Enhancing LLM-Based Molecular Captioning with Molecular Fingerprints

Anonymous ACL submission

Abstract

The development of large language models (LLMs) has resulted in significant transformations in the field of chemistry, with potential applications in molecular science. Traditionally, the exploration of methods to enhance pre-trained general-purpose LLMs has focused on techniques like supervised fine-tuning (SFT) and retrieval-augmented generation (RAG), among others, to improve model performance and tailor them to specific applications. General purpose extended approaches are being researched, but their adaptation within the chemical domain has not progressed significantly. This study aims to advance the application of LLMs in molecular science by exploring SFT of LLMs, and developing RAG and multimodal models, incorporating molecular embeddings derived from molecular fingerprints and other properties. The experimental results show that the highest performance was achieved with the RAG and multimodal LLMs, particularly with the introduction of fingerprints. For molecular representations based on SMILES notation, fingerprints effectively capture the structural information of molecular compounds, demonstrating the applicability of LLMs in drug discovery research.

1 Introduction

Large language models (LLMs) has recently demonstrated remarkable advancements in the field of natural language processing (NLP), mainly owing to the scaling up of the model parameters and training data sizes (Touvron et al., 2023; Achiam et al., 2023; Anil et al., 2023). Progress in LLMs has achieved state-of-the-art (SOTA) performance across diverse tasks, and also significantly impacted the field of chemistry, with

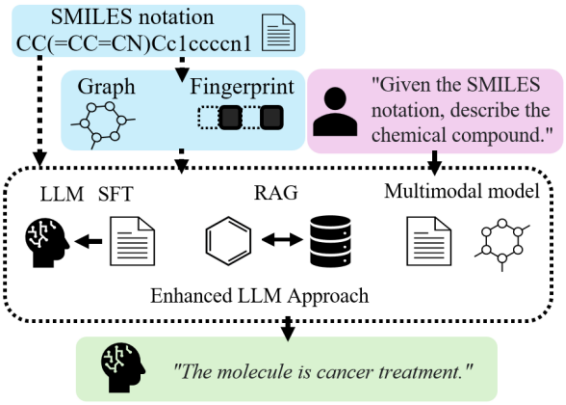


Figure 1: Overview of our molecular captioning task.

applications rapidly emerging in areas such as drug discovery and domain-specific information retrieval (Zheng et al., 2024; Zhang et al., 2024; Xiao et al., 2024). Molecular captioning is one of the representative tasks in the chemical application. In this task, a model takes chemical structure information, such as a Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988) or molecular graph, and generates a textual description of the compound's properties. It enables researchers to understand compound feature more easily, accelerating drug discovery. Generally, SMILES, a textual data format, is used as input for this task with LLMs (Edwards et al., 2022).

Improving the accuracy of LLMs for specialized tasks are classified into two strategies: model-centric improvements and prompt-centric improvements. The model-centric approach focuses on refining the LLM itself, for example through architectural changes, continual pre-training, or supervised fine-tuning (SFT). Especially, SFT is a promising technique due to its relatively low training cost compared to pre-training. The prompt-centric approach focuses on optimizing the input given to the model. This can involve techniques like prompt engineering, in-context learning or the use of Retrieval-Augmented

Generation (RAG) to retrieve and incorporate relevant information from external sources.

While considerable research explores for these approaches, their application to the molecular captioning task remains relatively unexplored. A key challenge in applying LLMs to chemistry is how to represent and input chemical structures for them. This critical question of optimal molecular representation within the LLM framework remains largely unaddressed.

In this study, we investigate the effectiveness of various approaches for improving LLM-based molecular captioning tasks with SMILES notation (Figure 1). The first approach involves SFT of a closed-source LLM, using SMILES text as the input and the corresponding descriptive text as the ground truth to create a specialized LLM for describing molecular compounds. Because closed-source LLMs are more powerful at understanding text due to their larger model parameters, this SFT approach can achieve more precise inference compared to fine-tuning open-source LLMs. The second approach employs RAG to leverage the similarity of SMILES strings to retrieve the related compound data. This is intended to allow the LLM to describe molecular compounds that may not have been sufficiently learned or have complex properties not present in the training data. In addition to conventional text embedding-based retrieval for RAG, we incorporate fingerprint-based retrieval using the Tanimoto coefficient (Bajusz et al., 2015) as a similarity metric to retrieve structurally similar compounds. The third approach utilizes multimodal-LLMs with molecular compound embeddings. In multimodal models, the way to embed new modal data is crucial. Here, we compare different types of embeddings: molecular fingerprint, graph neural network embedding, and language model embedding.

Experimental results on a benchmark dataset of molecular compounds show that, among molecular embeddings, the use of molecular fingerprints for RAG and the incorporation of molecular fingerprints as an integrated input for multimodal-LLM yielded the highest accuracy in each approach. Specifically, the latter multimodal model demonstrated the highest performance in this study. This suggests that molecular fingerprints best capture molecular property information compared with the other two embedding methods and it's more effective to use a general model with structural information (multimodal) than to improve unimodal model training methods. These findings suggest the potential to support the analysis of molecular compounds and improve the efficiency of drug discovery research.

