbose. While this verbosity aids comprehension and fosters conversation, it may inadvertently introduce redundancy, compromising the conciseness valued by similarity metrics and ROUGE-L precision. Conversely, GPT tends to produce more concise outputs, though it may sometimes repeat input context. While GPT aims to highlight key concepts of the question with a Answer Relevance score of 0.950, it risks reducing content diversity and accuracy, as assessed by Answer Similarity and SciBERT score. However, while verbosity and conciseness are often presented as opposites, the ultimate effectiveness of a language model depends on its ability to tailor its output to the context of the prompt and the user's expectations.

SMELLM aims to achieve balance by integrating system-defined prompts with information retrieved from processes. This approach enables the generation of responses closely aligned with user queries, displaying greater similarity to ground truth and achieving higher BLEU and ROUGE-L scores. Consequently, SMELLM responses tend to reflect expert terminology and adhere to the context expected in scientific content.

Specifically, SMELLM 4.0 effectively addresses key points while maintaining its superior ability to generate informative, nuanced, and academically valuable content, as shown in Figure A1. This comparative analysis not only highlights the incremental advancements in AI-generated content but also underscores the pivotal role these technologies play in disseminating and interpreting scientific knowledge.

5 Dataset

Establishing real-time and highly accurate domain datasets is the cornerstone of the entire work. In this section, we present a comprehensive overview of the systematic pipeline for acquiring and integrating the SME literature dataset into databases,

LLM	SciBert	Ragas		BLEU		ROUGE		
		AR	AS	1	4	R	P	F
SMELLM-3.5	**0.712**	**0.945**	**0.865**	0.078	0.014	0.457	0.115	**0.179**
SMELLM-4.0	**0.731**	0.891	0.822	**0.100**	**0.018**	0.453	0.091	0.149
GPT-3.5	0.680	**0.950**	0.823	0.093	0.016	0.392	0.113	0.170
GPT-4.0	0.708	0.870	0.819	0.075	0.013	0.393	0.097	0.145
ChatGPT-3.5	0.604	0.904	0.831	0.079	0.014	0.499	0.071	0.123
ChatGPT-4.0	0.638	0.896	0.834	0.078	0.014	0.499	0.068	0.118

Table 2: **Generation Metrics for GPT, ChatGPT, and SMELLM.** The GPT API, developed by OpenAI, serves as the interface for the GPT Model. The SciBERT Score quantifies the similarity between the ground truth and the answers generated by LLM. Answer Relevance (AR) and Answer Similarity (AS), are derived from Raga's automated evaluation. BLEU-1 and BLEU-4 are the precision of unigrams and 4-grams, respectively. Additionally, R (ROUGE-L recall), P (ROUGE-L precision), and F (ROUGE-L F-score) are overlap performance metrics.

as shown in Figure 5, domain experts provide initial domain-specific keywords and topics, and the subsequent data collection process and evaluation is automated. Notably, researchers can adapt the entire pipeline to any other field within a few weeks. Leveraging this dataset, a domain-specific chatbot could be developed with minimal resource consumption and expense.

5.1 Data Retrieval and Extraction

The data retrieval process commenced by querying relevant academic literature from the Web of Science (WoS) database using specific keywords outlined in Table S1. We exported the results as HTML files and extracted bibliographic details. Given the interdisciplinary nature of SME, the initial search yielded a significant volume of unrelated literature. To address this issue, we constructed a Neo4j graph database based on the citation information, encompassing the writing relationship between authors and literature, forming the basis for filtering.

5.2 Preliminary Publication Filtering

We employed the term "junction" as a strongly associated term to select 7,000 identified SME literature, along with an additional set of 20,000 associated references as the training dataset. Following this, we fine-tuned a BERT model using the training dataset as a classifier to identify SME literature. While not infallible, this model effectively eliminates unrelated literature based on their titles.

5.3 Integration of Influential Research Groups' Information

We pinpointed authors with a significant number of "WROTE" relationships in the Neo4j database. These authors often belong to influential research groups with websites containing accurate publication records. We collected publication details from these websites using web crawlers or copied and processed by ChatGPT into a table format, employing a fuzzy match technique to capture and integrate any potentially missed literature.

