ChEmREF: Evaluating Language Model Readiness for Chemical Emergency Response

Risha Surana **Qinyuan Ye** **Swabha Swayamdipta**
University of Southern California
{rsurana,qinyuany,swabhas}@usc.edu

Abstract

Emergency responders managing hazardous material (HAZMAT) incidents face critical, time-sensitive decisions, manually navigating extensive chemical guidelines. We investigate whether today's language models can assist responders by rapidly and reliably understanding critical information, identifying hazards, and providing recommendations. We introduce the Chemical Emergency Response Evaluation Framework (ChEmREF), a new benchmark comprising questions on 1,035 HAZMAT chemicals from the Emergency Response Guidebook and the PubChem Database. ChEmREF is organized into three tasks: (1) translation of chemical representation between structured and unstructured forms (*e.g.*, converting "C_2H_6O" to "ethanol"), (2) emergency response generation (*e.g.*, recommending appropriate evacuation distances) and (3) domain knowledge question answering from chemical safety and certification exams. Our best evaluated models received an exact match of 68.0% on unstructured HAZMAT chemical representation translation, a LLM Judge score of 52.7% on incident response recommendations, and a multiple-choice accuracy of 63.9% on HAZMAT examinations. These findings suggest that while language models show potential to assist emergency responders in various tasks, they require careful human oversight due to their current limitations.

1 Introduction

In a hazardous material (HAZMAT) incident, "the decisions made and actions taken in the first few minutes of a response will often establish the character of the overall response – and ultimately its success or failure [U.S. HHS, 2024b]." During the critical "golden hour," first responders must quickly assess the nature and scale of the incident, establish safety protocols, and request resources while protecting themselves and the public. In current practice, first responders rely extensively on the HAZMAT Emergency Response Guidebook (**ERG**; U.S. DoT, 2024)[1]: a printed manual of 400 pages that links common HAZMAT chemicals with their corresponding incident response guidelines. A major challenge in emergency response is navigating the extensive content of the ERG quickly and accurately to identify the appropriate response measures. In addition, first responders are often required to draw inferences from environmental cues, placards, databases, container labels, victim symptoms, and other contextual information. This adds complexity to the task, demanding skills and interdisciplinary knowledge that were not explicitly documented in the ERG.

Language models (LMs) today have acquired extensive knowledge through large scale pre-training and can synthesize vast amounts of information to generate responses that adhere to complex and dynamic contexts. Given this progress, we ask: can language models assist chemical emergency responders in HAZMAT incidents by automating parts of their workflow? We believe that the broad knowledge capacity, strong information synthesize ability and accessible interface of LMs open up

[1] See §C for a glossary of important terms related to HAZMAT and emergency response.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

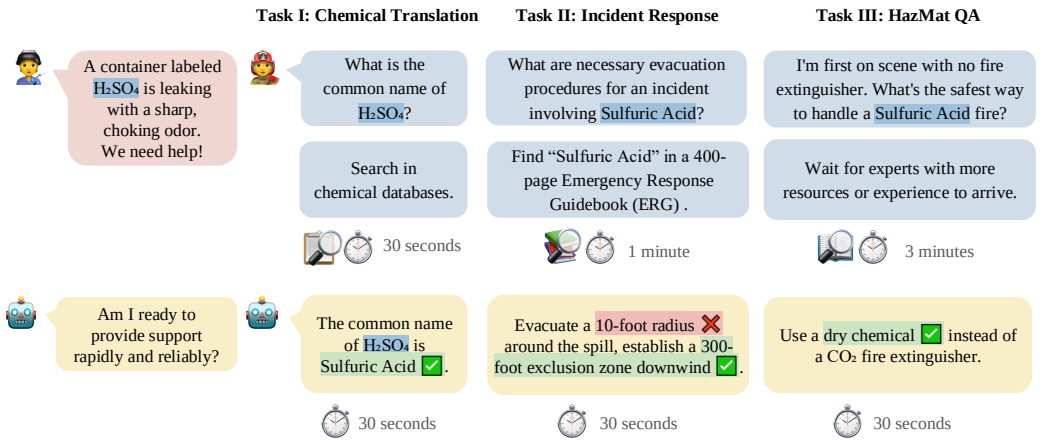


Figure 1: **LLM Assistants in Emergency Response.** In HAZMAT emergencies, timely decisions during scene size-up are critical [EMT Training Base, 2019]. This figure compares the traditional chemical response timeline with one aided by an LLM. While LLMs may accelerate decision-making, they can also produce incorrect guidance. These dynamics informed the design of ChEmREF.

Task I: Chemical Representation Translation (§A.1)
Question: What is the SMILES format for $C_4H_{10}O$? *Answer:* CC(C)(C)O
Task II: Emergency Incident Response (§A.2)
Scenario: A leak of *chlorine gas* has occurred near a populated area. Describe immediate public safety and evacuation steps necessary to manage the situation effectively. *Recommendation:* Stage emergency vehicles 300 feet away from the edge of the spill site...
Task III: HAZMAT Examination (§A.3)
Question: When transporting hazardous materials, when must the driver check the vehicle's tires? *Answer Choices:* A) Only at the start of the trip. B) Every 150 miles or every three hours, whichever comes first. C) Only if the tire pressure warning light turns on. D) After reaching the destination. *Answer:* B) Every 150 miles or every three hours, whichever comes first.

Table 1: **Overview of Three Core Tasks in ChEmREF (Section 2).** This table provides representative examples for each of the three core tasks evaluated in our framework.

new possibilities for supporting this task. Additionally, their fast information processing speed may help save valuable time in high-pressure, time-critical decision-making scenarios.