2 Related Works

2.1 Representation of molecules

For the three representations of molecules, the information content of the graphs and the SMILES is considered equivalent. SMILES is a simple notation representing molecular structures as a single string. It uses element symbols for atoms and symbols for bonds, making it easy to use in machine learning.

SMILES embeddings are typically obtained using language models. For SMILES embedding, molecular language models that extend Transformer-based models (Vaswani et al., 2017) like T5 (Raffel et al., 2020) or BERT (Devlin et al., 2019) for chemistry, such as molbert or MolT5, are used (Edwards et al., 2022; Fabian et al., 2020; Chithrananda et al., 2020; Ahmad et al., 2022).

Graphs are variable-length data structures capable of representing three-dimensional (3D) structural information. With advancements in deep learning, graph neural network (GNN)-based models (Zhou et al., 2020; Scarselli et al., 2008) are commonly used to generate graph embeddings like MolCLR (Wang et al., 2022).

Molecular fingerprints are vectors, typically binary, that are calculated from SMILES strings using algorithms (Rogers & Hahn, 2010). These vectors store information about the presence or absence of structural features in a compound. Their fixed-length nature allows them to be readily input into general-purpose machine learning models.

2.2 Molecule-text multimodality

nach0 (Livne et al., 2024), a T5-based model trained to acquire molecular chemistry knowledge, enables multimodal reasoning by distinguishing between SMILES and natural language text tokens. Furthermore, research has been conducted on models that perform contrastive learning after encoding chemical structures and text to solve downstream tasks such as property prediction (Su et al., 2022; Liu et al., 2023; Luo et al., 2023) , and on models that have been extended to include images as input (Liu et al., 2024). As an extension of LLMs, models that perform multimodal

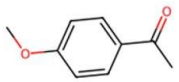
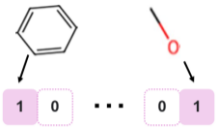
	Image	Feature	Convert method	Encoding Method
SMILES	COc1ccc(C(C)=O)cc1	Variable-length text	-	Molecular Language Model (molt5-large is used in this study)
Graph		Graph including node and edge	Rule based	Graph Neural Network (MolCLR is used in this study)
Molecular fingerprint		Fixed-length vector	Rule based (ECFP is used in this study)	-

Table 1: Three types of molecular representation. The right column represents the conversion methods from SMILES to their respective representation formats and the creation of embedding vectors employed in this study.

reasoning by adding molecular graphs as inputs to accurately capture the structural information of molecular compounds are also being developed (Liu et al., 2023; Cao et al., 2023). On the other hand, multimodal models utilizing molecular fingerprints, as well as comparative studies of these, have not been conducted.

3 Problem Settings

This study assumes two tasks using SMILES notations of molecular compounds. The first is the molecular captioning task, which involves explaining the properties of a molecular compound from its SMILES notation. For this task, it is desirable to appropriately describe the properties of the molecular compounds represented by the SMILES. The second task is the molecular property prediction and the experimental results of the second task are presented in detail in the Appendix, as part of additional validation.

We assumed that only SMILES is given as the data for molecular compounds, and cases in which molecular structure information is provided as data are not assumed. In the molecular embedding models, we used to be detailed in Table 1. RDKit was used for the transformation from SMILES to Graph and molecular fingerprints. Extended-Connectivity Fingerprints 4 (ECFP4) were adopted as the algorithm for the transformation to molecular fingerprints. Furthermore, molt5-large was used for SMILES embeddings, and MolCLR was used for Graph embeddings.

4 Proposed Methods

We proposed three approaches for predicting the properties of molecular compounds based on their SMILES text (Figure 2).

4.1 First Approach: SFT

For the first approach, we used SFT of a closed-source LLM to create an LLM specialized in generating descriptive text from SMILES text by providing the SMILES text of a molecular compound as input data and its descriptive text as output data to an LLM. Although open-source LLMs allow for easy parameter customization through the SFT, their parameter counts are lower than those of closed-source LLMs. Generally, models with more parameters tend to have higher performance in text generation tasks. Therefore, the SFT of open-source LLMs may not achieve sufficient performance in text generation tasks.

