

Tooling or Not Tooling? The Impact of Tools on Language Agents for Chemistry Problem Solving

Botao Yu^{†C}, Frazier N. Baker^{*CB}, Ziru Chen^{*C}, Garrett Herb^C, Boyu Gou^C,
Daniel Adu-Ampratwum^P, Xia Ning^{BCP}, Huan Sun^{†C}

^CDepartment of Computer Science and Engineering, OSU

^BDepartment of Biomedical Informatics, OSU ^PCollege of Pharmacy, OSU

Abstract

To enhance large language models (LLMs) for chemistry problem solving, several LLM-based agents augmented with tools have been proposed, such as ChemCrow and Coscientist. However, their evaluations are narrow in scope, leaving a large gap in understanding the benefits of tools across diverse chemistry tasks. To bridge this gap, we develop ChemAgent, an enhanced chemistry agent over ChemCrow, and conduct a comprehensive evaluation of its performance on both specialized chemistry tasks and general chemistry questions. Surprisingly, ChemAgent does not consistently outperform its base LLMs without tools. Our error analysis with a chemistry expert suggests that: For specialized chemistry tasks, such as synthesis prediction, we should augment agents with specialized tools; however, for general chemistry questions like those in exams, agents' ability to reason correctly with chemistry knowledge matters more, and tool augmentation does not always help.¹

1 Introduction

Large language models (LLMs) have demonstrated impressive problem-solving capabilities in many disciplines (Wang et al., 2024b; Yue et al., 2024; Grossmann et al., 2023). When it comes to chemistry, LLMs still face significant challenges, such as incorrect calculation, lack of domain knowledge, or inability to perform certain tasks like reaction prediction (Guo et al., 2023; Mirza et al., 2024). To address these limitations, LLM-based agents integrated with tools have been proposed to tackle chemistry-specific problems (Wang et al., 2024a; Ramos et al., 2024). For example, ChemCrow (M. Bran et al., 2024) expands LLMs' capabilities by incorporating 18 tools, ranging from web

search to chemical reaction prediction. Similarly, Coscientist (Boiko et al., 2023) integrates the control of cloud labs to enable LLMs to automate wet lab experiments.

Despite the promise of these tool-augmented agents, existing evaluations have been largely qualitative and limited in scope. For example, ChemCrow is assessed with only 14 individual tasks mainly focusing on compound synthesis, and Coscientist's evaluation involves merely six specific tasks. These narrow assessments leave a large gap in our understanding of how tool-augmented agents perform across diverse chemistry tasks in real-world applications.

In this work, we conduct a comprehensive evaluation of LLM-based agents on different chemistry tasks to grasp a deep understanding of their potential and limitations. To explore and enhance the capabilities of agents in diverse and complex chemistry scenarios, we introduce ChemAgent, a new chemistry agent capable of handling a wide spectrum of tasks. It leverages the ReAct framework (Yao et al., 2023) and integrates 29 tools, such as a search tool for PubChem (Kim et al., 2019), several molecular property predictors, as well as many practical tools present in ChemCrow. Then, we adapt two categories of real-world chemistry problems for systematic evaluation: specialized tasks and general questions. For specialized tasks, we use SMolInstruct (Yu et al., 2024), which contains 14 types of specialized molecule- and reaction-centric tasks. For general questions, we use MMLU-Chemistry and GPQA-Chemistry, which are chemistry-related subsets of the MMLU (Hendrycks et al., 2021) and GPQA (Rein et al., 2023) benchmarks, containing exam-like questions spanning high school, college, and graduate levels.

Through comprehensive experiments, we show that: While ChemAgent substantially outperforms ChemCrow on all chemistry tasks, it does not consistently outperform the base LLMs without

*Equal contribution.

†Correspondence to: {yu.3737, sun.397}@osu.edu.

¹<https://github.com/OSU-NLP-Group/ChemAgent>.

tools. In addition, the impact of tool augmentation is highly dependent on task characteristics. For specialized chemistry tasks involving professional molecular representations (e.g., SMILES (Weininger, 1988)) and specialized chemical operations (e.g., compound synthesis), augmenting LLMs with task-specific tools can yield substantial performance gains. Nonetheless, for general chemistry questions that require fundamental knowledge and extensive reasoning, ChemAgent cannot address these challenges adequately and underperforms the base LLMs. Further analysis along with a chemistry expert shows that ChemAgent’s underperformance on general chemistry questions is primarily due to delicate mistakes at intermediate stages of its problem-solving process, such as wrong logic and information oversight. Overall, our findings indicate that tool augmentation may introduce additional complexity that hinders LLMs’ reasoning and thus does not always help in chemistry problem-solving. Future research may improve LLM-based agents for chemistry by optimizing cognitive load and enhancing reasoning and information verification abilities.

2 ChemAgent

We introduce ChemAgent (Figure 1), a chemistry agent improved over ChemCrow (M. Bran et al., 2024) and equipped with enhanced tools for a wider range of tasks. It implements two essential cognitive abilities (Sumers et al., 2024) required for chemistry problem-solving: (1) **Reasoning**: This ability is required in the *Thought* step for comprehending user queries and tool outputs, assessing current status, and formulating subsequent steps. (2) **Grounding**: Based on the reasoning result (i.e., the “thought”), this ability determines the appropriate tool to execute and its corresponding input.

To enhance ChemAgent’s capabilities, we develop an extensive set of 29 tools (Appendix B), categorized into general, molecule, and reaction tools. General tools provide the agent with common problem-solving abilities, such as the execution of Python code for computations and various operations via PythonREPL. Molecule tools specialize in the analysis, prediction, and conversion of molecules and their properties. For example, FunctionalGroups can identify functional groups within a molecule, which is crucial for analyzing molecular characteristics. Lastly, reaction tools are instrumental in predicting chemi-

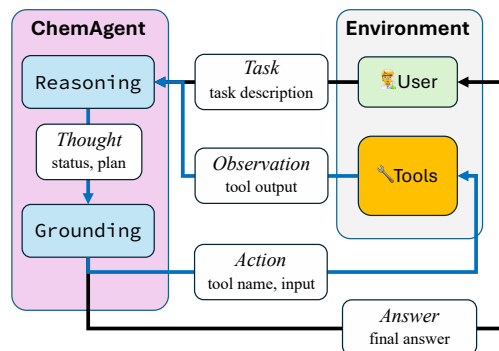


Figure 1: Our ChemAgent framework. Upon receiving a user *task*, the agent iterates through a three-step ReAct process (Yao et al., 2023): (1) *Thought* generation, analyzing the current situation and planning subsequent steps; (2) *Action* determination, selecting the appropriate tool and its input based on the generated thought; and (3) *Observation* obtaining, executing a tool in the environment and obtaining the results or feedback. This iterative cycle continues until task completion or conclusion, and the final *answer* is returned to the user.

cal reaction outcomes (ForwardSynthesis) and suggesting synthesis paths for desired products (Retrosynthesis), both of which are essential in applications like drug discovery (Berdigaliyev and Aljofan, 2020).

In this tool set, we create 16 new tools and enhance 6 existing ones in ChemCrow, which provides ChemAgent more comprehensive and robust abilities in solving chemistry problems. For example, we create PubchemSearchQA, which leverages an LLM to retrieve and extract authorized and comprehensive compound information from PubChem (Kim et al., 2019), and several molecular property predictors (BBBPPredictor, SideEffectPredictor, etc.), which employ neural networks (Zhou et al., 2023) for molecular property predictions. We also enhance WebSearch with an LLM-enhanced searching service to yield more comprehensive and flexible search results.

3 Experiments

3.1 Experimental Setup

Datasets. We use three well-established datasets (listed in Table 1) to thoroughly assess tool-augmented agents on two categories of chemistry problems: (1) **Specialized chemistry tasks** focus on experiment-like problems involving molecular manipulations, predictions, and representations. This category includes **SMolInstruct** (Yu et al., 2024), which contains 14 molecule- and

Category	Dataset	# Sample	Specific task type
Specialized tasks	SMolInstruct	700	Molecule- and reaction-centric tasks
General questions	MMLU-Chemistry	70	High school- and college-level questions
	GPQA-Chemistry	93	Graduate-level questions

Table 1: Datasets used in our experiments.

reaction-centric tasks and requires models to understand molecular representations like SMILES (Weininger, 1988) and perform specific chemical operations, such as predicting synthesis paths and converting chemical names (Figure C.1). (2) **General chemistry questions** resemble questions appearing in exams at different levels and test a wide range of fundamental knowledge and general reasoning in chemistry. This category includes **MMLU-Chemistry**, a manually verified chemistry subset of the MMLU benchmark (Hendrycks et al., 2021) that consists of questions at the high school and college level (Appendix C.2), and **GPQA-Chemistry**, the chemistry section of the GPQA-Diamond benchmark (Rein et al., 2023) that consists of difficult graduate-level questions.