To this end, we introduce ChEmREF: a framework for evaluating whether LMs have the capabilities to assist first responders in chemical emergencies. Our framework consists of three tasks: (1) Chemical Representation Translation, which tests whether models can accurately convert between different unstructured (*e.g.*, UN-Number, Common Name, and Synonyms) or structured (*e.g.*, Molecular Formula, InChI, SMILES, and IUPAC Name[2]) chemical representations (see Table 3); (2) Incident Response Recommendation, which measures the ability of LMs to generate relevant safety measures, evacuation distances, and containment strategies based on incident context; and (3) HAZMAT Examination, which assesses performance on general-knowledge multiple-choice questions from real-world certification exams for first responders and chemical safety tests. We present abbreviated examples of each tasks in Table 1. By covering 1,035 HAZMAT chemicals, 8 chemical representation types, 6 emergency response dimensions and 6 HAZMAT exam categories, ChEmREF provides a comprehensive evaluation of LMs in this high-stakes domain.

We evaluate six recent general-purpose and domain-specific large language models on ChEmREF Overall, our results suggest that language models show potential to assist first responders in HAZMAT emergencies, but must be used with great caution and under the supervision of trained professionals,

[2]IUPAC Name is set by the International Union of Pure and Applied Chemistry. InChI refers to International Chemical Identifier. SMILES refers to Simplified Molecular Input Line Entry System.

given the high-stakes nature of these situations. We will publicly release the data and code for ChEmREF and hope that future work will build on it to advance research in this critical domain.

2 Evaluating LMs on ChEmREF

In this section, we first introduce tasks in ChEmREF and their task-specific prompting strategies in §2.1. We describe experiment details such as evaluated models and evaluation metrics in §2.2. Finally, we report model performance on ChEmREF and summarize our key observations in §2.3.

2.1 Tasks in ChEmREF

Task I: Chemical Representation Translation. Task I comprises two settings: structured and unstructured translation (described in detail in subsection A.1). We conducted a pilot study (Appendix G.1.1) to determine the most suitable prompt format for this task.

For **structured** translation, we adopt **one-shot chain-of-thought prompting**, as intermediate reasoning steps often aid in breaking down complex chemical terms (*e.g.*, translating "sodium" to "Na" and "superoxide" to "O_2"). For **unstructured** translation, we employ **direct zero-shot prompting**, as we observed no consistent improvement from including demonstrations or reasoning steps. We provide prompt examples in Appendix E.

Task II: Incident Response Recommendation. We evaluate models on Task II with **direct 0-shot prompting** that queries models to provide recommendations in one of the six key dimensions (*e.g.*, fire or explosion hazards, public safety). We provide an example prompt in Appendix E.

Task III: HAZMAT Examination. Task III involves answering multiple-choice questions with inputs presented in a standardized format: *Question: ... Answer Choices: ... Answer:*. We use **direct zero-shot prompting** without examples or intermediate reasoning.

2.2 Experiment Details

Evaluated Models. We evaluate five open-weight LMs on ChEmREF: three general-purpose (PHI-3 3.8B [Abdin et al., 2024a], PHI-4 14B [Abdin et al., 2024b], LLAMA-3.1 70B [Grattafiori et al., 2024]) and two domain-specialized (CHEMLLM 7B [Zhang et al., 2024], MED42 70B [Christophe et al., 2024]). CHEMLLM is SFT from INTERN-LM-2 7B [Cai et al., 2024], and MED42 from LLAMA-2 70B [Touvron et al., 2023]. We also include GPT-4O [Hurst et al., 2024] as a closed-source baseline.

Evaluation Metrics. For **Task I: Chemical Representation Translation**, we use *Exact Match (EM)*: a prediction receives credit only if it exactly matches the target representation.

For **Task II: Incident Response Recommendation**, we adopt three complementary metrics: (1) *LLM Judge* for correctness (Incorrect/Partial/Correct), (2) *BERTScore-F1* for semantic similarity, and (3) *Cautiousness* via Mean Absolute Relative Error (MARE) for numerical outputs such as distances and durations. We also present a human evaluation on the model outputs as outlined in subsection B.2.

For **Task III: HAZMAT Examination**, we evaluate model performance using standard *accuracy* on multiple-choice questions.

2.3 Overall Results

GPT-4O is the best overall performer, with MED42 and LLAMA-3.1 following closely. We present the ChEmREF evaluation results in Table 2. In addition to performance on individual tasks, we compute an overall average score over all columns (excluding cautiousness MARE) for high-level comparison of different models. We found that GPT-4O achieves the highest overall average score of 71.4%, followed by MED42 (68.0%) and LLAMA-3.1 (67.2%).

Models exhibit distinct strengths and weaknesses. In Figure 2, we visualize the results in Table 2, with each metric column normalized to the range of [0,1] to enable clearer model-wise comparison.

Model	I. Translation		II. Incident Response			III. Exam	Avg. (↑)
	Struc. EM (↑)	Unst. EM (↑)	LLM Judge (↑)	BERTScore (↑)	MARE (↓)	Acc. (↑)	
Phi-3 (3.8B)	93.7	60.0	42.7	52.0	3.8	49.0	59.5
ChemLLM (7B)	79.9	56.8	46.3	79.7	120.0	47.3	62.0
Phi-4 (14B)	**97.2**	48.7	25.2	79.9	2.8	60.0	62.2
Med42 (70B)	88.7	61.9	50.8	**80.5**	3.7	58.0	68.0
Llama-3.1 (70B)	93.1	67.3	50.7	64.9	5.1	60.0	67.2
GPT-4o	92.4	**68.0**	**52.7**	80.2	**2.0**	**63.9**	**71.4**

Table 2: **ChEmREF Evaluation Results.** **Bold** and underlined values indicate the **best** and second-best performance in each column. EM stands for exact match. Average is computed over all columns except for MARE.