4.2 Second Approach: RAG

In the second approach using RAG, a dataset of pairs of training molecule SMILES text and their corresponding descriptive text is stored in a database in advance. The molecule that was most similar to the input test data of the SMILES text was retrieved from the database. To prevent data leakage during the search, the SMILES stored in the database are not used in the test data**.** In this study, we performed similarity searches for similar molecular compounds using retrievers based on the following:

- Similarity of embeddings by a SMILES Encoder using CLIP
- Cosine similarity of embeddings of SMILES by MolT5
- The cosine similarity of GNN embeddings for graph-represented molecules.
- Tanimoto coefficient of molecular fingerprints

The Tanimoto coefficient is most suitable for similarity comparison of molecules converted to fingerprints. In this study, we provided the top five

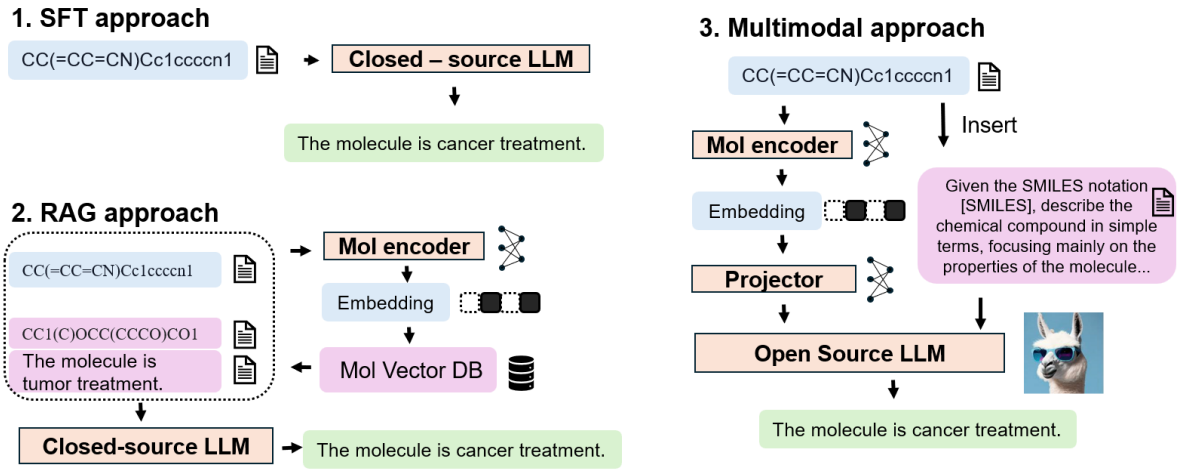


Figure 2: Details of our three approaches. Embeddings are created using three patterns: SMILES + MolT5, Graph + MolCLR, and Molecular fingerprint.

SMILES and caption pairs obtained through a similarity search of LLM and instructed it to generate an appropriate caption for the input SMILES.

4.3 Third Approach: Multimodal

The third approach involves multimodal-LLM using molecular fingerprints. This is an extension of the SFT method to the multimodal domain, where the LLM is given a molecular compound's SMILES text and fingerprint, enabling it to obtain structural information from SMILES and describe its properties. We implemented a multimodal LLM that processes instruction text and integrated inputs of SMILES, graph representations, or molecular fingerprints. The input SMILES undergoes a two-step branching process. First, it is converted into a molecular embedding by an encoder model. This embedding is then transformed via a projector into a vector with the same dimensionality as the LLM input and fed into the LLM. The other step involves embedding the SMILES string directly into the prompt as text. Finally, these inputs are integrated, and the LLM generates text. By including graph embeddings or fingerprints as inputs, the LLM is able to generate text while having captured the structural information of the molecular compounds.

	split train	split valid
Number of samples	126864	33696
Average SMILES length	108.5	105.4
Average number of sentence text words	30.39	29.53

Table 2: Dataset overview.

5 Experiments and Results

5.1 Dataset

We used the L+M-24 dataset,[1] an open dataset containing SMILES notation text of molecular compounds and text describing their properties. There are 3502 property names. the property can be divided into four categories: biomedical (=2032), light and electricity (=58), human interaction and organoleptic (=787), and agriculture and industry (=625). This is the most common dataset containing pairs of SMILES notations of molecular compounds and text describing their properties in English.

Table 2 provides the details of the dataset. The number of samples in the training data (split_train) is 126,864, and the number of samples in the validation data (split_valid) is 33,696. Each molecular compound is unique to all the samples. In this study, owing to computational time constraints, we randomly extracted 1000 samples (seed value: 42) from the validation data

[1] https://huggingface.co/datasets/language-plus-molecules/LPM-24_train

(split_valid) as test data for evaluation. Additionally, the training dataset was divided into training and validation sets in an 8:2 ratio. The divided training data were used to train the proposed methods, and the validation data were used to evaluate the checkpoint with the highest accuracy.

5.2 LLMs

For SFT approach, we utilized the custom tuning feature of Vertex AI Studio in a Google Cloud environment and used the gemini-1.0-pro-002 model of the closed-source LLM. Also, we used molt5-large, meta-llama-3-8B, meditron-7b, and nach0_large as the SFTs of the open-source LLMs. In addition, when training the LLM parameters, we used Lora to achieve lightweight fine-tuning. The computational environment for these experiments was an NVIDIA A100 40GB computer connected to Google Cloud Workstations.