LLMs and Agents. We compare our ChemAgent with two baselines: (1) State-of-the-art (SoTA) base LLMs, including GPT-4o (OpenAI, 2024) and Claude-3.5-Sonnet (Anthropic, 2024), which have shown superior capabilities in chemistry problem-solving among existing LLMs (Wang et al., 2024b). (2) ChemCrow (M. Bran et al., 2024), a pioneering chemistry-focused agent equipped with 18 expert-designed tools. For ChemCrow and ChemAgent, we utilize GPT-4o or Claude-3.5-Sonnet as the backbone language models, and refer to them as GPT and Claude, respectively.

3.2 Overall Performance

Specialized Chemistry Tasks. Models are evaluated on 50 randomly selected samples from the test set of SMolInstruct for each task, and the results on four selected tasks are presented in Table 2 (see Appendix C.1.2 for the full results). We can observe that: (1) ChemAgent exhibits substantial improvements over its base LLM counterparts, highlighting the critical role of domain-specific tools in augmenting LLMs’ capabilities on the specialized tasks in SMolInstruct. (2) Compared to ChemCrow, ChemAgent demonstrates superior performance. Our analysis suggests that the disparity is attributed to ChemCrow’s limited tool set and

Model	NC-S2I EM	PP-SIDER Acc	FS EM	RS EM
GPT-4o	0.0	44.0	12.0	0.0
Claude-3.5-Sonnet	2.0	62.0	22.0	0.0
ChemCrow (GPT)	2.0	36.0	72.0	8.0
ChemCrow (Claude)	2.0	32.0	70.0	22.0
ChemAgent (GPT)	70.0	70.0	78.0	42.0
ChemAgent (Claude)	70.0	68.0	80.0	42.0

Table 2: The results (%) on the SMolInstruct dataset. EM (exact match) and Acc (accuracy) are the metrics.

Model	MMLU-Chemistry	GPQA-Chemistry
GPT-4o	80.5	40.5
Claude-3.5-Sonnet	76.7	52.3
ChemCrow (GPT)	43.3	27.5
ChemCrow (Claude)	68.6	35.2
ChemAgent (GPT)	71.0	33.8
ChemAgent (Claude)	70.0	45.9

Table 3: The accuracy scores (%) on the MMLU-Chemistry and GPQA-Chemistry datasets, averaged over three runs.

the potential lack of robustness in its tool implementations. For instance, ChemCrow’s apparent deficiency in molecular property prediction tools and its limited web search capabilities seem to hinder its performance in property prediction tasks. In contrast, ChemAgent’s tool set (Appendix B) is more comprehensive and robust for LLMs to leverage effectively.

General Chemistry Questions. As presented in Table 3, contrary to our expectations, the ChemAgent variants underperform their base LLM counterparts. This trend persists across both datasets and is also observed with ChemCrow, suggesting a common issue in tool-augmented agents for chemistry. This observation challenges the intuition that tool augmentation would invariably enhance the performance of LLMs by providing additional information (Schick et al., 2023; Qu et al., 2024), and shows that both agents cannot fall back to base LLMs’ capabilities when tools offer no advantage, calling for a second thought on applying such agents on different tasks with more thorough experiments.

3.3 Error Analysis

To examine the errors made by ChemAgent, we use SMolInstruct and MMLU-Chemistry as representatives from their respective categories and conduct a manual error analysis. For all the samples where ChemAgent (GPT) fails in our experiments, we

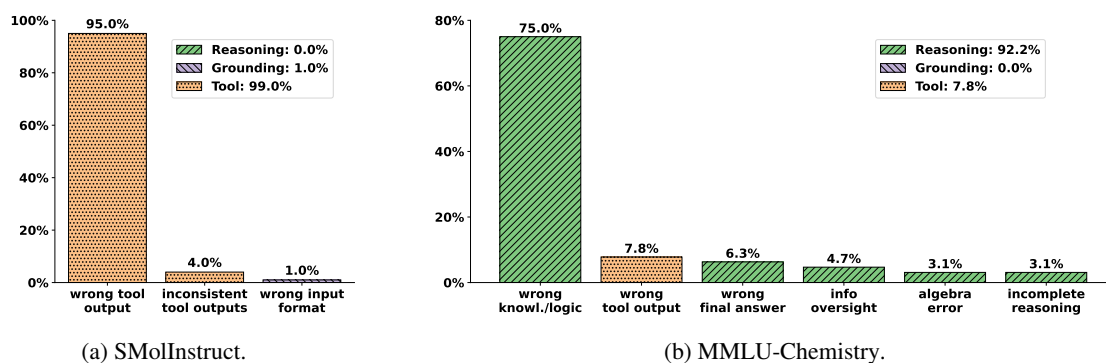


Figure 2: The error statistics of ChemAgent (GPT) on SMolInstruct (102 errors) and MMLU-Chemistry (64 errors).

engage a chemistry expert to analyze the errors, which are then classified into three types, namely **reasoning error**, **grounding error**, and **tool error**, based on the components (the cognitive abilities and the environment) responsible for the errors. The definitions of the errors identified during our experiment are as follows:

Reasoning errors. Errors made by the “reasoning” ability, where the agent inaccurately assesses the situation or devises an incorrect plan for subsequent steps, such as misinterpreting tool outputs or suggesting incorrect methodologies. Specifically, they include the following errors:

- Wrong knowledge/logic: an error where agent makes a mistake in applying chemistry knowledge or makes a conclusion that does not logically follow from the previous information.
- Wrong final answer: an error where the analysis process is correct but the final answer is wrong.
- Information oversight: an error where the agent neglects to consider relevant information given in the question or the previous steps.
- Algebra error: an error in algebraic manipulation or simplification, such as the incorrect solving of equations or misapplication of algebraic axioms.
- Incomplete reasoning: An error where the reasoning process is not fully developed, such as when solving a problem but omitting necessary steps or details.

Grounding errors. These occur during tool invocation, such as selecting an inappropriate tool, using an incorrect input format, or providing erroneous inputs to a tool. Specifically:

- Wrong input format: an error arising from data being provided in a format that the tool cannot process, resulting in failures or incorrect results.

Tool errors. These errors originate from the environment (i.e., the tools used in this study), where the tools either fail to execute properly or return

inaccurate information. Specifically:

- Wrong tool output: an error occurring when a tool produces incorrect or unexpected results, leading to faulty conclusions or actions.
- Inconsistent tool outputs: an error where multiple tools return inconsistent information, leading to faulty conclusions or actions.

As illustrated in Figure 2, the error distributions are very different on the two datasets. On SMolInstruct (Figure 2a), tool errors account for over 99.0% of all errors. These errors mainly stem from the neural networks-based tools (e.g., ForwardSynthesis, BBBPPredictor), which inherently possess imperfect accuracy. For these specialized tasks where dedicated tools exist, the agent can easily pinpoint and correctly use the needed tools (Appendix E), resulting in limited or no reasoning and grounding errors. In contrast, on MMLU-Chemistry (Figure 2b), reasoning errors constitute over 90.0%. This is because MMLU questions require broader knowledge and more intricate chemical reasoning and rely less on external tools. Our analysis indicates that all the observed reasoning errors manifest as delicate mistakes at intermediate stages of problem-solving, rather than incorrect overall methods. For instance, an inaccurate chemistry knowledge is applied or a mistaken conclusion is made (the wrong knowledge/logic error), or a wrong final option is selected despite of the correct analysis process (the wrong final answer error). Specific cases showcasing the errors can be found in Appendix D. Compared to LLMs without tools, the tool-augmented agent appears more prone to such delicate mistakes.