We observe that individual models often excel in specific tasks while under-performing in others. For example, PHI-4 achieves the highest score in Structured Translation (97.2% EM) but performs the worst in Incidence Response (25.2% LLM Judge Score). These trade-offs manifest as skewed or imbalanced shapes in Figure 2. While GPT-4O, MED42, and LLAMA-3.1 show promising overall performance, it is important to remain aware of each model's distinct weaknesses. Their outputs should always be used with human oversight to ensure safe and reliable decision-making in emergency response.

Limitations of domain-specific training. CHEMLLM and MED42 are two models we evaluated that have undergone domain-specific supervised fine-tuning (SFT). Our results suggest that such domain-specific training *does not* always guarantee successful application in chemical emergency response. Notably, despite being trained on 7 million chemistry-related QA pairs, ChemLLM ranks the worst in structured chemical translation and second-worst in unstructured chemical translation.

Notably, general-purpose models including GPT-4O, LLAMA-3.1 and PHI-4 outperform domain-specialized models on Task III: HAZMAT Exam. We hypothesize that this is due to the interdisciplinary nature of the task, as it may require knowledge spanning chemistry, medicine, emergency response, numerical reasoning, and more. General-purpose models likely benefit from its broader knowledge coverage, whereas domain-specialized models may lose access to such general knowledge after domain-specific training. These results also highlight the complexity of chemical emergency response and the limitations of current domain-specific training technique. We leave the exploration of more effective domain or task specific training strategies as future work.

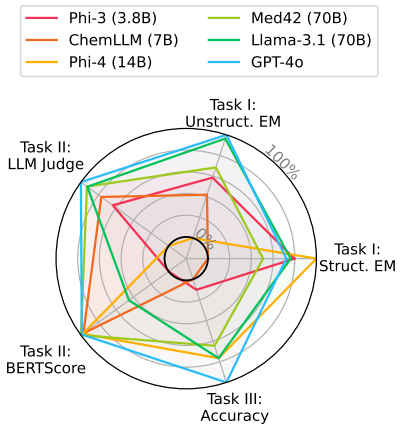


Figure 2: **ChEmREF Evaluation Results With Per-Metric Normalization.** To facilitate clearer model comparison, we normalize each column in Table 2 to the [0,1] range.

3 Conclusion

In this work, we introduce ChEmREF, an evaluation framework for LLMs covering three tasks related to chemical emergencies. Our results show that while models perform well with structured representations, used in the Task I: Chemical Representation Translation and the Task II: Incident Response Recommendation, they struggle with the chemical representations that are less structured but more commonly used in human colloquial communication. Addressing this human-AI communication gap should be a key consideration in the design of future AI-assisted emergency response systems. Our results on Task III: Hazmat Examination, suggest that current domain-specific post-training approaches may lead to a loss of general capabilities, making them less suitable for highly interdisciplinary domains like chemical emergency response. More broadly, we believe LLMs have the potential to assist with chemical identification and incident response, but they should be used with caution and positioned as a complement to, rather than a replacement for, human experts in emergency response.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024a. URL https://arxiv.org/abs/2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024b. URL https://arxiv.org/abs/2412.08905.

Tatiana Anikina. Towards efficient dialogue processing in the emergency response domain. In Vishakh Padmakumar, Gisela Vallejo, and Yao Fu, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 212–225, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-srw.31. URL https://aclanthology.org/2023.acl-srw.31/.

Jianxiong Cai, Jiawei Hou, Yiren Lu, Hongyu Chen, Laurent Kneip, and Soren Schwertfeger. Improving cnn-based planar object detection with geometric prior knowledge. In *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, page 387–393. IEEE, November 2020. doi: 10.1109/ssrr50563.2020.9292601. URL http://dx.doi.org/10.1109/SSRR50563.2020.9292601.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling, 2023. URL https://arxiv.org/abs/2301.12586.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. Med42 – evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches, 2024. URL https://arxiv.org/abs/2404.14779.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL https://aclanthology.org/2022.emnlp-main.26/.

EMT Training Base. Scene size-up: The first step of ems calls, 2019. URL https://emttrainingbase.com/scene-size-up/. Accessed: 2025-02-14.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks, 2023. URL https://arxiv.org/abs/2305.18365.

Yang Han, Ziping Wan, Lu Chen, Kai Yu, and Xin Chen. From generalist to specialist: A survey of large language models for chemistry. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1106–1123, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.74/.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Miltiadis Marios Katsakioris, Ioannis Konstas, Pierre Yves Mignotte, and Helen Hastie. Learning to read maps: Understanding natural language instructions from unseen maps. In Malihe Alikhani, Valts Blukis, Parisa Kordjamshidi, Aishwarya Padmakumar, and Hao Tan, editors, *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 11–21, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.splurobonlp-1.2. URL https://aclanthology.org/2021.splurobonlp-1.2/.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=e2TBb5y0yFf.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelmi, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, Caroline T. Holick, Tanya Gupta, Mehrdad Asgari, Christina Glaubitz, Lea C. Klepsch, Yannik Köster, Jakob Meyer, Santiago Miret, Tim Hoffmann, Fabian Alexander Kreth, Michael Ringleb, Nicole Roesner, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. Are large language models superhuman chemists?, 2024. URL https://arxiv.org/abs/2404.01475.