5.4 Evaluation and Metadata Enrichment

Within the Neo4j database, the relevance of each SME paper was assessed by its local citation count, achieving a node connection proportion of 0.948. Building upon this foundational information and following methodologies outlined by (Koneru et al., 2022), missing metadata were supplemented from Semantic Scholar[1], CrossRef[2], and Google Scholar[3] to enhance citation details and abstracts. As of the most recent update, the SME database encompasses 5,715 SME literature, accompanied by 81,719 references. Future updates necessitate the utilization of the same retrieval strategy in WoS, sorting by date to integrate the latest literature and selecting literature surpassing the local citation threshold in the Neo4j Database.

6 Conclusion

In this work, we propose a simple and effective paradigm for integrating domain knowledge into LLMs via the RAG framework. The computational resource requirements for implementing and maintaining the proposed SMELLM framework are crucial considerations that require further discussion. We employ the OpenAI's API and the Pinecone cloud vector database to streamline the RAG process, primarily relying on CPU resources for model

[1]https://www.semanticscholar.org
[2]https://www.crossref.org/
[3]https://scholar.google.com/

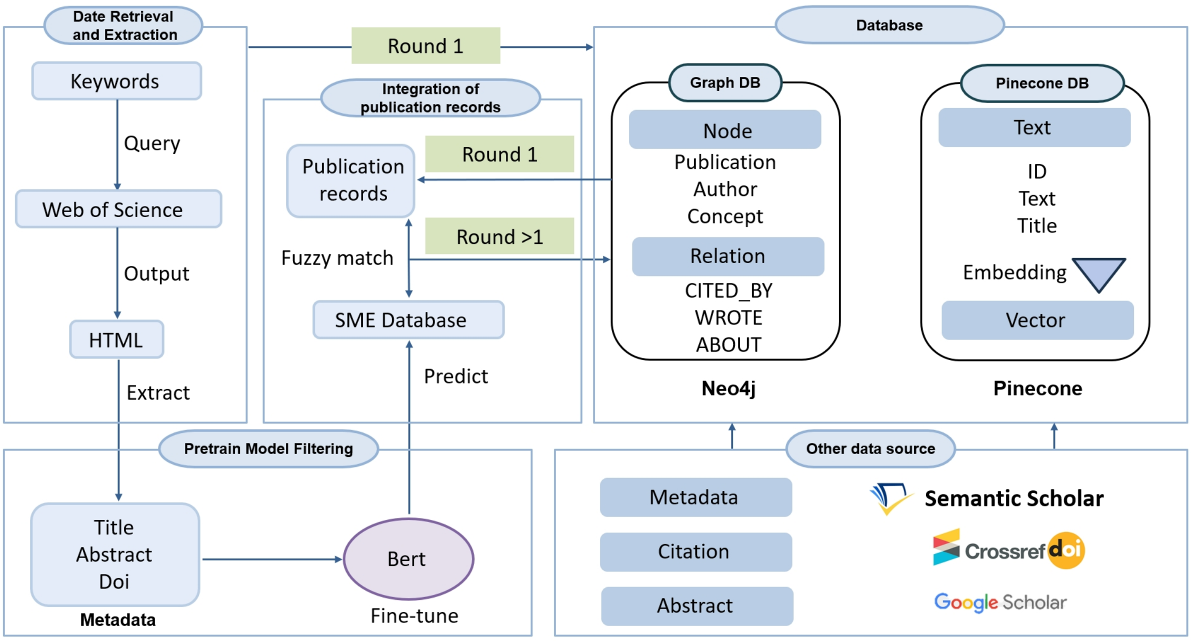

Figure 5: **The Process of Constructing a High-Accuracy Domain Literature Dataset.** The construction of a precise domain-specific dataset entails ongoing refinement and filtering. In ROUND 1, influential authors and SME literature are identified through relational metrics within the Neo4j database, which is utilized for subsequent publication record matching and pre-trained model fine-tuning.

execution. We have strategically tailored the framework to operate efficiently within such constraints, such as setting a minimum string length during vector construction and using importance-ranking. We meticulously curated SME dataset with substantial domain relevance, employing it to establish both graph and knowledge vector databases. Our approach offers practical solutions for researchers challenged by restricted computational resources and insufficient technical accumulation, significantly facilitating the customization of domain-specific QA systems.