We hypothesize that the errors occur for two reasons: (1) **Increased cognitive load of agents:** The backbone LLM in the agent is tasked with multiple responsibilities, including task comprehension, tool selection, and tool output interpretation. This

necessitates frequent role-switching (Qiao et al., 2024), which, along with more complex contexts, may hinder the LLM’s ability to maintain a holistic and consistent approach to the main task (Verma et al., 2024), resulting in more reasoning errors. (2) **Potentially confusing tool outputs:** The tool outputs may occasionally be inaccurate or conflict with the model’s internal knowledge (Xie et al., 2024). This discrepancy can introduce confusion and lead to reasoning and tool errors. To address these issues, future research could focus on developing new agent frameworks that **reduce cognitive load and context distractions for LLMs**. This might involve *building multi-agent systems* to distribute the workload (Chen et al., 2023, 2024a), or *filtering out irrelevant information* to enhance task-focused reasoning (Shi et al., 2023; Yuan et al., 2024; Ouyang et al., 2023). Additionally, exploring **information verification mechanisms** could help LLMs resolve discrepancies from multiple sources, improving the accuracy of the final output.

4 Conclusion

In this paper, we conducted a comprehensive evaluation of tool-augmented language agents for chemistry problem-solving. We introduce ChemAgent, an enhanced chemistry agent with an enhanced tool set, and assess its performance across diverse chemistry problems, including specialized tasks and general questions. Our findings reveal that the impact of tool augmentation is highly dependent on task characteristics: While ChemAgent demonstrates significant improvements on specialized tasks, it does not surpass the base LLMs without tools on general questions. The manual error analysis highlights that tool errors predominate in specialized tasks, whereas reasoning errors are more frequent in general questions due to delicate mistakes in the problem solving process. To minimize reasoning errors and enhance performance in general questions, future agent design should focus on optimizing the cognitive load of LLMs and improving their ability to reason and verify information, especially when resolving inconsistencies from multiple sources.

Limitations

This study focuses on evaluating the performance of tool-augmented agents across various chemistry tasks. While it presents the most comprehensive evaluation on this topic to the best of our knowledge, there are several limitations:

- Our evaluation uses GPT-4o and Claude-3.5-Sonnet as the primary models for comparison and as backbones for the agents. This selection does not encompass a broader range of potential LLMs, such as Llama-3.2 (Meta, 2024) or Qwen2 (Yang et al., 2024). Although these additional models might exhibit different performance patterns, our choice of state-of-the-art (SoTA) models ensures strong baselines without significant loss of generality.
- While our study addresses both specialized tasks and general questions in the field of chemistry, it may not entirely represent the vast array of real-world chemistry problems. The scope might overlook certain nuanced challenges and scenarios encountered in practical applications.
- The manual error analysis, while thorough, is conducted by a chemistry expert together with two PhD students with chemistry knowledge and is limited in scale. This constraint could introduce potential biases and may not capture all subtle variations in error types or frequencies.

Acknowledgments

The authors would thank colleagues from the OSU NLP group and the OSU Ning Lab for constructive feedback. This research was supported in part by NSF CAREER #1942980, NLM R01LM014385, and NSF IIS-2133650. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

References

- Anthropic. 2024. Introducing claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>. Accessed: 2024-10-15.
- Nurken Berdigaliyev and Mohamad Aljofan. 2020. An overview of drug discovery and development. *Future medicinal chemistry*, 12(10):939–947.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.
- Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

- Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. 2024a. Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. arXiv preprint arXiv:2402.01620.
- Kexin Chen, Junyou Li, Kunyi Wang, Yuyang Du, Jiahui Yu, Jiamin Lu, Lanqing Li, Jiezhong Qiu, Jianzhang Pan, Yi Huang, Qun Fang, Pheng Ann Heng, and Guangyong Chen. 2024b. Chemist-x: Large language model-empowered agent for reaction condition recommendation in chemical synthesis. arXiv preprint arXiv:2311.10776.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024c. Teaching large language models to self-debug. In Proceedings of International Conference on Learning Representations (ICLR).
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. 2024d. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. arXiv preprint arXiv:2410.05080.
- MATTHEW CHERUKARA, ALKATERINI VRIZA, HENRY CHAN, TAO ZHOU, VARUNI KATTI SASSTRY, and MICHAEL PRINCE. 2024. Calms: Context-aware language model for science. Technical report, Argonne National Laboratory (ANL), Argonne, IL (United States).
- Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 375–413.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2024. Are we done with mmlu? arXiv preprint arXiv:2406.04127.
- Alireza Ghafarollahi and Markus J Buehler. 2024. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. Digital Discovery.
- Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. Science, 380(6650):1108–1109.
- Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In Proceedings of Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In Proceedings of International Conference on Learning Representations (ICLR).
- Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024. Crispr-gpt: An llm agent for automated design of gene-editing experiments. arXiv preprint arXiv:2404.18021.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.
- Yeonghun Kang and Jihan Kim. 2023. Chatmof: An autonomous ai system for predicting and generating metal-organic frameworks. arXiv preprint arXiv:2308.01423.
- Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2019. Pubchem 2019 update: improved access to chemical data. Nucleic acids research, 47:D1102–D1109.
- Mario Krenn, Florian Häse, A Nigam, Pascal Friederich, and Alán Aspuru-Guzik. 2019. Selfies: a robust representation of semantically constrained graphs with an example application in chemistry. arXiv preprint arXiv:1905.13741, 1(3).
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. Nature Machine Intelligence, pages 1–11.
- Andrew D McNaughton, Gautham Ramalaxmi, Agustin Krueel, Carter R Knutson, Rohith A Varikoti, and Neeraj Kumar. 2024. Cactus: Chemistry agent connecting tool-usage to science. arXiv preprint arXiv:2405.00972.
- Meta. 2024. Llama 3.2. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2. Accessed: 2024-10-15.
- Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelm, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, et al. 2024. Are large language models superhuman chemists? arXiv preprint arXiv:2404.01475.
- Odhran O’Donoghue, Aleksandar Shtedritski, John Ginger, Ralph Abboud, Ali Essa Ghareeb, Justin Booth, and Samuel G Rodrigues. 2023. Bioplanner: automatic evaluation of llms on protocol planning in biology. arXiv preprint arXiv:2310.10632.
- OpenAI. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

- OpenAI. 2024. Hello, gpt-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed: 2024-10-15.
- Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Jiawei Han, and Lianhui Qin. 2023. Structured chemistry reasoning with large language models. *arXiv preprint arXiv:2311.09656*.
- Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, chengfeiv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch via self-planning. In *ICLR Workshop on Large Language Model (LLM) Agents*.
- Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2024. Tool learning with large language models: A survey. *arXiv preprint arXiv:2405.17935*.
- Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2021. Stout: Smiles to iupac names using neural machine translation. *Journal of Cheminformatics*, 13(1):1–14.
- Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. 2024. A review of large language models and autonomous agents in chemistry. *arXiv preprint arXiv:2407.01603*.
- RDKit. 2023. *Rdkit: Open-source cheminformatics*. Accessed on 27 Jan 2024.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.
- Yixiang Ruan, Chenyin Lu, Ning Xu, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, Xiaodong Shen, Ning Ye, et al. 2024. Accelerated end-to-end chemical synthesis development with large language models. *ChemRxiv preprint*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, volume 36, pages 68539–68551.
- Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 31210–31227.
- Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.
- Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, et al. 2025. Chemagent: Self-updating library in large language models improves chemical reasoning. In *Proceedings of International Conference on Learning Representations (ICLR)*.
- Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. On the brittle foundations of react prompting for agentic large language models. *arXiv preprint arXiv:2405.13966*.
- Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2714–2730.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of Neural Information Processing Systems (NeurIPS)*, 35:24824–24837.
- David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of International Conference on Learning Representations (ICLR)*.

- Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. LLaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In Proceedings of Conference on Language Modeling (COLM).
- Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Kan Ren, Dongsheng Li, and Deqing Yang. 2024. EASYTOOL: Enhancing LLM-based agents with concise tool instruction. In Proceedings of ICLR Workshop on Large Language Model (LLM) Agents.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9556–9567.
- Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In International Conference on Learning Representations (ICLR).