Aryan Naim, Ryan Alimo, and Jay Braun. Ai agents in emergency response applications, 2021. URL https://arxiv.org/abs/2109.04646.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Amir Sharifi, Ahmadreza Zibaei, and Mahdi Rezaei. Deephazmat: Hazardous materials sign detection and segmentation with restricted computational resources, 2020. URL https://arxiv.org/abs/2007.06392.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Camilo Thorne and Saber Akhondi. NLP for chemistry – introduction and recent advances. In Roman Klinger, Naozaki Okazaki, Nicoletta Calzolari, and Min-Yen Kan, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 45–49, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-tutorials.8/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

U.S. DoT. Emergency response guidebook (erg), 2024. URL https://www.phmsa.dot.gov/training/hazmat/erg/emergency-response-guidebook-erg. Accessed: 2025-02-14.

U.S. EPA. What is the cameo software suite?, 2024. URL https://www.epa.gov/cameo/what-cameo-software-suite. Accessed: 2025-02-14.

U.S. HHS. Chemical incident management: Chemm-ist and chemm-ist for first responders, 2024a. URL https://chemm.hhs.gov/chemmist.htm. Accessed: 2025-02-14.

U.S. HHS. The golden first minutes — initial response to a chemical hazardous materials incident, 2024b. URL https://chemm.hhs.gov/detailedinfo.htm. Accessed: 2025-02-14.

Yotaro Watanabe, Kentaro Inui, Shingo Suzuki, Hiroko Koumoto, Mitsuhiro Higashida, Yuji Maeda, and Katsumi Iwatsuki. Computer-assisted structuring of emergency management information: A project note. In Kentaro Inui, Hideto Kazawa, Graham Neubig, and Masao Utiyama, editors, *Proceedings of the Workshop on Language Processing and Crisis Information 2013*, pages 10–18, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL https://aclanthology.org/W13-4502/.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. Chemllm: A chemical large language model, 2024. URL https://arxiv.org/abs/2402.06852.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

A Extended ChEmREF Description and Metrics

A.1 Task I: Chemical Representation Translation

Unstructured Representations		Structured Representations	
Common Name	tert-Butyl alcohol	IUPAC Name	2-methylpropan-2-ol
UN-Number	1120	Molecular Formula	$C_4H_{10}O$
Synonyms	Trimethylcarbinol	InChI	InChI=1S/$C_4H_{10}O$/c1-4(2,3)5/h5H,1-3H3
	2-Methyl-2-propanol	SMILES	CC(C)(C)O

Table 3: **Different Chemical Representations of tert-Butyl alcohol, a sample hazardous material.** In Task I: Chemical Representation Translation (subsection A.1) we evaluate LMs on translating between these representations, with a focus on translating to/from Common Name, which is most frequently used in chemical emergency correspondence and guidebooks.

On receiving a dispatch call, first responders must rapidly and accurately identify the emergency chemicals involved, often based on descriptions of placards, container labels, and other cues. However this process can be complicated by the existence of multiple representations for the same chemical. In Table 3, we provide 7 different representations for the HAZMAT chemical named *tert-Butyl alcohol*. In practice, first responders may spend a few minutes consulting the ERG or electronic databases to obtain the representation required for subsequent decision-making. Task I investigates whether LLMs can assist in this critical task at a fast speed.

To build this task, we first subsample 100 random chemicals from the 1,035 hazardous materials listed in the 2024 Emergency Response Guidebook (ERG) and collect their corresponding chemical representations by cross-referencing the PubChem database. For each chemical, we consider both structured and unstructured translation settings, resulting in 19 source-target translation pairs in total.

- **Structured translation** (12 pairs) involves bi-directional, pairwise translation among the 4 structured representations (IUPAC Name, Molecular Formula, InChI and SMILES[3]). This setting evaluates whether LLMs can reason over the structure of hazardous chemicals.
- **Unstructured translation** (7 pairs) focuses on translating to/from Common Name, as it is most frequently used in emergency communication. This includes translations between {UN-Number, IUPAC Name, Molecular Formula} $\leftrightarrows$ Common Name (6 pairs) and Synonyms $\rightarrow$ Common Name (1 pair).

We evaluate model performance using exact match (EM). A prediction receives full credit when the output matches the target chemical representation exactly, and zero otherwise.

A.2 Task II: Incident Response Recommendation

Once the chemical involved in the incident has been identified and the suitable chemical representation is found (Task I), emergency responders must analyze the incident and provide appropriate safety

[3]Regarding SMILES evaluation, we rely on the single SMILES string that PubChem now provides for every compound. PubChem has deprecated the separate canonical and isomeric fields in favor of a unified SMILES that is both canonicalized and stereo-/isotope-explicit. For each compound, this hybrid and unique canonicalized form is the isomeric SMILES and was used verbatim during both training and evaluation. As such, the task is designed to reproduce PubChem's published representation, not to normalize across equivalent non-canonical variants which do not have unique representations (i.e., one non-cannocalized form can represent multiple chemicals).

recommendations. For example, in the event of an anhydrous ammonia leak (commonly referred to as "ammonia gas"), the responder may advice on safe evacuation distances (*e.g.*, 330 feet), protective gear (*e.g.*, self-contained breathing apparatus, or SCBA) and containment procedures (*e.g.*, applying water spray to reduce vapors). Task II evaluates whether LMs can assist first responders in drafting emergency response recommendations.

To construct this task, we extract the official recommendations from the 2024 Emergency Response Guidebook (ERG; U.S. DoT 2024). For each chemical, the ERG provides detailed guidance across six key dimensions: fire or explosion, health, public safety, protective clothing, spill or leak, and first aid. Following this, we prompt LLMs to generate recommendations for each of these six aspects through separate queries. For efficient evaluation, we randomly subsample 100 chemicals across the 4 structured representation types from the 1,035 chemicals in Task I, resulting in a total of 2,400 queries for Task II.