7 Limitation

Despite significant advancements, more efforts will continue to be dedicated to future research due to some limitations. Our attention has been primarily directed towards domain-specific QA tasks, underpinned by the need to confront the subtleties within specialized knowledge domains. Tailored LLMs necessitate comprehensive input from subject experts to guarantee the accuracy and integrity of multifaceted domain data. Moreover, our dependency on existing databases and tools underscores the essentiality of consistent updates and maintenance to ensure the precision and relevance of integrated domain knowledge. Future research endeavors could explore optimizations to make the framework more resource-efficient and adaptable to diverse applications, thus alleviating potential barriers to wider adoption and allowing LLMs to independently invoke, execute, and process algorithms relevant to the prediction and analysis of experimental data.

8 Code and Data Availability

The code and data will be made publicly available upon publication.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 22325303, and 22250003). We also thank the support from the experts in SME for their contributions to this work.

References

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, et al. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv e-prints*, pages arXiv–2401.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific

Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3613–3618, Hong Kong, China. Association for Computational Linguistics.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. ArXiv:2304.05376 [physics, stat].

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. ArXiv:2307.03109 [cs].

H. Chase. 2022. Langchain. https://github.com/hwchase17/langchain.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. ArXiv:2309.01431 [cs].

Saeyeon Cheon and Insung Ahn. 2022. Fine-Tuning BERT for Question and Answering Using PubMed Abstract Dataset. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 681–684, Chiang Mai, Thailand. IEEE.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. ArXiv:2205.09712 [cs].

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated Evaluation of Retrieval Augmented Generation. ArXiv:2309.15217 [cs].

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv:2312.10997 [cs].

Google. 2023. Gemini: A family of highly capable multimodal models.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. ArXiv:2310.19736 [cs].

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Sai Dileep Koneru, David Rench McCauley, Michael C. Smith, David Guarrera, Jenn Robinson, and Sarah Rajtmajer. 2022. The evolution of scientific literature as metastable knowledge states. ArXiv:2202.12913 [cs].

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking Cognitive Biases in Large Language Models as Evaluators. ArXiv:2309.17012 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. 2023. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. ArXiv:2305.18703 [cs].

Jerry Liu. 2022. LlamaIndex.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2023. ChatGPT-powered Conversational Drug Editing Using Retrieval and Domain Feedback. ArXiv:2305.18090 [cs, q-bio].

Claude Models. Year. Model card and evaluations for claude models.

Arantxa Otegi, Iñaki San Vicente, Xabier Saralegi, Anselmo Peñas, Borja Lozano, and Eneko Agirre. 2022. Information retrieval and question answering: A case study on COVID-19 scientific literature. *Knowledge-Based Systems*, 240:108072.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

Odhran O'Donoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Ghareeb, and Samuel Rodriques. 2023. BioPlanner: Automatic Evaluation of LLMs on Protocol Planning in Biology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2676–2694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Paul G. Schmidt and Amnon J. Meir. 2023. Using Generative AI for Literature Searches and Scholarly Writing: Is the Integrity of the Scientific Discourse in Jeopardy? ArXiv:2311.06981 [cs, math].

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Linfeng Song and Lin Zhao. 2016. Domain-specific Question Generation from a Knowledge Base. ArXiv:1610.03807 [cs] version: 1.

Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Cite arxiv:2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. ArXiv:2308.08155 [cs].

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning. ArXiv:2402.04333 [cs].

Benfeng Xu, An Yang, Junyang Lin, Quang Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *ArXiv preprint arXiv:2305.14688*, 2.

Kevin G Yager. 2023. Domain-specific chatbots for science using embeddings. *Digital Discovery*, 2(6):1850–1861.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Domain Adaptation for Question Answering via Question Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1776–1790, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T.N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2023a. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*.

Zhiling Zheng, Zichao Rong, Nakul Rampal, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. A GPT-4 Reticular Chemist for Guiding MOF Discovery.

Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. 2023b. Chatgpt chemistry assistant for text mining and prediction of mof synthesis. *arXiv preprint arXiv:2306.11296*.

A1 Tools

Graph_DB, constructed on Neo4j [4], integrates three types of nodes—Authors, Publications, and Concepts—complemented by three relationship categories—CITED_BY, WROTE, and ABOUT. This graph database delineates both the collaborative writing dynamics among authors and the citation networks within the SME.

Pinecone_Vector_DB is established on Pinecone [5], a cloud-based repository for text vectors. PDFs are obtained via bulk download, followed by paragraph extraction through CerMine(Tkaczyk et al., 2015). Subsequently, we employ the hkunlp/instructor-xl to generate text embeddings, facilitating the retrieval of original text data through similarity searches with the stored vectors in Pinecone.