A Related Work

Recent advancements in large language models (LLMs) have led to the development of sophisticated language agents capable of assisting in various aspects of chemical research (Ramos et al., 2024). These agents, such as ChemCrow (M. Bran et al., 2024) and Coscientist (Boiko et al., 2023), have demonstrated the ability to automate routine chemical tasks and accelerate molecular discovery. ChemCrow, for instance, integrates LLMs with common chemical tools to perform a wide range of chemistry-related tasks, consistently outperforming GPT-4 (OpenAI, 2023) in accuracy. Similarly, Coscientist exemplifies the integration of semi-autonomous robots in planning and executing chemical reactions with minimal human intervention. Other notable agents include Chemist-X (Chen et al., 2024b), which focuses on designing chemical reactions to achieve specific molecules, and ProtAgent (Ghafarollahi and Buehler, 2024), a multi-agent system designed to automate and optimize protein design. In the realm of experimental planning, several agents have been developed to bridge the gap between virtual assistants and physical laboratory environments. CALMS (CHERUKARA et al., 2024) enhances laboratory efficiency by operating instruments and managing complex experiments through conversational LLMs. BioPlanner (O'Donoghue et al., 2023) improves experimental efficiency by creating pseudocode representations of procedures, while CRISPR-GPT (Huang et al., 2024) assists in designing gene editing experiments iteratively with constant human feedback. LLM-RDF (Ruan et al., 2024) takes this a step further by automating every step of the synthesis workflow, from literature search to product purification. Cheminformatics tasks have also been significantly impacted by LLM-based agents. CACTUS (McNaughton et al., 2024) automates the application of multiple cheminformatics tools while maintaining human oversight in molecular discovery. ChatMOF (Kang and Kim, 2023) focuses on predicting and generating Metal-Organic Frameworks, integrating MOF databases with its predictor module. IBM ChemChat augments LLMs with common APIs and Python packages used in cheminformatics research, facilitating tasks such as de novo drug design and property prediction. Most recently, Tang et al. (2025) concurrently proposed an agent framework with a dynamic and self-updating memory mechanism to improve LLMs in solving general chemistry questions, and Chen et al. (2024d) examine agents' performance in data-driven scientific discovery via coding and achieve better performance with domain knowledge insertion and self-debug (Chen et al., 2024c). These advancements collectively demonstrate the transformative potential of AI agents in chemical research, streamlining processes, enhancing efficiency, and accelerating scientific discovery.

Although these above agents have been proposed to tackle specific chemistry applications, there lacks a comprehensive evaluation on how tool-augmented agents perform on various chemistry problems. This study aims to address this issue and provide actionable insights to shed the light for future directions.

B Tool Set

B.1 Tools

The tool set contains 29 tools ranging from general tools, molecule tools, to reaction tools. This section introduces all the tools in detail.

General tools: Provide broad information retrieval, web searching, and computational.

- AiExpert: A general-purpose LLM prompted to answer any questions when other tools cannot handle. We use GPT-4o or Claude-3.5-Sonnet in our experiments, identical to the backbone models of ChemAgent.
- PythonREPL: Executes Python commands and allows for package installation.
- WebSearch: Searches the internet for both general and domain-specific information, providing concise summaries of relevant content. This involves an LLM-based search service² that uses LLMs to summarize the search result, providing more straightforward and organized results.
- WikipediaSearch: Searches Wikipedia and provides summaries of related content.

Molecule tools: Offer various analyses, predictions, and conversions related to chemical compounds and their properties.

²<https://tavily.com/>.

- **BBBPPredictor**: Predicts the probability of a compound penetrating the blood-brain barrier using the Uni-Mol model (Zhou et al., 2023).
 - **CanonicalizeSMILES**: Converts SMILES representation to its canonical form with RDKit (RDKit, 2023).
 - **CompareSMILES**: Determines if two molecule SMILES representations are identical.
 - **CountMolAtoms**: Counts the number and types of atoms in a molecule.
 - **FunctionalGroups**: Identifies functional groups present in a molecule.
 - **GetMoleculePrice**: Retrieves the cheapest available price for a purchasable molecule.
 - **HIVInhibitorPredictor**: Predicts the probability of a compound inhibiting HIV replication using the Uni-Mol model.
 - **IUPAC2SMILES**: Converts IUPAC names to SMILES representation by searching PubChem, ChemSpace, or using the neural network based STOUT model (Rajan et al., 2021).
 - **LogDPredictor**: Predicts the octanol/water distribution coefficient (logD) at pH 7.4 using the Uni-Mol model.
 - **MolSimilarity**: Computes the Tanimoto similarity between two molecules.
 - **MoleculeCaptioner**: Generates a textual description of a molecule using neural networks using the MolT5 model (Edwards et al., 2022).
 - **MoleculeGenerator**: Creates SMILES representations based on molecular descriptions using neural networks using the MolT5 model (Edwards et al., 2022).
 - **Name2SMILES**: Converts common molecule names to SMILES representation.
 - **PatentCheck**: Verifies if a molecule is patented.
 - **PubchemSearchQA**: Searches and retrieves molecule/compound information from PubChem, a comprehensive database of chemical molecules and their activities. Given the information of a molecule/compound (SMILES, IUPAC name, or common name) and a related question, it retrieves the corresponding document from PubChem, and applies an instructed LLM (GPT-4o in our experiments) to briefly answer the input questions. Instead of directly returning the whole document, which is typically very long, this QA design reduces the irrelevant information in the context, so as to avoid distractions and length limit violation for LLMs.
 - **SELFIES2SMILES**: Converts SELFIES (Krenn et al., 2019) to SMILES representation.
 - **SMILES2Formula**: Derives the molecular formula from SMILES representation using fixed algorithm implemented with RDKit.
 - **SMILES2IUPAC**: Converts SMILES representation to IUPAC name by searching PubChem, ChemSpace, or using the neural network based STOUT model.
 - **SMILES2SELFIES**: Converts SMILES representation to SELFIES representation.
 - **SMILES2Weight**: Calculates the molecular weight from SMILES representation.
 - **SideEffectPredictor**: Predicts the probabilities of a compound causing various side effects across 20 different categories using the Uni-Mol model.
 - **SolubilityPredictor**: Predicts the log solubility of a compound in mol/L using the Uni-Mol model.
 - **ToxicityPredictor**: Predicts the probability of a compound being toxic using the Uni-Mol model.
- Reaction tools**: Predict products of chemical reactions and suggest potential reactants for synthesizing given products.
- **ForwardSynthesis**: Predicts the products of a chemical reaction based on given reactants and reagents using IBM RXN for Chemistry³.
 - **Retrosynthesis**: Conducts single-step retrosynthesis, suggesting potential reactants for a given product using IBM RXN for Chemistry.

B.2 Improvements of Tool Set

We build an improved tool set based on that of ChemCrow (M. Bran et al., 2024). The purposes are:

- **Addressing non-functional tools**: Some tools in ChemCrow were not well-maintained and failed to function properly, so we fixed their bugs and made them function again. For example, RXNPredict

³<https://rxn.res.ibm.com/rxn>.

encountered a “project_id not set” error due to implementation issues, and we fixed this issue, resulting in our ForwardSynthesis tool.

- **Improving tool usability:** Certain tools did not meet practical needs, so we modified them to make them more practical and robust. For instance, the original web search tool in ChemCrow has limitations, such as inflexible input handling and truncation of output when results are lengthy, often cutting off critical information. We resolved this by replacing it with an LLM-based search service, which supports flexible input and generates summarized outputs.
- **Filling gaps in functionality:** The original tool set lacked support for many important chemistry tasks, so we created tools to enable it on more diverse tasks. For example, accessing information from PubChem, a comprehensive database of molecules and compounds, is not supported. To address this, we implemented the PubChemQA tool. Chemical representation conversion is another critical capability missing from the original tool set. To fill this gap, we added multiple tools for the conversion.

After these improvements, the tool set became significantly more robust and comprehensive. This is reflected in the performance comparisons in Table 2 and Table 3, where ChemAgent consistently and substantially outperforms ChemCrow that uses the original tool set.