For evaluation, we compare the model generated recommendations and the ground-truth recommendations in the ERG using two metrics: **(1) LLM Judge** [Zheng et al., 2023]: We use GPT-4-Turbo as an automatic evaluator for the quality of each generation. It assigns either **Incorrect (0)**, **Partially Correct (0.5)**, or **Perfect (1)** to each generation. Details on the judging prompts and setup can be found in the §F. **(2) BERTScore-F1** [Zhang et al., 2020]. We use BERTScore-F1 as a secondary metric to capture semantic similarity and surface-level variation in phrasing.

We introduce a separate metric named **(3) Cautiousness** for queries involving numerical outputs such as length, time, and volume (*e.g.*, recommending a **330**-foot evacuation distance, recommending not to enter the accident site for **24** hours). This metric quantifies how closely model-generated values align with the ERG. Specifically, we compute an Mean Absolute Relative Error (MARE) metric between the predicted ($\hat{y}_i$) and ground truth (y_i) values.

$$\text{MARE} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \;=\; \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right|.$$

A.3 Task III: HAZMAT Examination

Beyond translation and emergency response recommendation, a fundamental understanding of HAZMAT concepts is also essential for first responders. Such knowledge can be evaluated through question-answering tasks that more closely reflect real-world scenarios. In practice, first responders, along with other professionals like lab technicians, truck drivers, and warehouse managers, are required to pass HAZMAT certification exams. These exams offer a valuable benchmark for evaluating a model's understanding of HAZMAT chemicals.

In Task III, we collect a total of **865 multiple-choice questions** from 46 publicly available quizzes on ProProfs.[4] These quizzes cover a wide range of topics, including HAZMAT Awareness, Workplace Safety, Transportation Safety and more. Each quiz provides answer keys and explanations that have been reviewed by both educators and learners preparing for the exam. Additionally, we manually filtered and verified the questions to ensure quality. We evaluate LLMs on these questions and report standard accuracy.

B Analyzing LLM performance on ChEmREF

In this section, we further break down model performance on the three core tasks and analyze common failure cases and limitations. We highlight several key observations, with additional discussion deferred to Appendix G.

B.1 Analyzing Task I: Chemical Representation Translation

In Figure 3, we break down the performance on Task I by source and target chemical representations. The reported performance is averaged with all six evaluated models. We have the following observations:

[4] `https://www.proprofs.com/quiz-school/`

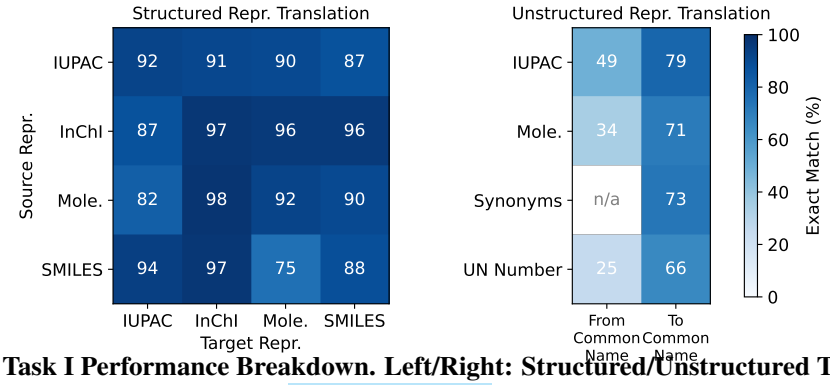


Figure 3: **Task I Performance Breakdown. Left/Right: Structured/Unstructured Translation.** For brevity, we use "Mole." to denote Molecular Formula. We report EM scores for each source-target pair, averaged across all evaluated models.

Structured Translation vs. Unstructured Translation. We observe that models perform significantly better when translating between *structured* chemical formats such as IUPAC Name, Molecular Formula, InChI, and SMILES, compared to translation involving *unstructured* forms like Common Name, UN-Name, and Synonyms. For example, PHI-4 achieves a EM of 97.2% on structured translation, but the performance drops to 48.7% on unstructured translation. Similar trends are observed on other models, highlighting the model's limitation on grounding and normalization when working with less structured or colloquial names.

We have several hypotheses for this performance gap. (1) The structured representations are rule-based and systematic, making them more suitable for step-by-step reasoning through chain-of-thought prompting. In contrast, unstructured translation tends to rely more on the model's memorization than on logical reasoning. (2) Structured representations are defined by international standards and are likely to appear more frequently in chemical databases and online resources, making them more prevalent in the model's pretraining data. In contrast, unstructured forms like Synonyms are often long-tail, ambiguous, and geographically dependent. In general, while unstructured representations are more accessible to humans, structured representations are more machine-actionable. This observation should inform the future design of AI-assisted emergency response systems and effective human-AI collaboration.

To Common Name vs. From Common Name. Translation to Common Name is a critical step in the emergency response workflow, as it often serves as the entry point for subsequent actions such as consulting the Emergency Response Guidebook (see Figure 1). In Figure 3 (Right), we observe that models perform better when translating to Common Name than from it, which we view as a promising sign. However, the average EM score on these translation pairs is 72.3%, which remains insufficient.

Self-translation as a sanity check. Self-translations (*e.g.*, from SMILES to SMILES) is a simple extension from the translation task. We additionally evaluate models on self-translation on the four structured representations, as a sanity check. We report the detailed model-specific results in Fig. 7. LLAMA-3.1 achieved the highest EM score (98.6%), followed closely by PHI-4 (97.8%) and MED42 (95.0%). Surprisingly, CHEMLLM achieves a lower score of 86.3% on this task, suggesting a lack of basic understanding of chemical representation types. We also observe that model rankings on self-translation are consistent with their ranking on structured translation in Table 2, suggesting that this simple check can serve as an early-stage evaluation for filtering out less suitable models.