Our methodology incorporates Google Scholar for targeted literature searches using specific keywords, yielding search outcomes accompanied by concise summaries. Google is utilized to search for relevant results from various sources across the web, offering up-to-date information. Wiki stands as an accessible online encyclopedia containing an extensive repository of definitions. PubChem provides information on chemical structures, properties, and biological activities.

A2 Domain Expertise

In this study, whether for utilization within SME or for transfer to other specific domains, the contribution of domain experts is essential for gathering fundamental domain-related information that enhances the precision of the retrieval process. This encompasses providing domain-specific keywords, identifying top journals, defining research topics, and recognizing highly influential research groups.

A2.1 Domain-specific keywords

During the literature collection process, it is crucial to distinguish between similar yet distinct domains. The suggested keywords stem from the specificity of testing techniques (e.g., break junction) and prevalent applications (e.g., device). SME experts recommend the following set of keywords for effectively retrieving literature within the Web of Science database. Keywords[6] Shown in Table A1 and Table A2

Keyword
molecular electronics
break junction
molecular conductance
conductance histogram
graphene electrode
metal-molecule-metal junction
single molecular transport
single-molecular device
single molecular magnet
single molecular switch
single molecular transistor
single molecular memristor
single molecular diode
single molecular sensor
single molecular trace

Table A1: **Domain specific keywords of SME.**

KeywordPair_1	KeywordPair_2
junction	molecular
conductance	molecular
gold	junction
electronic	single molecular
tunnel	single molecular
device	single molecular
charge transport	single molecular
conductance	trace
au	junction
molecular-scale	charge

Table A2: **Domain specific keyword pairs of SME.**

A2.2 Top journal

Employing a fuzzy matching approach, we assessed the representation of literature from top journals within the SME domain. This quantitative analysis serves as a qualitative benchmark to validate the accuracy of journal data. Adopting this method ensures the credibility of the chosen journals and their alignment with the subject domain.

the Top journal is including: Nature, Nature Materials, Nature Nanotechnology, Nature Chemistry, Nature Reviews Chemistry, Nature Reviews Materials, Nature Reviews Physics, Nature Communications, Science, Science Advances, Journal of the American Chemical Society, Angewandte Chemie-International Edition, Proceedings of the National Academy of Sciences of the United States of Amer-

[4]https://neo4j.com
[5]https://www.pinecone.io
[6]single molecular can be replaced by single molecule and single-molecule

ica, Chemical Reviews, Chemical Society Reviews, Chem, Matter, Accounts of Chemical Research, ACS Nano, Advanced Materials, Nano Letters.

A2.3 Domain research topics

Here are the academic explanations of the main topics of SME and their associated keywords:

Thermoelectricity and Thermal Conductance: Therm, Seebeck, ZT (Figure of Merit for thermoelectric materials), Heat, Phonon. Mechanical Manipulation: Mechanics, Force, AFM (Atomic Force Microscopy), Deformation, Elasticity, Stiffness. Spin Dynamics: Spin, Kondo Effect, Radical, Magnetism, Spin Crossover, Giant Spin. Optoelectronics: Opto, Light, Photonic, Plasmon, Vibration, Exciton. Electric-field Control: Electric Field, Electrostatics. Structure and Reaction Analysis: Bridging, Linker, Backbone, Anchor, Skeleton, Synthesis, Symmetric, Conjugation, Geometry, Configuration, Length, Aromaticity, Covalent Bonds, Ligand, Complexation, Supramolecular, Supercritical, Catalysis, Reaction Pathways. Data Mining and Analysis: Clustering, Classification, Cluster Analysis, Classification, Supervised Learning, Data Analysis, Algorithms, Models. Theoretical Calculations and Simulations: Density Functional Theory (DFT), Orbital, First Principles, Non-equilibrium Green's Function (NEGF), Calculation, Simulation, Modeling.

A2.4 Research groups

Influential research groups can be identified through the preliminary selection of SME papers using the Step of 5.2. This involves calculating the count of "WROTE" relationships for each author. Subsequently, domain experts are engaged to validate and refine the list by adding or removing authors as necessary.