C Experiment Details

C.1 SMolInstruct

C.1.1 Dataset and Evaluation Setups

Name Conversion	Property Prediction
<p>IUPAC to Molecular Formula (NC-I2F) Query: What is the molecular formula of the compound with this IUPAC name <IUPAC> 2,5-diphenyl-1,3-oxazole </IUPAC> ? Response: <MOLEFORMULA> C15H11NO </MOLEFORMULA></p> <p>IUPAC to SMILES (NC-I2S) Query: Could you provide the SMILES for <IUPAC> 4-ethyl-4-methylxolan-2-one </IUPAC> ? Response: Of course. It's <SMILES> CCC1(C)COC(=O)C1 </SMILES></p> <p>SMILES to Molecular Formula (NC-S2F) Query: Given the SMILES representation <SMILES> S-P1(N(CCC)CCC)NCCO1 </SMILES>, what would be its molecular formula? Response: It is <MOLEFORMULA> C7H15Cl2N2OPS </MOLEFORMULA> .</p> <p>SMILES to IUPAC (NC-S2I) Query: Translate the given SMILES formula of a molecule <SMILES> CCC(C)C1NCCCNC1 </SMILES> into its IUPAC name. Response: <IUPAC> 3-butan-2-yl-1,5-diazocane </IUPAC></p>	<p>ESOL (PP-ESOL) Query: How soluble is <SMILES> CC(C)Cl </SMILES> ? Response: Its log solubility is <NUMBER> -1.41 </NUMBER> mol/L.</p> <p>LIPO (PP-LIPO) Query: Predict the octanol/water distribution coefficient logD under the circumstance of pH 7.4 for <SMILES> NC(=O)C1=CC=CC=O </SMILES> . Response: <NUMBER> 1.090 </NUMBER></p> <p>BBBP (PP-BBBP) Query: Is blood-brain barrier permeability (BBBP) a property of <SMILES> CCNC(=O)/C=C/C1=CC=CC(Br)=C1 </SMILES>? Response: <BOOLEAN> Yes </BOOLEAN></p> <p>ClinTox (PP-ClinTox) Query: Is <SMILES> COC[C@@H](NC(C)=O)C(=O)NCC1=CC=CC=C1 </SMILES> toxic? Response: <BOOLEAN> No </BOOLEAN></p> <p>HIV (PP-HIV) Query: Can <SMILES> CC1=CN(C2C=CCCC2O)C(=O)NC1=O </SMILES> serve as an inhibitor of HIV replication? Response: <BOOLEAN> No </BOOLEAN></p> <p>SIDER (PP-SIDER) Query: Are there any known side effects of <SMILES> CC1=CC(C)=C(NC(=O)CN(CC(=O)O)CC(=O)O)C(C)=C1Br </SMILES> affecting the heart? Response: <BOOLEAN> No </BOOLEAN></p>
Molecule Description	Chemical Reaction
<p>Molecule Captioning (MC) Query: Describe this molecule: <SMILES> CCOC(=O)C1=CN=CN[C@@H](C)C1=CC=CC=C1 </SMILES> Response: The molecule is an imidazole derivative with short-acting sedative, hypnotic, and general anesthetic properties. Etomidate appears to have gamma-aminobutyric acid (GABA) like effects, mediated through GABA-A receptor. The action enhances the inhibitory effect of GABA on the central nervous system by causing chloride channel opening events which leads to membrane hyperpolarization.</p> <p>Molecule Generation (MG) Query: Give me a molecule that satisfies the conditions outlined in the description: The molecule is a member of the class of tripyrroles that is a red-coloured pigment with antibiotic properties produced by Serratia marcescens. It has a role as an antimicrobial agent, a biological pigment, a bacterial metabolite, an apoptosis inducer and an antineoplastic agent. It is a tripyrrole, an aromatic ether and a ring assembly. Response: Here is a potential molecule: <SMILES> CCCCC1=C(C)NC(C=C2N=C(C3=CC=CN3)C=C2OC)=C1 </SMILES></p>	<p>Forward Synthesis (FS) Query: <SMILES> NC1=CC=C2OCOC2=C1.O=CO </SMILES> Based on the reactants and reagents given above, suggest a possible product. Response: A possible product can be <SMILES> O=CNC1=CC=C2OCOC2=C1 </SMILES></p> <p>Retrosynthesis (RS) Query: Identify possible reactants that could have been used to create the specified product. <SMILES> CC1=CC=C(N)N=C1N </SMILES> Response: <SMILES> CC(C#N)CC#N.N </SMILES></p>

Figure C.1: Tasks in SMolInstruct (Yu et al., 2024).

SMolInstruct (Yu et al., 2024) contains 14 molecule- and reaction-centric tasks, which, along with the task name abbreviations and examples, are illustrated in Figure C.1.

We evaluate the models on 50 randomly selected samples from the test set for each task. For reference, in the following detailed results, we also include the SoTA non-LLM models used in Yu et al. (2024), and LlaSMol⁴, which is a Mistral model (Jiang et al., 2023) fine-tuned on SMolInstruct. For SoTA non-LLM models and LlaSMol, we adopt their own formats of input and output. For other models, we prompt

⁴osunlp/LlaSMol-Mistral-7B.

them to think step by step, i.e., using chain-of-thought (CoT) (Wei et al., 2022; Wang et al., 2022), and wrap their final answers with “<ANSWER>” and “</ANSWER>” to facilitate answer extraction. The evaluation metrics are adopted from Yu et al. (2024).

C.1.2 Detailed Results

Model	NC					PP						
	I2F		I2S		S2F	S2I	ESOL	Lipo	BBBP	Clintox	HIV	SIDER
	EM	EM	Valid	EM	EM	RMSE↓	RMSE↓	Acc	Acc	Acc	Acc	
SoTA non-LLM models	96.0	68.0	100.0	100.0	54.0	0.808	0.527	88.0	90.0	94.0	70.0	
GPT-4o	12.0	0.0	66.0	8.0	0.0	1.315	1.264	70.0	36.0	86.0	44.0	
Claude-3.5-Sonnet	4.0	10.0	70.0	4.0	2.0	1.443	1.267	78.0	50.0	88.0	62.0	
LlaSMol	92.0	60.0	96.0	96.0	34.0	1.062	1.164	82.0	98.0	94.0	74.0	
ChemCrow (GPT)	18.0	10.0	18.0	88.0	2.0	4.376	2.061	46.0	62.0	74.0	36.0	
ChemCrow (Claude)	16.0	14.0	18.0	42.0	2.0	2.025	1.179	60.0	34.0	92.0	32.0	
ChemAgent (GPT)	100.0	64.0	100.0	100.0	70.0	0.812	0.529	90.0	82.0	94.0	70.0	
ChemAgent (Claude)	100.0	68.0	100.0	100.0	70.0	1.131	0.531	90.0	58.0	92.0	68.0	

Table C.1: The results on SMolInstruct for name conversion and property prediction tasks. All the metrics except RMSE are in percentage.

Model	MC	MG			FS			RS		
	METEOR	EM	FTS	Valid	EM	FTS	Valid	EM	FTS	Valid
SoTA non-LLM models	0.539	32.0	75.7	96.0	78.0	91.7	100.0	42.0	80.5	100.0
GPT-4o	0.152	10.0	57.5	84.0	12.0	46.3	84.0	0.0	36.0	84.0
Claude-3.5-Sonnet	0.211	12.0	67.5	90.0	22.0	60.9	98.0	0.0	45.7	90.0
LlaSMol	0.426	22.0	67.0	98.0	56.0	83.4	100.0	26.0	70.3	100.0
ChemCrow (GPT)	0.195	34.0	79.9	68.0	72.0	92.5	92.0	8.0	49.0	74.0
ChemCrow (Claude)	0.255	40.0	81.0	86.0	70.0	90.5	92.0	22.0	0.0	90.0
ChemAgent (GPT)	0.510	28.0	76.8	90.0	78.0	92.1	98.0	42.0	78.0	98.0
ChemAgent (Claude)	0.443	44.0	83.5	100.0	80.0	92.2	100.0	42.0	78.6	100.0

Table C.2: The results on SMolInstruct for the MC, MG, FS, and RS tasks. All the metrics except METEOR score are in percentage.

The detailed results on SMolInstruct are presented in Table C.1 and Table C.2. We can see that: (1) The SoTA LLMs, GPT-4o and Claude-3.5-Sonnet, demonstrate relatively low performance across all evaluated tasks, which underscores the persistent challenges faced by general-purpose LLMs in specialized chemistry domains, particularly in handling molecular representations such as SMILES and executing specialized chemical operations. (2) On all tasks, ChemAgent achieves the best performance or close, confirming the benefits of specialized tools for the SMolInstruct tasks. (3) While Claude-3.5-Sonnet generally outperforms GPT-4o, their performance as ChemAgent backbones is comparable. This parity can be attributed to the nature of the SMolInstruct tasks, which primarily require effective tool utilization rather than extensive knowledge or complex reasoning abilities inherent to the LLMs themselves. Both LLMs demonstrate proficiency as “tool users,” effectively leveraging the provided resources to address the given tasks.