B.2 Analyzing Task II: Incident Response Recommendation

Evaluating quantitative recommendations and cautiousness. Many emergency response actions rely on quantitative recommendations, *e.g.*, advising an **330**-foot evacuation radius. While examining the model outputs in Task II, we first observe that models often neglect to generate quantitative recommendations when expected. Among queries where the ground-truth recommendation includes a

numerical value, models produce values in the correct category (*e.g.*, length, mass, time, concentration) in only 20% of cases.

For outputs that include valid numerical values, we further compute the mean absolute relative error (MARE) metric (defined in §A.2) as a measurement of model's cautiousness. CHEMLLM demonstrates the most cautious behavior, with an MARE of **120.03**. Our manual inspection reveals its tendency to overestimate values, meaning that the quantities recommended by ChemLLM are on average 120 times higher than the groundtruth, which could place unnecessary strain on response efforts. GPT-4O has the lowest MARE of 2.0, which is significantly lower than 120.0. While overestimation still occurs, it is far less severe and likely more manageable in practical settings.

Our analysis on cautiousness also emphasize the limitations of automatic metrics like LLM Judge or BERTScore. In high-stakes domains like chemical emergency response, it is essential to assess not only the overall plausibility of a recommendation, but also whether it produces a quantitative value when needed, and whether that value is accurate.

Verbose outputs may hinder time-sensitive decision-making. In our manual review of 20 recommendations generated by GPT-4O, we find that the model frequently produces outputs that extend beyond the scope of the prompt. For example (§G.2.2), when asked specifically about the fire or explosion hazards of a given chemical, GPT-4O additionally generates guidance on personal protective equipment (PPE) and continuous assessment of hazards, resulting in outputs that are longer and less focused. While such comprehensive responses demonstrate the model's broad knowledge, this verbosity may pose challenges in time-sensitive scenarios, where concise and actionable information is preferred.

Human evaluation by a first responder shows alignment with the LLM-Judge. We emphasize that Tasks 1 and 3 have verifiable answers that can be evaluated deterministically: Task 1 uses Exact/Partial Match on chemical representations, while Task 3 relies on Multiple Choice Answer Keys. Only Task 2 requires long-form responses, where the gold-standard outputs were not LLM-generated but instead obtained directly from the federally standardized *Emergency Response Guidebook* (ERG), the authoritative reference for emergency responders.

For Task 2, we adopt a combination of Exact Match, BERTScore, and LLM-based judgment as a starting point. We acknowledge, however, that these automated metrics cannot fully substitute for human validation. In particular, using GPT-4 as a judge when GPT-4o is one of the evaluated models raises potential bias concerns. To mitigate this, one of the authors—a certified EMT with HAZMAT training and five years of field experience—conducted multiple rounds of testing to refine and validate outputs. While large-scale human validation was beyond the scope of this initial study, we consider it a critical next step as we expand to more complex emergency scenarios.

The LLM Judge score measures alignment between a model's response and the ERG reference. For example, a score of 52.7% indicates that while models can capture the broad contours of appropriate guidance (e.g., recognizing that PPE is required), they often lack specificity or completeness (e.g., omitting gloves or eye protection). Such partial correctness can present real-world risks, underscoring the importance of human oversight for emergency response applications.

To further probe reliability, we conducted a small-scale human validation study of Task 2 responses. We sampled 96 responses across six prompt types, focusing on the top two performing models and the lowest performing model. Both the LLM-Judge and a HAZMAT-certified EMT independently reviewed each prompt, model output, and corresponding ERG section. Additional relevant but accurate information was accepted, while omissions of significant ERG content or introduction of inaccuracies were penalized.

Model	Agreement	LLM-Annotator Avg (%)	Human Avg (%)
Llama3	88.5	69.8	62.0
Phi4	94.8	86.5	84.4
GPT4o	76.0	69.3	68.8

Table 4: Comparison of Task 2 evaluation between LLM-Judge and human validator (HAZMAT-certified EMT with 5 years of experience).

B.3 Analyzing Task III: HAZMAT Examination

Category	Average Accuracy	Number of Exams
CBRN & Chemical Safety	46.5	3
HAZMAT Awareness & Operations	51.9	15
HAZMAT Lesson Plans	70.8	8
Hazard Communication & Workplace Safety	**75.6**	9
IMDG & Transportation Safety	66.7	1
Incident Management & Reporting	62.6	10
Total Exam Cohort	62.0	46

Table 5: **Task III Accuracy by Category.** We report mean accuracy across all evaluated models. Models perform better on categories with instructional content, but struggle with technical and risk-related topics, highlighting their limitations in these specialized domains.

Performance varies across HAZMAT exam categories. When gathering HAZMAT-related exams from ProProfs, we manually group them into six major exam categories. In Table 5, we report the accuracy on different exam categories. Models perform well on instructional content, such as Hazard Communication and Workplace Safety (Accuracy: 75.6%) and HAZMAT Lesson Plans (Accuracy: 70.8%), but struggle with more technical and risk-related topics, including CBRN & Chemical Safety (Accuracy: 46.5%) and HAZMAT Awareness & Operations (Accuracy: 51.9%). This discrepancy highlights limitations in current models' ability to handle complex, high-stakes topics that require deeper domain understanding.

Contamination check. To ensure that model performance reflects understanding and generalization instead of memorization, we conducted exact-match searches using the Infini-gram engine [Liu et al., 2024] over several open pretraining corpora—`Dolma-v1.7` [Soldaini et al., 2024], `RedPajama` [Weber et al., 2024], `Pile-train` [Gao et al., 2020], and `C4-train` [Raffel et al., 2020]. No overlaps were found with our multiple-choice QA dataset.