A3 Metrics

A3.1 BLEU

BLEU (Bilingual Evaluation Understudy)(Papineni et al., 2002) evaluates the quality of machine-translated text by comparing it to one or more reference translations. BLEU-1 and BLEU-4 are variants of BLEU that focus on comparing unigrams (individual words) and 4-grams (sequences of 4 words) respectively.

$$\text{BLEU-1} = \text{BP} \times \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log p_n\right)$$

N is the maximum n-gram order considered. p_n is the precision of n-grams.

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \le r \end{cases}$$

is the Brevity Penalty to penalize short translations, c is the length of the candidate translation, and r is the effective reference length (closest length of the reference to the candidate)

A3.2 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)(Lin, 2004) evaluates the quality of summaries generated by automatic summarization systems. ROUGE-L (Longest Common Subsequence) is one of the ROUGE metrics, which focuses on the longest common subsequence between the generated summary and the reference summary.

$$\text{ROUGE-L} = \frac{\text{LCS}(X, Y)}{\text{R}}$$

$\text{LCS}(X, Y)$ is the length of the longest common subsequence of words between the candidate summary X and the reference summary Y, and R is the total number of words in the reference summary.

A3.3 Ragas

Ragas (Retrieval Augmented Generation Assessment)(Es et al., 2023) is a framework for evaluating Retrieval Augmented Generation (RAG) pipelines. Specifically, Ragas including following Metrics:

Faithfulness is the factual consistency of the generated answer with the retrieved context. A generated answer is considered faithful when all claims it presents can be inferred from the provided context.

$$\text{Faithfulness} = \frac{\text{Could be Inferred from context}}{\text{Total of claims}}$$

Answer Relevancy is the relevance of generated answers to the questions. This metric disregards the factual accuracy of answers and penalizes incompleteness or redundant information.

$$\text{Answer Relevancy} = \frac{1}{N}\sum_{i=1}^{N}\text{sim}(Q_i, Q_{\text{orig}})$$

Where N is the LLM generates questions for the generated answer. Q_i is the i-th generated question. Q_{orig} denotes the original question. $\text{sim}(Q_i, Q_{\text{orig}})$

represents the cosine similarity between Q_i and Q_{orig}.

Context Precision evaluates whether all chunks in retrieved contexts relevant to the question are ranked higher in items, and refers to the ratio of chunks that are helpful in answering a question evaluated by LLM of all chunks.

$$\text{Context Precision} = \frac{\sum \text{P@}k}{\text{relevant chunks in the top } K}$$

$$\text{P@}k = \frac{\text{true P@}k}{\text{true P@}k + \text{false P@}k}$$

Context Relevancy is similar with Context Precision, and calculated based on the question and contexts in sentence level, S is the number of helpful sentences.

$$\text{Relevancy} = \frac{S}{\text{sentences in retrieved context}}$$

Context Recall is calculated based on the ground truth and the retrieved context

$$\text{Recall} = \frac{\text{GT sentences that can be attributed}}{\text{Number of sentences in GT}}$$

Answer Semantic Similarity evaluates the quality of generated answer by using a cross-encoder model to calculated with ground truth. In this work, we use sentence-transformers/all-MiniLM-L6-v2

Besides, Ragas include metrics such as Answer Correctness and Aspect Critiques, which are not used in this study, that can be refereed for detailed information on Metrics.

A4 Prompt

A4.1 QA Generation

> Construct a dataset for domain-specific QA tasks. Each entry in the dataset should be a python list of dictionaries with 'question' and 'answer' fields. Only keep the list.
> {Context}
> Consider the following key points: The answers should provide clear and brief information. Ensure that the questions cover various aspects such as definitions, motivations, challenges, solutions, techniques, and notable contributions in the field.

A4.2 QA Filtering

> Your task is to review each question and answer pair, filtering out any entries that:
> 1.contain ambiguous references
> 2.delve excessively into specific experimental setups
> 3.ask very simple concepts, like What is single-molecule electronics?
> Please prioritize clarity, relevance, and generalizability in your filtering process

A4.3 ChatGPT-4 Evaluator

> Compare and contrast the responses from large language models to the following question: Question Evaluate the accuracy, depth, and coherence of each model's response. Consider the scientific accuracy, clarity, and relevance of information presented by each model in addressing the question. Additionally, highlight any notable differences or unique insights offered by each model in their responses.
> Answers
> Finally, summarize which model provides the better answer, mentioning only the model's name.