For RAG approach, because a certain maximum length of the input context is required, we used Gemini-1.5-pro-002, which is a closed-source LLM. This allowed us to input all SMILES, similar to the input SMILES and their caption pairs into the LLM for captioning. In the RAG using CLIP, we used a distilbert-base-uncased text encoder to perform lightweight and high-speed training. The embedding dimensionality is 768. Because the SMILES fingerprint is represented by a 2048-dimensional vector, the SMILES encoder used a linear layer with an input of 2048 dimensions and an output of 2048 dimensions. During CLIP training, it is necessary to unify the dimensionality of these embeddings, we added a projector with 256 output dimensions for CLIP training. The input to the projector for the SMILES encoder was 2048 dimensions and the input to the projector for the text encoder was 768 dimensions.

For multimodal approach, from the perspective of high instruction-following ability and trainable parameters, meta-llama/Meta-Llama-3-8B was used as the base model for the multimodal model. As the dimensionality of the hidden layer embeddings of this model is 4096, the projection from the SMILES Encoder to the LLM used a linear layer with an input of 2048 dimensions and an output of 4096 dimensions. The parameters and computational environment for the training were the same as those for the SFT conducted with open-source LLMs. The LLM (Llama3) and projector are trained simultaneously, while the Mol encoder uses a pre-trained model with frozen parameters.

5.3 Evaluation Metrics

Following the paper that created the L+M-24 Dataset (Edwards et al., 2024), we used machine translation (MT) evaluation metrics, which have been used in NLP tasks, such as machine translation and text summarization, and property-specific scores, which calculate whether the generated text includes property-specific words of molecular compounds. For the MT evaluation metrics, we performed evaluations using natural language generation metrics such as BLEU-2/4 (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), and ROUGE-1/2/L (Lin, 2004). These metrics are similar to those used by Edward et al. (Edwards et al., 2022). Property-specific scores are calculated by matching tokenized names within the generated captions, specifically using macro-F1, precision, and recall.

Comparing MT evaluation metrics and property-specific scores, MT evaluation metrics are influenced by how grammatically similar they are to the ground-truth text. Therefore, the score may be high even if the characteristics of the molecular compound are not properly expressed. Property-specific scores are more appropriate evaluation metrics for assessing whether the characteristics of molecular compounds have been correctly captured.

5.4 Results

We evaluated the performance of our three proposed approaches compared with baselines including MolT5 (Edwards et al., 2022) and Meditron (Chen et al., 2023) , which were utilized in the L+M-24 dataset, and nach0, a high-scoring model from ACL 2024.

Figure 3 compares models using property-scores on the y-axis and models on the x-axis, and Table 3 delineates the model characteristics and evaluation metrics for each model. We compared against MolCLR, in Figure 3, represents a non-LLM, GNN-based predictive model which leverages the three-dimensional structure of molecules. It predicts the presence or absence of property related words and calculates only property-specific scores. Among the baselines, Meditron achieved the highest scores in MT evaluation, while nach0 achieved the highest property-specific score.

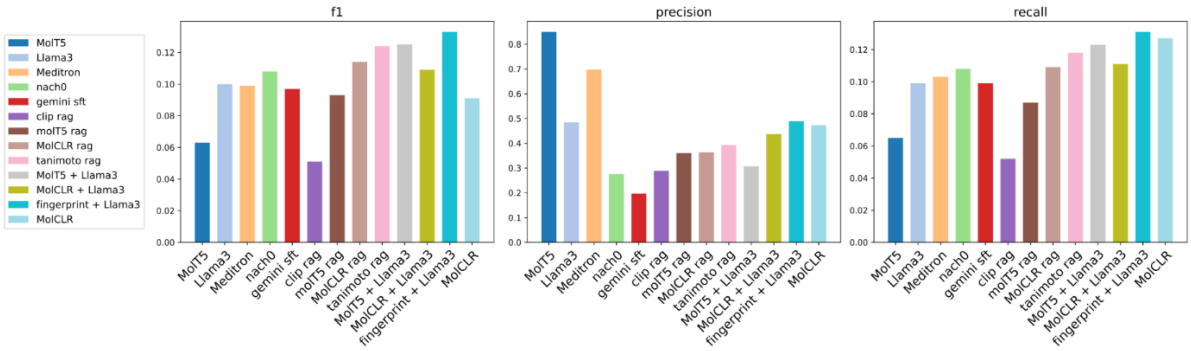


Figure 3: Overall property-specific score for molecular captioning using LLMs. Evaluation Metrics: macro-f1 score (f1), precision, recall. The model used for verification is the same as the one shown in Table 3.

Our proposed Closed-source LLM, Gemini (Team Gemini et al., 2023), fine-tuned through SFT (Gemini SFT), surpassed MolT5 and Llama3 in both MT evaluation and property-specific f1 scores. However, it did not outperform Meditron or nach0. The underperformance is likely due to Gemini's lack of specialization in chemical text generation and its inability to effectively distinguish SMILES strings from regular alphabet sequences during tokenization. This suggests that for domain-specific tasks with LLMs, a domain-specific training approach is more vital than model parameter size.