C Glossary

Term	Definition
Hazardous Materials (HAZMAT)	Substances that pose a risk to health, property, or the environment, often requiring special handling and regulations.
Emergency Response Guidebook (ERG)	A resource used by first responders to identify hazardous materials and guide their response during transport incidents.
International Union of Pure and Applied Chemistry (IUPAC)	A global organization that sets standards for chemical nomenclature, terminology, and measurement.
Simplified Molecular Input Line Entry System (SMILES)	A notation system that allows the representation of a molecular structure using a linear string of text.
International Chemical Identifier (InChI)	A textual identifier that provides a unique representation of chemical substances to facilitate data sharing and searchability.
UN-Number	A UN number is a four-digit code used to identify hazardous materials for safe transportation and emergency response, regulated by frameworks such as the International Maritime Dangerous Goods (IMDG) Code, International Air Transport Association (IATA) Dangerous Goods Regulations, U.S. Department of Transportation (DOT) Regulations, and the European Agreement concerning the International Carriage of Dangerous Goods by Road (ADR).
Commercial Driver's License (CDL)	A specialized license required for individuals operating large, heavy, or hazardous material-carrying commercial vehicles.
Compound Identifier (CID)	A unique numerical identifier assigned to chemical substances in the PubChem database for reference and research purposes.
Personal Protective Equipment (PPE)	Gear and clothing designed to protect individuals from hazardous materials, contamination, or injury in various environments.
National Registry of Emergency Medical Technicians (NREMT)	A certification organization that ensures emergency medical personnel meet national training and competency standards.

Table 6: **Key Definitions.** A glossary of important terms related to HAZMAT and emergency response.

D Related Work

Evaluating LMs on Chemistry Knowledge. Language technologies have long been applied to the chemistry domain, with Thorne and Akhondi [2024] providing a comprehensive review. Recent benchmarks like ChemLLMBench [Guo et al., 2023] and ChemBench [Mirza et al., 2024] evaluate large language models (LLMs) on a range of chemistry knowledge and skills, demonstrating their potential while also highlighting limitations such as difficulties with SMILES representations, hallucinations, and overconfident predictions. Complementary to these works, our work offers a distinct perspective by focusing on the critical and specific domain of HAZMAT chemicals. Furthermore, we move beyond memorizing chemicals properties to actively reasoning about and applying them in real-world HAZMAT emergency response.

Adapting and Augmenting LMs for Chemistry. While general-purpose LMs show promise in chemistry, further specialization is possible through techniques like continued pre-training or instruction tuning [Zhang et al., 2024, Christofidellis et al., 2023]. Separately, researchers have explored incorporating molecular structure as a distinct modality [Edwards et al., 2022] with contrastive learning [Liu et al., 2023], or augmenting general-purpose LMs with chemistry-specific tools [M. Bran et al., 2024]. Han et al. [2025] provide a recent overview of this area. In this work, we focus on single-modality LMs due to their prevalence. We benchmark both general-purpose LMs and domain-specialized LMs to investigate the impact of domain-specific training on HAZMAT-related tasks.

NLP for Emergency Response. The NLP community has a history of exploratory contributions to various aspects of emergency response. For example, Watanabe et al. [2013] developed a system to automatically extract key information from emergency management logs for database construction. Katsakioris et al. [2021] investigated the problem of converting natural language descriptions of an incident location into GPS coordinates. Other work, such as Anikina [2023], has focused on dialogue act classification and slot tagging for emergency response dialogues. These works demonstrate the potential of language technology in the domain of emergency response. With the advent of large language models, we revisit this potential and explore their applicability to HAZMAT emergency response.

Existing Software. While automated tools have been developed to offer support in emergency response, they are often limited to specific cases or lack integration with the latest language technologies. For example, CHEMM [U.S. HHS, 2024a] provides an automated chemical identification tool based on patient vitals, such as pupil size and skin condition. However, this functionality is limited to later steps in the response sequence and excludes earlier-stage scenarios like dispatch communications [EMT Training Base, 2019]. Another tool, CAMEO [U.S. EPA, 2024], offers web interfaces with database search, threat zones modeling and incident site mapping, but lacks modern language technologies like flexible keyword matching or a question-answering interface. To this day, first responders still largely resort to the physical Emergency Response Guidebook (ERG). LLMs, with their extensive knowledge capacity and user-friendly interfaces, have the potential to bridge this gap and enhance decision-making during these critical moments.

Efforts in CV and Robotics Research Communities. Beyond the scope of natural language processing and language models, fields such as computer vision and robotics also hold promise for improving emergency response. Naim et al. [2021] discuss the integration of AI agents in emergency response applications, emphasizing their potential to enhance situational awareness and decision-making during critical incidents. Sharifi et al. [2020] introduced DeepHAZMAT, a deep learning-based robot designed for hazardous materials sign detection and segmentation under constrained computational resources. This approach uses convolutional neural networks (CNNs) to identify hazardous materials in complex environments, facilitating faster and more accurate scene assessments. Furthermore, Cai et al. [2020] demonstrated the effectiveness of incorporating geometric prior knowledge into CNN-based planar object detection, which can significantly improve the accuracy of identifying hazardous materials and related objects in emergency scenarios.

E Example Prompts and Outputs

Previously in Table 1, we used redacted examples due to space limit. In this section, we provide examples of full inputs and expected outputs (marked in blue) for each of the three tasks in ChEmREF.

E.1 Translation Task

E.1.1 Unstructured Translation

You are an expert chemist. Question: What is the molecular formula of Sodium superoxide?
Answer: NaO_2.

E.1.2 Structured Translation

You are an expert chemist. Question: What is the molecular formula of Sodium superoxide?
Answer: Let's think step by step.
Sodium (Na) is an alkali metal with a +1 oxidation state.
Superoxide is an ion with the formula O_2^- and a -1 charge.
To balance the charges, one sodium ion (Na^+) will pair with one superoxide ion (O_2^-).
Therefore, the molecular formula is NaO_2.