A5 Comparison Example

In consideration of accuracy, depth, and coherence, SMELLM-4.0 emerges as the superior choice, offering a detailed, scientifically accurate, and comprehensive overview of SME domain-specific tasks (see A1).

Question: What are the potential advantages of carbon electrode-molecule junctions over ordinary metal electrode-based molecular junctions?

Answer of SMELLM-3.5:
The potential advantages of carbon electrode-molecule junctions over ordinary metal electrode-based molecular junctions include the following:
1. Chemical Inertness: Carbon electrodes are chemically inert, making them particularly advantageous in biomolecular sensing applications or fast DNA sequencing where specific or nonspecific binding of biomolecules to metal electrodes is not desirable.
2. Anchor Possibilities: Carbon electrodes allow for a large variety of possibilities to anchor diverse molecules as compared to metallic electrodes, providing more flexibility in molecular junction construction.
3. Stability and Compatibility: Carbon-based electronic devices offer enhanced stability due to strong covalent bonding in the disordered carbon materials, resisting electromigration or penetration into the molecular layer. The greater compatibility between the carbon electrode and molecule leads to better construction of the molecular junction.
4. Contact Flexibility: Carbon-based electrodes provide contact flexibility, allowing for covalent bonding or – stacking configurations, providing more versatile options for forming molecular junctions.
5. Stability of Carbon-Molecule Interface: The interface between carbon electrodes and molecules is much stable than that of metal-thiol interface, leading to highly reproducible, robust, and high-yield device fabrication.
These advantages make carbon electrode-molecule junctions an appealing platform for molecular electronics, offering stability, compatibility, and diverse anchoring possibilities for a wide range of applications.
TERMINATE

Answer of GPT-3.5:
1. Enhanced stability: Carbon electrodes are typically more stable and have a higher chemical resistance compared to metal electrodes, which can lead to longer-lasting and more reliable molecular junctions.
2. Reduced reactivity: Carbon electrodes are generally less reactive compared to metal electrodes, which can help prevent unwanted side reactions or degradation of the molecules being studied in the junction.
3. Tunable properties: Carbon electrodes can be functionalized or modified to tailor their properties and create specific interactions with molecules, allowing for better control and manipulation of the junction.
4. Widely available and cost-effective: Carbon electrodes are relatively inexpensive and readily available, making them a more cost-effective option for creating molecular junctions compared to specialized metal electrodes.
5. Compatibility with a wide range of molecules: Carbon electrodes are versatile and can be used with a wide variety of molecules, allowing for more flexibility in the types of studies that can be conducted with carbon electrode-molecule junctions.

Answer of SMELLM-4.0:
Carbon electrode-molecule junctions have numerous potential advantages over ordinary metal electrode-based molecular junctions.
1. Chemical Inertness: Unlike metal electrodes, carbon electrodes are chemically inert. This property makes carbon-based electronic devices more suitable for applications such as biomolecular sensing or fast DNA sequencing where unwanted binding of biomolecules on metal electrodes is undesirable.
2. Enhanced Stability: The strong covalent bonding in disordered carbon materials bolsters its resistance against electromigration or penetration into the molecular layer, providing enhanced stability.
3. Compatibility: Most molecules are carbon-based organic compounds, thereby creating better compatibility with carbon electrodes, leading to improved construction of molecular junctions.
4. Durability and Operational Stability: Carbon-based molecular junctions have shown commendable durabilities, thermal and operational stabilities, and longer device lifetimes.
5. Versatility: Carbon electrodes offer a wide range of anchoring possibilities for diverse molecules, making it more versatile than metal electrodes.
6. Reduced Defects: Compared to metallic electrodes, carbon-based electrodes may allow a reduction of the effects of defects in the electrodes through the use of extended planar anchor groups.
7. Flexibility: Carbon electrodes tend to be more flexible in contact - e.g., in a covalent bonding or pi–pi stacking configuration.
8. Microfabrication Compatibility: Some forms of carbon electrodes, like Silicon, can be compatible with most semiconductor fabrication procedures.
9. Resistance to Oxidation: Carbon electrodes are resistant to oxidation, crucial when high bias is required.
10. Simplicity and High Yield: The fabrication process of carbon-electrode based junctions is simpler and affords stable molecular devices in high yield, making them potentially more practical for numerous applications.
Taken together, these advantages can render carbon electrode-molecule junctions as a reliable platform for molecular electronics towards practical applications.
TERMINTATE