C.1.3 Potential Data Leakage via Tools

To ensure a fair evaluation on SmolInstruct (Yu et al., 2024), we account for potential data leakage in our experiments.

For tools based on our self-trained models, we mitigated data leakage by training exclusively on samples from the SmolInstruct training set. This applies to MoleculeCaptioner, MoleculeGenerator, SolubilityPredictor, LogDPredictor, BBBPPredictor, ToxicityPredictor, HIVInhibitorPredictor, and SideEffectPredictor. Since test examples were excluded from the training set, these tools cannot leak test data.

For ForwardSynthesis and Retrosynthesis, which rely on IBM’s APIs, determining the original training data of their backend models (Molecular Transformer (Schwaller et al., 2019)) is challenging, making potential data leakage uncertain. However, the SmolInstruct paper retrained Molecular Transformer on the training set and reported test set performance consistent with ChemAgent’s results. This suggests that if data leakage exists, it is minimal; otherwise, ChemAgent’s performance would significantly surpass the SmolInstruct benchmarks.

C.2 MMLU-Chemistry

C.2.1 Dataset and Evaluation Setups

To effectively and efficiently evaluate the models, we build MMLU-Chemistry, a subset of 70 chemistry question samples derived from the widely-used MMLU dataset (Hendrycks et al., 2021). Specifically, to increase the difficulty and differentiation of the questions, while avoiding erroneous samples presented in the original MMLU, we select samples that appear in both MMLU-Pro (Wang et al., 2024b) and MMLU-Redux (Gema et al., 2024). These two datasets are verified versions of MMLU, and MMLU-Pro has extended the answer options from 4 to 10 to introduce more challenges. When the gold standard answers from both sources match, we utilize the 10 options from MMLU-Pro. In cases of discrepancies, we manually review and correct any potential issues. To reduce the cost of evaluation, we eliminated samples where all models performed correctly in our preliminary experiments. This results in a final set of 70 questions, divided evenly between 35 high school-level and 35 college-level questions.

In our evaluation, all the models are prompted to generate a CoT solution and close the solution with “the answer is ...” to facilitate answer extraction. To mitigate randomness, we run each sample three times and report the average accuracy.

In addition, to understand the influence of in-context examples, in the following detailed results, we also introduce a 5-shot setting in comparison with 0-shot for the base LLMs and ChemAgent. The questions of the in-context examples are originally from MMLU’s and MMLU-Pro’s development set, and we manually construct CoT solutions for the base LLMs and tool-using step-wise solutions for ChemAgent. The order of the examples is randomized for each test sample.

C.2.2 Detailed Results

Model	High school	College	Overall
GPT-4o (0-shot)	88.6	72.4	80.5
GPT-4o (5-shot)	85.7	72.4	79.0
Claude-3.5-Sonnet (0-shot)	83.8	69.5	76.7
Claude-3.5-Sonnet (5-shot)	83.8	73.3	78.6
ChemCrow (GPT, 0-shot)	47.6	39.0	43.3
ChemCrow (Claude, 0-shot)	69.5	67.6	68.6
ChemAgent (GPT, 0-shot)	81.9	60.0	71.0
ChemAgent (GPT, 5-shot)	87.6	63.8	75.7
ChemAgent (Claude, 0-shot)	73.3	66.7	70.0
ChemAgent (Claude, 5-shot)	86.7	79.0	82.9

Table C.3: Accuracies (%) on MMLU-Chemistry, averaged over three runs.

The results are presented in Table C.3. (1) While ChemAgent achieves the highest overall performance in one specific configuration (Claude, 5-shot), it demonstrates inferior performance compared to the base LLMs in all other configurations. This trend persists across both high school and college questions and is also observed with ChemCrow, suggesting a consistent pattern rather than an isolated occurrence. (2) Comparing 0-shot and 5-shot performance, the addition of examples (5-shot) yields minimal improvement for base LLMs but results in significant enhancement for ChemAgent. This disparity may be attributed to the extensive pre-training of base LLMs on general chemistry questions, potentially rendering additional examples redundant for task comprehension. Conversely, for ChemAgent, the step-wise demonstration examples appear to effectively guide the LLMs in reasoning and tool utilization and potentially reduce the

cognitive overload for LLMs, thereby optimizing the problem-solving process. This finding suggests that incorporating examples can be a valuable strategy for enhancing the performance.

C.3 GPQA-Chemistry

We use GPQA-Chemistry, the 93 chemistry multi-choice questions from the expert-verified GPQA-Diamond subset of the GPQA dataset (Rein et al., 2023) to evaluate models' abilities in high-level chemistry knowledge and complex reasoning. All the evaluated models are prompted to generate CoT solutions and close their output with "the answer is ..." to facilitate answer extraction. We report the average accuracy across 3 runs. The results are presented in Table 3.

D Error Cases

This section presents a specific case of each error. For reasoning errors, the cases are shown in Appendix D.1, Appendix D.2, Appendix D.3, and Appendix D.4. For grounding errors, the case is presented in Appendix D.5. For tool errors, the cases are presented in Appendix D.6 and Appendix D.7.

D.1 Reasoning: Wrong Knowledge/Logic and Incomplete Reasoning

Explanation

This is an example of two reasoning errors: wrong knowledge/logic and incomplete reasoning. The question is from MMLU-Chemistry, and ChemAgent answers it by firstly calculating the amount of the reactants in moles with the PythonREPL tool and then analyzing it based on its internal knowledge. The calculation with PythonREPL is correct, but the **highlighted** conclusion is incorrect. While the agent concludes that the limiting reactant is bromine, the rubidium is actually the limiting reactant in this case. Additionally, the model does not verify whether the amounts of the remaining reactants and the product are larger than 0.10 mg, showcasing an incomplete reasoning.

Question

If 1.0 g of rubidium and 1.0 g of bromine are reacted, what will be left in measurable amounts (more than 0.10 mg) in the reaction vessel?

- (A) RbBr only (B) RbBr, Rb, Br₂, and Rb₂Br (C) RbBr and Rb₂Br only
(D) RbBr, Rb, and Br₂ (E) Rb and Br₂ only (F) Br₂ only
(G) RbBr and Rb only **[predicted]** (H) Rb only
(I) Nothing will be left in measurable amounts (J) RbBr and Br₂ only **[Correct]**

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

To determine what substances will be left in measurable amounts, I need to conduct a stoichiometric calculation based on the reaction between rubidium (Rb) and bromine (Br₂) to form rubidium bromide (RbBr). First, I'll calculate the moles of rubidium and bromine to verify which one is the limiting reagent. Then I will determine whether any excess reactant is left after the reaction. Given:

- Atomic weight of Rubidium (Rb) = 85.47 g/mol
- Molecular weight of Bromine (Br₂) = 159.808 g/mol (since Br = 79.904 g/mol)

Let's start by calculating the moles of rubidium and Bromine.

Action

Tool: PythonREPL
Tool input:

```
1 rubidium_molar_mass = 85.47 # g/mol for Rubidium (Rb)
2 bromine_molar_mass = 159.808 # g/mol for Bromine (Br2)
3 # Calculate moles
4 rubidium_moles = 1.0 / rubidium_molar_mass
5 bromine_moles = 1.0 / bromine_molar_mass
6 rubidium_moles, bromine_moles
```

Observation

Tool output: (0.0117000117000117, 0.006257509010812976)

Step 2

Thought

The moles of rubidium are approximately 0.0117 mol, and the moles of bromine are approximately 0.0063 mol. The reaction between rubidium and bromine to form rubidium bromide (RbBr) is as follows: $2\text{Rb} + \text{Br}_2 \rightarrow 2\text{RbBr}$. From the stoichiometry of the reaction, 2 moles of rubidium react with 1 mole of bromine. Therefore, **bromine is the limiting reagent** since 0.0117 moles of rubidium would require 0.00585 moles of bromine (less than 0.0063 moles provided). Since bromine is the limiting reagent, it will be completely consumed, while excess rubidium will remain. Now I can determine what substances will be present in measurable amounts (> 0.10 mg) based on the stoichiometric calculation:

1. Bromine will be fully consumed.
2. Rubidium bromide (RbBr) will be formed.
3. Excess rubidium will remain unreacted.

Therefore, RbBr and excess Rb will be present in measurable amounts.

Answer

The answer is (G) RbBr and Rb only.