Conversely, the RAG approach, which does not involve SFT, yielded lower scores, failing to fully grasp the characteristics of captioning. Upon examining generated texts, we observed significant variations in grammar and phrasing compared to the ground truth, as well as instances of overly lengthy text. This is likely due to the LLM not having learned the structure of ground truth texts. This issue might be mitigated by adjusting the system prompt to encourage outputs that follow the ground truth text structure. For example, captions in the L+M-24 dataset often begin with "The Molecule is," a pattern not always captured by RAG-generated text. When comparing the property specific score, the Molecular fingerprint Tanimoto coefficient-based RAG model (Tanimoto rag) had the highest f1 score among the RAG approaches. This RAG model also surpassed the performance of the other unimodal models using SFT, suggesting that RAG can be a feasible approach when computational resources are limited.

Multimodal LLM captioning consistently achieved the highest prediction accuracy overall in both MT and property-specific scores across all three approaches. Particularly, the multimodal model using fingerprint embeddings achieved the highest scores overall.

The superior performance of models that incorporate molecular structure information, either via multimodal methods or molecular fingerprints in RAG, suggests that accurately representing chemical structure is paramount for LLMs. Our results show that correctly encoding chemical structure allows general-purpose LLMs like Llama3 to outperform domain-specific unimodal models in tasks such as molecular captioning. The strong performance of models using molecular fingerprints in both RAG and multimodal settings underscores that text encoder-based representations like those in MolT5 and nach0 may not always fully capture crucial molecular features like the presence of atoms, bonds, and rings. If MolCLR or MolT5 cannot produce embeddings that adequately capture these structural aspects, the prediction accuracy may suffer. In contrast, molecular fingerprints explicitly represent the local characteristics of molecules, enabling models to easily discern meaningful features.

Based on these findings, we conclude that for the molecular captioning task, multimodal models with SFT are the most effective approach. Furthermore, when computational resources are constrained, RAG offers a viable alternative for generating descriptions based on similar molecules. Across all methods, molecular fingerprint representations, which explicitly encode structural information as vectors, consistently yielded the best results.

(a) SFT approach	BLEU-2	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
MolT5 (baseline)	0.048	0.036	0.310	0.427	0.325	0.402
Meditron (baseline)	0.754	**0.545**	**0.713**	**0.767**	**0.580**	**0.551**
nach0 (baseline)	**0.756**	0.543	0.707	0.745	0.544	0.525
Llama3 (baseline)	0.721	0.521	0.700	0.755	0.565	0.545
Gemini SFT	0.745	0.533	0.694	0.731	0.530	0.512

(b) RAG approach	BLEU-2	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
CLIP-rag	0.103	0.030	0.202	0.195	0.052	0.155
Molt5-rag	**0.190**	**0.087**	**0.336**	**0.299**	**0.108**	**0.201**
MolCLR-rag	0.182	0.081	0.327	0.291	0.099	0.196
Tanimoto-rag	0.176	0.079	0.323	0.286	0.099	0.199

(c) MM approach	BLEU-2	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
MolCLR + Llama3	0.766	0.552	0.725	0.771	0.573	0.549
MolT5 + Llama3	0.727	0.525	0.714	0.770	0.575	0.555
fingerprint + Llama3	**0.776**	**0.560**	**0.738**	**0.785**	**0.587**	**0.563**

Table 3: MT scores for molecular captioning using LLMs of (a) SFT approach, (b) RAG approach, and (c) multimodal (MM) approach, respectively. The best performing model for each metric is shown in bold.

5.5 RAG vs. multimodal model SFT

In a general-purpose LLM approach, SFT often requires repeated training to memorize specific information. In contrast, RAG can predict information does not present in the training data with few-shot learning by externally inserting knowledge into the prompt. To confirm this in our study, we compared Tanimoto-rag (our best performing RAG model) with fingerprint + Llama3 (our best performing fine-tuned multimodal model). Figure 4 plots the frequency of property words within the training data against the accuracy of those words appearing in the generated text. The left side of the figure plots words with a training data frequency below 100, while the right-side plots words with a frequency above 100.

As shown in Figure 4, for properties with a limited number of samples in the training data, multimodal models tend to struggle with accurate predictions, while Retrieval-Augmented Generation models show higher accuracy. Therefore, the performance of multimodal models relies on high-frequency properties. For instance, properties with a frequency exceeding 10,000, such as "alcohol," "fatty," and "catalyst," achieved accuracy above 99% across all models that underwent supervised fine-tuning, except for MolT5.

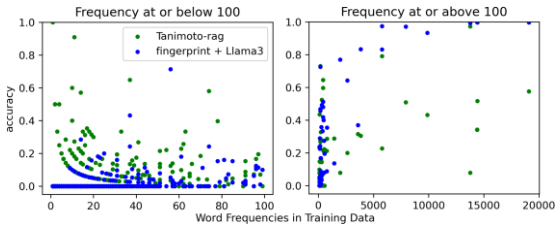


Figure 4: Training data property count and generated text accuracy. Molecular fingerprint is used for both Tanimoto-rag and fingerprint + Llama3.