Answer of GPT-4.0:

1. Enhanced Stability: Carbon electrode-molecule junctions have shown to be more stable compared to traditional metal electrode-based junctions. This is mainly due to the strong covalent bonds that can form between carbon and other atoms.
2. Wide Energy Gap: Carbon atoms have a wide energy gap which allows for better control over the molecule's electronic properties. This can lead to more precise tuning of the molecule's behavior and, ultimately, the creation of more efficient devices.
3. Flexibility: Carbon has a very high degree of flexibility, allowing it to be manipulated into various shapes and structures. This makes carbon electrode-molecule junctions highly adaptable to different applications.
4. Resistance to Oxidation: Carbon is more resistant to oxidation than many metals, which can give devices based on carbon electrode-molecule junctions a longer lifespan.
5. Better Thermal Conductivity: Carbon, especially in its graphene form, has excellent thermal conductivity making it an excellent choice for applications where heat dissipation is a concern.
6. Higher Sensitivity: Some studies suggest that carbon electrode-molecule junctions can be more sensitive than their metal counterparts, making them potentially more effective in applications such as sensors.
7. Incorporation with organic molecules: It is easier to connect carbon electrodes with organic molecules which can lead to the creation of more complex and functional devices.
8. Environmental Impact: The use of carbon-based devices can be environmentally more preferable as it reduces reliance on metal mining and processing, which can have high environmental costs.

Automatic Evaluation with ChatGPT-4.0

Upon examining the responses from GPT-3.5, SMELLM-3.5, GPT-4, and SMELLM-4.0 regarding the potential advantages of carbon electrode-molecule junctions over ordinary metal electrode-based molecular junctions, several observations can be made regarding their accuracy, depth, and coherence.

1. GPT-3.5 provides a coherent and accessible answer, highlighting enhanced stability, reduced reactivity, tunable properties, cost-effectiveness, and compatibility with a wide range of molecules. The response is clear and relevant, focusing on practical benefits such as stability, cost, and versatility without delving deeply into the specific scientific or molecular interactions.

2. SMELLM-3.5's response emphasizes chemical inertness, anchor possibilities, stability and compatibility, contact flexibility, and the stability of the carbon-molecule interface. This response adds unique insights into the suitability of carbon electrodes for biomolecular sensing and DNA sequencing, offering a slightly deeper dive into specific applications and molecular junction construction than GPT-3.5.

3. GPT-4 expands on the advantages by discussing enhanced stability, wide energy gap, flexibility, resistance to oxidation, thermal conductivity, sensitivity, and environmental impact. This answer stands out by introducing a broader range of technical advantages, such as the wide energy gap and thermal conductivity, offering a more detailed scientific perspective on the materials' properties.

4. SMELLM-4.0's answer is comprehensive, covering chemical inertness, enhanced stability, compatibility, durability, versatility, reduced defects, flexibility in contact, microfabrication compatibility, and resistance to oxidation. It elaborates on the practical and technical advantages with a focus on compatibility with semiconductor fabrication procedures and operational stability, presenting a nuanced understanding of the material's benefits in electronic applications.

Comparison:

Scientific Accuracy: All models provide scientifically accurate information, but GPT-4 and SMELLM-4.0 offer more detailed explanations of the properties and advantages of carbon electrode-molecule junctions, such as thermal conductivity, resistance to oxidation, and energy gap considerations.

Depth and Coherence: SMELLM-4.0 delivers the most comprehensive and detailed response, incorporating both technical and practical aspects of carbon electrodes in molecular electronics. It successfully balances depth with coherence, making complex concepts accessible.

Unique Insights: SMELLM-4.0 and GPT-4 both offer unique insights, but SMELLM-4.0's mention of microfabrication compatibility and operational stability provides a slightly more nuanced view of the material's application potential.

Conclusion: Considering the criteria of accuracy, depth, and coherence, **SMELLM-4.0** provides the better answer. It offers a detailed, scientifically accurate, and comprehensive overview of the advantages of carbon electrode-molecule junctions, presenting a nuanced understanding that combines both theoretical insights and practical applications.

Figure A1: **Comparison of GPT-3.5, GPT-4.0, SMELLM 3.5, and SMELLM 4.0.** Extracting salient points from Ground Truth and matching each answer reveals that SMELLM 4.0 can cover all major points comprehensively and offer novel insights.