D.2 Reasoning: Wrong Final Answer

Explanation

This is an example of a reasoning error: wrong final answer. The question is from MMLU-Chemistry, and the agent answers it with the LLM's internal knowledge and no tool. Although it clearly rules out option (B) as **highlighted** in its analysis, it still selects (B) as the final answer, which is wrong.

Question

Which of the following is a true statement about optical isomerism of complexes containing achiral ligands?

- (A) Square planar complexes can display optical isomerism only if all four ligands are identical.
- (B) Tetrahedral complexes never display optical isomerism. **[predicted]**
- (C) Linear complexes can display optical isomerism when both ligands are different.
- (D) Octahedral complexes of monodentate ligands can display optical isomerism only when they have at least three different ligands. **[correct]**

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

To answer this question, we need to understand the coordination geometries and when optical isomerism can occur in complexes. Let's briefly consider each statement:

- (A) Square planar complexes: Optical isomerism occurs when there is no plane of symmetry in the molecule. Square planar complexes typically don't show optical isomerism unless they're part of a larger chiral structure or have different ligands.
- (B) Tetrahedral complexes: **Tetrahedral complexes can exhibit optical isomerism** if they have four different ligands, forming a chiral center analogous to asymmetric carbon in organic chemistry.
- (C) Linear complexes: Linear complexes generally don't exhibit optical isomerism, as they require stereocenters which are not possible with only two ligands.
- (D) Octahedral complexes: Octahedral complexes can show optical isomerism when they have a specific arrangement of different ligands that break symmetry and create chirality.

The only true statement regarding optical isomerism in complexes that could contain achiral ligands is (B).

Answer

The answer is (B).

D.3 Reasoning: Information Oversight

Explanation

This is an example of a reasoning error: information oversight. The question is from MMLU-Chemistry, and the agent answers it by using PythonREPL to calculate the result. However, the model ignores "13C" in the question, which indicates that the frequency should be approximately 1/4 of the frequency of the NMR instrument. This oversight leads to the wrong final answer.

Question

The ^{13}C chemical shifts of benzene and chloroform are 128.4 ppm and 77.2 ppm respectively. What is the difference in the ^{13}C NMR frequencies of the two compounds on a 600 MHz spectrometer?

- (A) 7.73 kHz [correct] (B) 60.3 kHz (C) 122 kHz (D) 500 kHz (E) 15.4 kHz
(F) 30.7 kHz [predicted] (G) 183.2 kHz (H) 367.2 kHz (I) 91.6 kHz (J) 244 kHz

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

To find the difference in frequencies, we need to calculate the frequency difference corresponding to the chemical shift difference. This can be calculated using the formula: frequency difference (in Hz) = chemical shift difference (in ppm) \times spectrometer frequency (in MHz).

The chemical shift difference between benzene and chloroform is (128.4 ppm - 77.2 ppm), and the spectrometer frequency is 600 MHz. Let's calculate the frequency difference in kHz.

Action

Tool: PythonREPL

Tool input:

```
1 chemical_shift_difference = 128.4 - 77.2
2 spectrometer_frequency_mhz = 600
3 # Calculating the frequency difference in kHz
4 frequency_difference_khz = chemical_shift_difference *
   spectrometer_frequency_mhz
5 frequency_difference_khz
```

Observation

Tool output: 30720.0

Step 2

Thought

The frequency difference between benzene and chloroform on a 600 MHz spectrometer is 30.72 kHz. This corresponds to option (F) 30.7 kHz.

Answer

The answer is (F).

D.4 Reasoning: Algebra Error

Explanation

This is an example of a reasoning error: algebra error. The question is from MMLU-Chemistry, and the agent answers it by firstly using Name2SMILES to understand the molecular structures and then calculating the ratio on its own. Everything is correct until the **highlighted** algebra calculation step fails to correctly cross-multiply to solve for the ratio, leading to the wrong final answer.

Question

The ^1H spectrum of a mixture of dimethylsulphoxide (DMSO) and acetonitrile (AN) contains lines with relative intensities α and 3α , respectively. What is the ratio of the two concentrations, $[\text{DMSO}]:[\text{AN}]$?

- (A) 3:2 (B) 1:6 [correct] (C) 1:9 [predicted] (D) 1:1 (E) 2:3
(F) 1:3 (G) 3:1 (H) 2:1 (I) 6:1 (J) 1:2

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

The relative intensities in the ^1H NMR spectrum of the mixture of DMSO and acetonitrile are given as α and 3α , respectively. To find the concentration ratio $[\text{DMSO}]:[\text{AN}]$, the relative number of nuclei contributing to these intensities must be considered. Both DMSO and acetonitrile contribute distinct numbers of protons. I will determine the number of protons contributing to the NMR spectrum for each compound.

Action

Tool: Name2SMILES
Tool input: Dimethylsulphoxide

Observation

Tool output: CS(=O)C

Step 2

Thought

Dimethylsulphoxide (DMSO) has the SMILES representation CS(=O)C, meaning it has 6 equivalent hydrogen atoms (as CH_3 groups contribute their protons). Now, I need to find out the SMILES for acetonitrile (AN) to determine its number of hydrogens.

Action

Tool: Name2SMILES
Tool input: Acetonitrile

Observation

Tool output: CC#N

Step 3

Thought

Acetonitrile (AN) has the SMILES representation CC#N, meaning it has 3 equivalent hydrogen atoms (from the CH_3 group). Dimethyl sulfoxide (DMSO) thus contributes 6 protons to the NMR, whereas acetonitrile (AN) contributes 3 protons. The NMR intensity is proportional to the number of protons, thus the ratio of concentrations can be calculated considering this proportionality.

- The intensity ratio given in the NMR spectrum is α (for DMSO) and 3α (for AN).
- Let the concentration of DMSO be $[\text{DMSO}]$ and that of AN be $[\text{AN}]$.

Based on proportionality:

$$\frac{[\text{DMSO}] \times 6}{[\text{AN}] \times 3} = \frac{\alpha}{3\alpha}$$

$$\frac{[\text{DMSO}]}{[\text{AN}]} = \frac{1}{9} \quad (1)$$

The concentration ratio of DMSO to AN is therefore 1:9.

Answer

The answer is (C) 1:9.

D.5 Grounding: Wrong Input Format

Explanation

This is an example of a grounding error: wrong input format. This question is from SMolInstruct, and the agent initially calls the dedicated tool, SideEffectPredictor, to predict the potential side effect. However, it inputs an extra "SMILES:" (highlighted), causing the tool to malfunction. The same issue occurs with the CanonicalizeSMILES tool.

Question

Are there any known side effects of <SMILES>CC(C)(C)NC[C@H](O)COC1=NSN=C1N1CCOCC1.CCN[C@H]1CN(CCCOC)S(=O)(=O)C2=C1C=C(S(N)(=O)=O)S2 </SMILES> affecting the hepatobiliary system?

Correct: No. **Predicted:** Yes.

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

To determine if there are any known side effects of the given compound affecting the hepatobiliary system, I will first use the SideEffectPredictor tool to obtain the probabilities of different side effects caused by the compound. After that, I will specifically look for any effects related to the hepatobiliary system.

Action

Tool: SideEffectPredictor

Tool input: **SMILES:** CC(C)(C)NC[C@H](O)COC1=NSN=C1N1CCOCC1.CCN[C@H]1CN(CCCOC)S(=O)(=O)C2=C1C=C(S(N)(=O)=O)S2

Observation

Tool output: Error: Invalid SMILES.

Step 2

Thought

The error indicates that the provided SMILES is invalid. I'll canonicalize the SMILES and then re-attempt using the SideEffectPredictor tool.

Action

Tool: CanonicalizeSMILES

Tool input: **SMILES:** CC(C)(C)NC[C@H](O)COC1=NSN=C1N1CCOCC1.CCN[C@H]1CN(CCCOC)S(=O)(=O)C2=C1C=C(S(N)(=O)=O)S2

Observation

Tool output: Error: Invalid SMILES.

Step 2

Thought

Since there's an issue with the SMILES provided and attempts to canonicalize it failed, I'll use the AiExpert tool to get an analysis regarding the potential side effects based on the description of the molecule.

Action

Tool: AiExpert

Tool input: Are there any known side effects affecting the hepatobiliary system for a compound with both isothiazolone and sulfonamide-like structures?