Table 4 gives the macroF1 scores of RAGs and multimodal approach for each categorized property. All model's categorized property specific scores are listed in the Appendix. As indicated in Table 4, the performance categorized by different properties generally favors multimodal models. However, for properties related to "Light and electricity" category, RAG approach exhibit better performance. This can be attributed to the relatively low frequency of properties within the "Light and electricity" category, with the maximum frequency being around 500, suggesting that the supervised fine-tuning of multimodal models was not successful for these properties. The study showed similar trends to those seen in general-purpose LLMs, and it is expected that applying RAG to chemistry-specific LLM that have undergone SFT,

	Biomedical	Human Interaction and Organoleptics	Agriculture and Industry	Light and electricity
Tanimoto-rag	0.254	0.061	0.030	**0.151**
fingerprint + Llama3	**0.281**	**0.073**	**0.064**	0.113

Table 4: Categorized property-specific score (macro-f1).

can lead to the creation of more robust models, even for properties with insufficient sample data.

6 Conclusions

This study explored three enhancement approaches, SFT, RAG, and multimodal LLMs for predicting molecular compound properties from SMILES notation. In the SFT approach, we fine-tuned a closed-source LLM using the Gemini API, achieving superior MT evaluation scores compared to MolT5 and Llama3, although it did not surpass the performance of Meditron or nach0. The RAG-based model exhibited property-specific scores comparable to those achieved by the SFT-trained model. Notably, both RAG and multimodal LLMs demonstrated higher scores when processing molecular fingerprints as input, rather than SMILES or graph representations. Specifically, a multimodal model with fingerprint inputs to the LLM achieved the highest overall performance.

These findings highlight the potential of LLMs in drug discovery research and suggest their promise for improving the efficiency of future pharmaceutical development.

Future research directions include optimizing prompt systems to improve RAG performance, investigating methods for combining SFT and RAG, developing effective techniques for integrating molecular fingerprints into multimodal models, exploring regression tasks within LLM-based molecular property prediction, leveraging more sophisticated representations based on molecular descriptors and structural information, and evaluating the applicability of these techniques to real-world drug discovery and other relevant tasks. Additionally, further investigation is required into the impact of increasing modalities, such as multimodal models that simultaneously input molecular fingerprints and graphs.

Limitations

The model used in our research had approximately 8 billion parameters, suggesting the need for experimentation with models that have a significantly larger parameter count. However, due to limitations in the memory capacity of our available computing machines, we were unable to explore models with substantially larger parameter counts at this study. In addition, due to the training time required for fine-tuning and data privacy concerns, we were only able to conduct experiments using a single open dataset.

References

OpenAI. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. arXiv preprint arXiv:2209.01712.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and others. 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403.

David Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics, volume 7, pages 1-13.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. arXiv preprint arXiv:2311.16208.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, and others. 2023. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. Journal of the Association for Computing Machinery, 28(1):114-133.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. ChemBERTa: large-scale self-

supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. : Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 375–413. Association for Computational Linguistics.

Carl Edwards, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+ M-24: Building a Dataset for Language+ Molecules@ ACL 2024. arXiv preprint arXiv:2403.00791.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. arXiv preprint arXiv:2011.13230.

Team Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and others. 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023. MolCA: Molecular Graph-Language Modeling with Cross-Modal Projector and Uni-Modal Adapter. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15623–15638.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. Computers in biology and medicine.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multi-modal molecule structure--text model for text-based retrieval and editing. Nature Machine Intelligence, 1447–1457.

Micha Livne, Zulfat Miftahutdinov, Elena Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, Alán Aspuru-Guzik, and others. 2024. nach0: Multimodal natural and chemical languages foundation model. Chemical Science, 8380–8389.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. Molfm: A multimodal molecular foundation model. arXiv preprint arXiv:2307.09484.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, volume 21, pages 1–67.

RDKit: Open-source cheminformatics. https://www.rdkit.org

David Rogers and Mathew Hahn. 2010. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling, volume 50, pages 742-754.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. IEEE transactions on neural networks, 19(1):61-80.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. arXiv preprint arXiv:2209.05481.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothèe Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and others. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc.,

David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, volume 28, pages 31–36.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, volume 4, pages 279–287.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2024. A comprehensive survey of large language models and multimodal large language models in medicine. arXiv preprint arXiv:2405.08603.

Yanxin Zheng, Wensheng Gan, Zefeng Chen, Zhenlian Qi, Qian Liang, and Philip S Yu. 2024. Large language models for medicine: a survey. International Journal of Machine Learning and Cybernetics, volume 16, pages 1015–1040.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, and others. 2024. Scientific large language models: A survey on biological & chemical domains. arXiv preprint arXiv:2401.14656.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. AI open, volume 1, pages 57–81.

A Categorized property-specific score

Table 5 displays the categorized property-specific scores. We observed that biomedical properties were generally easier to predict. The performance was largely consistent regardless of the molecular representation method used (SMILES, graph, or molecular fingerprints).