Observation

Tool output: Compounds that contain isothiazolone and sulfonamide-like structures can be complex in nature ... (omit the long analysis)

Given the potential for systemic toxicity, especially with ongoing exposure, caution is warranted, and hepatobiliary monitoring should be part of safety assessments for such a compound. Always consult and follow established guidelines and protocols for safety evaluation during drug development or chemical utilization phases.

Step 3

Thought

The AiExpert tool suggests there could be potential hepatobiliary side effects for compounds containing isothiazolone and sulfonamide-like structures due to possible hepatotoxicity, liver enzyme changes, or hypersensitivity reactions involving the liver. Therefore, there's a likelihood of such effects occurring in the given compound.

Answer

Yes.

D.6 Tool: Wrong Tool Output

Explanation

This example shows the tool: wrong tool output error. The question is from MMLU-Chemistry, and the agent answers it by searching the internet with the WebSearch tool. The tool provides an incorrect result (**highlighted**), which leads to the wrong final answer.

Question

Of the following atoms, which has the lowest electron affinity?

(A) F (B) Si [**predicted**] (C) O (D) Ca [**gold**]

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

Electron affinity generally increases across a period and decreases down a group in the periodic table. Considering this trend, I need to examine the electron affinities of the given atoms: F, Si, O, and Ca. I will search for electron affinities to accurately determine which atom has the lowest value.

Action

Tool: WebSearch
Tool input: Which atom has the lowest electron affinity among F, Si, O, and Ca?

Observation

Tool output: **The atom with the lowest electron affinity among F, Si, O, and Ca is Silicon (Si).** It has an electron affinity value of 134.06 kJ/mol, which is lower compared to the other atoms mentioned.

Step 2

Thought

Based on the electron affinity values obtained, Silicon (Si) has the lowest electron affinity among the listed options.

Answer

The answer is (B) Si.

D.7 Tool: Inconsistent Tool Outputs

Explanation

This is an example of a tool error: inconsistent tool outputs. The question is from SMolInstruct, and the agent answers it by predicting the answer with ToxicityPredictor and verifying the answer with PubchemSearchQA. However, the two tools give different answers and the agent chooses to believe the latter, which leads to the wrong final answer.

Question

Is `<SMILES>C[C@H](/C=C/[C@H](C)C(C)(C)O[C@H]1CC[C@H]2/C(=C/C=C3C[C@@H](O)C[C@H](O)C3)CCC[C@]12C</SMILES>` classified as a hazardous substance with documented toxicity?

Correct: No. **Predicted:** Yes.

Problem Solving Process of ChemAgent (GPT)

Step 1

Thought

To determine if the given SMILES represents a hazardous substance with documented toxicity, I should first predict its toxicity profile using the ToxicityPredictor tool.

Action

Tool: ToxicityPredictor
Tool input: C[C@H](/C=C/[C@H](C)C(C)(C)O)[C@H]1CC[C@H]2/C(=C/C=C3[C@@H](O)C[C@H](O)C3)CCC[C@]12C

Observation

Tool output: The probability of the compound to be toxic is 7.85%, which means it's unlikely to happen. Note that the result is predicted by a neural network model and may not be accurate. You may use other tools or resources to obtain more reliable results if needed.

Step 2

Thought

The ToxicityPredictor tool suggests that the probability of this compound being toxic is relatively low (7.85%), indicating it is unlikely to be classified as a hazardous substance based on this prediction alone. However, considering the neural network model's limitations, it would be prudent to verify this against more authoritative data sources like PubChem for documented toxicity.

Action

Tool: PubchemSearchQA
Tool input: SMILES: C[C@H](/C=C/[C@H](C)C(C)(C)O)[C@H]1CC[C@H]2/C(=C/C=C3[C@@H](O)C[C@H](O)C3)CCC[C@]12C
Question: Is this compound classified as a hazardous substance with documented toxicity?

Observation

Tool output: Yes, paricalcitol is classified as a hazardous substance with documented toxicity. It has several GHS hazard classifications, including being fatal if swallowed (H300) and toxic if swallowed or in contact with skin (H301+H311). Additionally, it is suspected of causing cancer (H351) and may cause harm to breast-fed children (H362). Paricalcitol has acute toxicity both orally and dermally, and is recognized for causing damage to organs through prolonged or repeated exposure (H372).

Furthermore, toxicological information indicates potential drug-induced liver injury (ambiguous concern), and toxicity symptoms such as hypercalcemia can be severe. It also shows potential for interactions that could exacerbate its toxic effects. These factors collectively contribute to its classification as a hazardous substance with significant documented toxicity.

Step 3

Thought

Based on the information obtained from PubChem, the compound represented by the given SMILES, identified as paricalcitol, is indeed classified as a hazardous substance with documented toxicity. It has several GHS hazard classifications and potential health risks.

Answer

Yes.

E Tool Usage Analysis

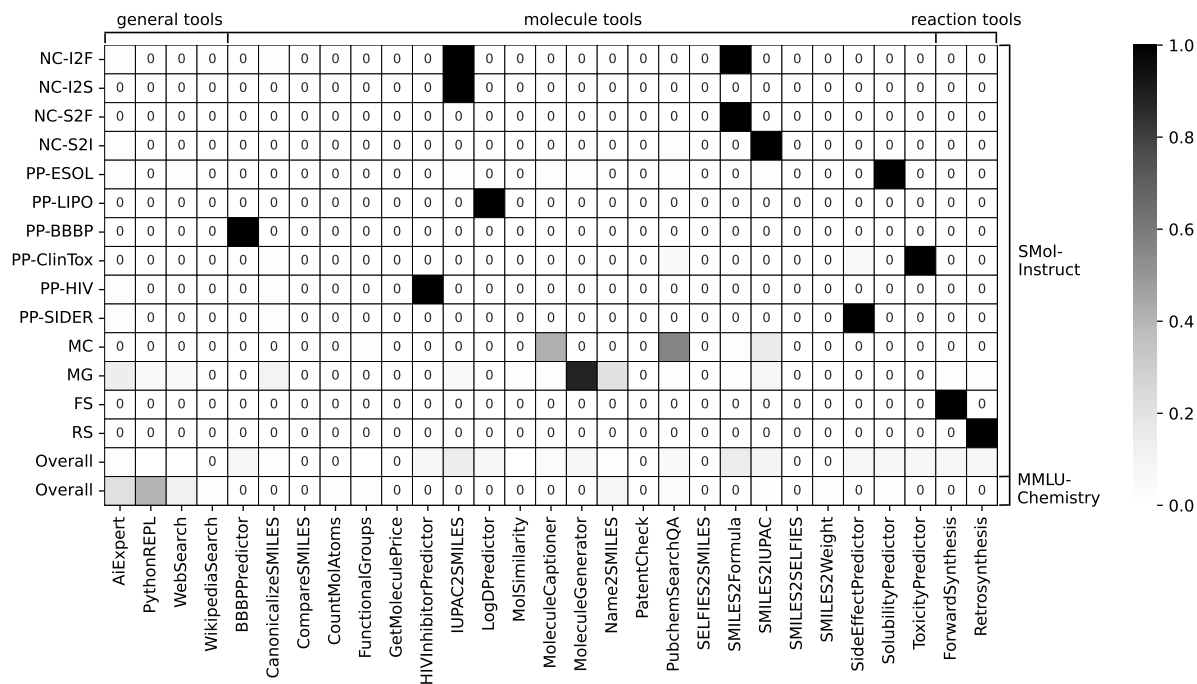


Figure E.2: The statistics of tool usage by ChemAgent (GPT). The cell values represent the ratios of number of samples where the corresponding tools are used out of the total number of samples, and the number “0” indicates the value is 0 (the tool is not used).

To examine the tool usage by the agent, we perform a statistical analysis of the tools used on both SMolInstruct and MMLU-Chemistry, as depicted in Figure E.2.

On SMolInstruct, the agent predominantly invokes the tools designed for the tasks. For instance, in the NC-I2F task, the agent commonly utilizes IUPAC2SMILES and SMILES2Formula to facilitate the conversion from IUPAC names to molecular formulas, with these tools often achieving values near 1.0. The molecular captioning (MC) and molecular generation (MG) tasks are notable exceptions. Since they are more open-ended, the agent opts for a variety of tools.

Conversely, on MMLU-Chemistry, the agent typically resorts to general-purpose tools (e.g., Python-REPL for calculations, WebSearch for knowledge gathering), due to the nature of the questions and the absence of task-specific tools.