B Molecular Property Prediction

Molecular property prediction involves predicting the property labels of a molecular compound using SMILES notation. For this task, accurate prediction of the property labels of the molecular compound represented by SMILES is desirable. For instance, blood‑brain barrier penetration (BBBP) can occur.

The input SMILES is given as text, and if the molecular compound given in SMILES can penetrate the blood–brain barrier, the output will be "Yes," otherwise, it will be "No." In this study, molecular property prediction solves only binary classification tasks, where whether the molecular compound exhibits a certain property is represented in a binary format; it does not solve regression tasks. This is because, given that the LLMs output tokens probabilistically in the forward direction, numerical regression tasks are challenging. In contrast, classification tasks are easier to solve because probabilistically outputting tokens is equivalent to multiclass classification.

	Biomedical			Human Interaction and Organoleptics			Agriculture and Industry			Light and electricity		
	p	r	f1	p	r	f1	p	r	f1	p	r	f1
MolT5	0.886	0.200	0.203	**0.990**	0.001	0.001	**0.960**	0.022	0.025	0.564	0.038	0.021
Llama3	0.568	0.255	0.259	0.377	0.031	0.037	0.790	0.048	0.051	0.204	0.061	0.052
Meditron	0.868	0.255	0.258	0.413	0.045	0.044	0.914	0.056	0.058	**0.592**	0.058	0.036
nach0	0.536	0.263	0.265	0.315	0.055	0.054	0.190	0.059	0.059	0.064	0.053	0.054
Gemini SFT	0.355	0.256	0.248	0.251	0.033	0.035	0.111	0.057	0.050	0.073	0.050	0.054
CLIP rag	**0.895**	0.192	0.193	0.158	0.010	0.007	0.102	0.004	0.005	0.000	0.000	0.000
Molt5 rag	0.649	0.211	0.220	0.381	0.058	0.057	0.161	0.024	0.027	0.254	0.054	0.068
MolCLR rag	0.651	0.219	0.232	0.380	**0.071**	0.065	0.170	0.029	0.031	0.252	0.118	0.129
Tanimoto rag	0.765	0.234	0.254	0.345	0.070	0.061	0.187	0.028	0.030	0.276	**0.139**	**0.151**
MolCLR + Llama3	0.578	0.277	0.280	0.315	0.062	0.066	0.200	0.055	0.059	0.134	0.097	0.096
MolT5 + Llama3	0.554	0.268	0.269	0.371	0.052	0.052	0.734	0.059	0.060	0.089	0.067	0.055
fingerprint + Llama3	0.572	**0.280**	**0.281**	0.484	**0.071**	**0.073**	0.707	**0.063**	**0.064**	0.194	0.111	0.113

Table 5: Categorized property-specific score. p is precision, r is recall, f1 is macro-f1 score.

	BBBP	Clintox	HIV	bace
Detail of task	Binary labels of blood-brain barrier penetration (permeability).	Qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.	Experimentally measured abilities to inhibit HIV replication.	Quantitative (IC50) and qualitative (binary label) binding results for a set of inhibitors of human β-secretase 1(BACE-1).
Number of samples	2039	1480	41127	1513
Positive label ratio	0.765	0.936	0.035	0.458
Task Type	Binary Classification	Binary Classification	Binary Classification	Binary Classification

Table 6: Molecule Net dataset overview.

B.1 Dataset

For molecular property prediction, we used five datasets released by Molecule Net[2], a large-scale benchmark that organizes several public datasets for molecular machine-learning evaluation. All datasets used in this research were for binary classification tasks that express whether a compound exhibits an arbitrary property in a binary format, and datasets for solving regression tasks were not used. To preprocess the datasets, all samples containing SMILES that could not be converted to fingerprint notation via rdkit were removed. Table 6 shows the types of datasets used and their basic statistics.

These datasets were divided into training, validation, and test data in a ratio of 6:2:2. Similar to molecular captioning, the divided training data were used to train the proposed methods, and the validation data were used to evaluate the c heckpoints with the highest accuracy. All the parameters used for the experiments were the same as those used for molecular captionin.

B.2 Results

As a baseline, we converted the SMILES into fingerprints and performed predictions using linear regression (LR), XGBoost (XGB), support vector machine (SVM) and Neural Network (NN). Additionally, we used knn_tanimoto, which performs nearest-neighbor searches based on the

[2] https://moleculenet.org/

Tanimoto coefficient of fingerprint embeddings, and knn_MolT5, which performs predictions based on the cosine similarity of the MolT5 embeddings. This is equivalent to the retrieve operation performed when doing RAG with an LLM.

We also performed predictions via encoder models, such as molbert, MolT5, and nach0. Furthermore, research is currently underway to perform binary classification based on LLMs, and by fine-tuning an LLM to ask for either "Yes" or "No," evaluation on the basis of the probability distributions of "Yes" or "No" outputs is possible.

Owing to the API specifications, we did not conduct experiments using closed-source models because it is difficult to output the probability distributions of words. We verified a multimodal model by encoding with MolT5 and a multimodal model via fingerprints. We used the predictions made via fine-tuned Llama3 as the baseline for the LLM SFT.

Tables 7 and 8 show the ROC-AUC and PR-AUC scores for binary classification for each dataset. The prediction model using MolCLR has not achieved accuracy surpassing that of text-based models. As with molecular captioning, this is likely due to the loss of information, such as the representation of isomers in SMILES notation, when it is converted into a molecular graph. Comparing the two KNN models, there are cases where the Tanimoto coefficient of the fingerprint and the cosine similarity of the MolT5 embedding are better, depending on the task. Because the Tanimoto coefficient indicates structural similarity, the similarity of molecular structure may not considerably affect some molecular properties.

It can also be seen that transformer encoder-based models, such as molbert, MolT5, and nach0 (T5 base), are more accurate than the Llama3-based models, including the multimodal model. This is apparent from the fact that transformer decoder models, such as Llama3, are designed with an emphasis on text generation and are not suitable for classification and that Llama3 cannot properly tokenize molecules expressed in SMILES. By contrast, the Llama3 multimodal model, which uses fingerprints, achieved an accuracy similar to that of the other transformer encoder models. This shows that even without properly tokenizing the SMILES, fingerprints contain sufficient molecular information.

	BBBP	clintox	HIV	bace
fingerprint + LR	0.910	0.627	0.755	0.904
fingerprint + XGB	0.929	0.675	0.802	0.922
fingerprint + SVM	0.897	0.631	–	0.889
fingerprint + NN	0.917	0.640	0.785	0.903
MolCLR	0.894	0.766	0.773	0.816
Molbert	0.957	0.998	0.759	0.863
MolT5	0.958	0.996	0.661	0.626
nach0	0.963	0.999	0.785	0.895
Llama3	0.812	0.822	0.746	0.720
MolT5 + Llama3	0.956	0.994	0.789	0.841
fingerprint + Llama3	0.953	0.981	0.774	0.878
MolCLR + Llama3	0.884	0.824	0.763	0.715

Table 7: ROC-AUC of molecule property

	BBBP	clintox	HIV	bace
fingerprint + LR	0.967	0.952	0.260	0.855
fingerprint + XGB	0.972	0.956	0.421	0.891
fingerprint + SVM	0.964	0.957	-	0.844
fingerprint + NN	0.969	0.960	0.374	0.849
MolCLR	0.958	0.980	0.077	0.752
Molbert	0.987	1.000	0.355	0.818
MolT5	0.988	1.000	0.101	0.513
nach0	0.990	1.000	0.381	0.857
Llama3	0.929	0.984	0.205	0.688
MolT5 + Llama3	0.987	1.000	0.345	0.778
fingerprint + Llama3	0.986	0.999	0.341	0.825
MolCLR + Llama3	0.952	0.979	0.269	0.609

Table 8: PR-AUC of molecule property prediction.

B Output Text

Figure 5 shows the text generated by each molecule captioning method, along with the ground truth. The Gemini SFT and multimodal models exhibited high lexical recall against the ground truth, whereas the Tanimoto RAG, which lacks SFT, produced

texts with larger word counts and more technical terms. As mentioned in the main text, the RAG performance is attributed to the model itself not having learned vocabulary or phrasing.

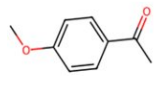
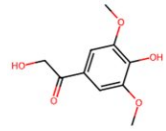
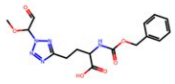
Molecule	Ground Truth	Gemini SFT	Tanimoto RAG	Multimodal Model
 COc1ccc(C(C)=O)cc1	When heated to decomposition it emits acrid smoke and irritating fumes. The molecule has both a Bitter and unpleasant taste and a Pleasant odor.	When heated to decomposition it emits acrid smoke and irritating vapors.	The compound is an aromatic molecule with a ketone group and a methoxy group attached. It likely has some lipophilic character due to the aromatic ring and ketone. Further properties cannot be reliably predicted based solely on the SMILES notation without additional information.	The molecule is a energy source, energy storage, emulsifier, surfactant, nutrient, membrane stabilizer.
 COc1cc(C(=O)CO)cc(OC)c1O	The molecule is a flavoring agent, savory, meaty, and sulfur.	The molecule is a nutrient.	The molecule is a substituted phenol with ester and ether groups. It likely possesses properties related to emulsification, membrane stabilization, and potentially acts as an energy source or nutrient.	The molecule is a nutrient.
 COC(C=O)n1nnc(CCC(NC(=O)OCc2ccccc2)C(=O)O)n1	The molecule is a anti-cancer.	The molecule is a cancer treatment.	The molecule is a complex nitrogen-containing ring system with amide and ester groups, likely exhibiting polar properties and potentially acting as a drug due to its structural similarities to known pharmaceutical compounds.	The molecule is a anti-inflammatory.

Figure 5: Output caption of models. Molecular fingerprinting is used for multimodal model.