E.2 Incident Response Task

Scenario: A fire has started involving Mercuric nitrate (IUPAC-name).
Prompt: Provide recommendations on fire control, explosion risks, and necessary evacuation procedures for the substance involved, referencing the hazmat guide where appropriate.
Answer:
Fire Control
Use water spray or fog to cool the area and prevent fire spread.
Avoid direct water streams that may spread contamination.
Use dry chemical powder or CO_2 for fire suppression.

Explosion Risks
Mercuric nitrate is a strong oxidizer that can intensify fires.
Keep it away from combustible materials.

Evacuation Procedures
Evacuate all non-essential personnel at least 100 meters away.
Ensure emergency responders wear PPE, including self-contained breathing apparatus.

E.3 Hazmat Examination Task

Reason through this question carefully and provide the answer as the correct multiple-choice letter.
Question: What amount of chlorine being transported requires the driver to stop before railroad crossings?
A. An amount requiring a Table 2 placard.
B. Any amount.
C. Any amount when combined with explosives A.
D. 1000 lbs or more.
Answer: The correct answer is B. Any amount.

F LLM Judge Setup and Evaluation Protocol

To evaluate model-generated responses against official guidelines, we developed an automated evaluation pipeline that uses GPT-4o as an LLM judge. The task involved comparing model responses to official guidance from the Emergency Response Guidebook (ERG) in six categories:

- **Fire or Explosion Hazard:** Focuses on whether the model correctly identifies flammability, reactivity, or explosiveness.

- **Health Hazard:** Assesses accuracy in toxicity, corrosiveness, or inhalation danger details.
- **Public Safety and Evacuation:** Evaluates evacuation distances, containment zones, and timing, considering whether the model is overly cautious or not cautious enough.
- **Protective Clothing:** Ensures that the model recommends appropriate personal protective equipment (PPE) as indicated in the guide.
- **Spill or Leak:** Focuses on containment strategies such as diking, neutralization, or absorption.
- **First Aid:** Evaluates whether the advice aligns with ERG guidance for eye, skin, or inhalation exposure.

F.1 Evaluation Prompt Design

Each evaluation prompt was designed to guide GPT-4o in judging the model's output relative to the ERG recommendations. The system prompt positioned the LLM as a hazardous materials safety expert tasked with verifying whether the model output correctly covered the major information required under the specified category.

```
You are a hazardous materials safety expert.  Your task is
to evaluate a model-generated hazmat safety recommendation
against the official Emergency Response Guide (ERG) guid-
ance.  Your goal is to check whether the model covers the
major information from the ERG under the given category.  It
is acceptable if the model includes extra relevant information
not in the ERG, as long as it is accurate.  Only penalize if
something significant from the ERG is missing or if the model
includes incorrect information.

Use one of these labels:

  • Incorrect:  Key points are missing or wrong.
  • Partial:  Some important points are correct, but others
    are missing or incorrect.
  • Correct:  Most or all important ERG details are present
    and accurate.
```

For each one of the six categories mentioned above, we supplement the LLM Judge with specific judging instructions. For example, the judging instructions for "Fire or Explosion Hazard" category encourages the LLM Judge to prioritize flammability, reactivity, and explosiveness in its assessment.

G Extended Results

In this section, we provide supplementary figures and analysis for the Task I: Translation and Task II: Incident Response.

G.1 Analyzing Task I: Chemical Representation Translation

G.1.1 Prompting Methods

For Task I, we ran a pilot experiment using four different prompting strategies, including (1) zero-shot chain-of-thought ("Let's think step by step"; Kojima et al. 2022), (2) zero-shot chain-of-thought with directions, (3) one-shot chain-of-thought prompting [Wei et al., 2022], and (4) one-shot chain-of-thought prompting with directions. Combining the four prompting strategies with the 16 source-target translation pairs by chemical representation type, we experiment with 64 settings for each model for each chemical. In Figure 4 and Figure 5, we report the average of the exact match over the 16 source-target pairs.

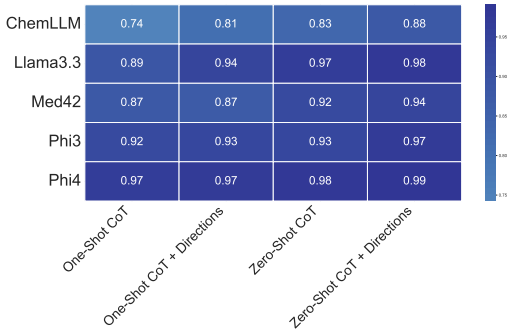


Figure 4: **Translation Task Model Performance Across Prompt Types:** Performance heatmap using HAZMAT data.

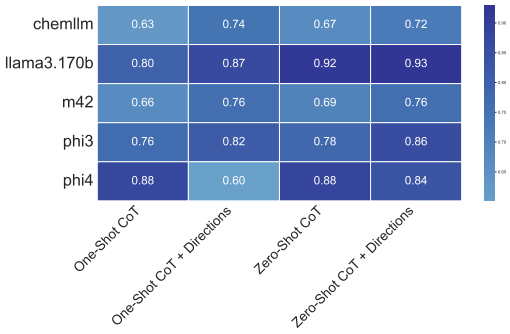


Figure 5: **Translation Task Model Performance Across Prompt Types:** Performance heatmap using Non-Hazmat chemical data.

Performance is slightly sensitive to the prompt format across models; the best prompt format varies with the model. As shown in Figure 4, the impact of prompt format on performance varies significantly by model. For example, ChemLLM achieves 62% accuracy with directions included in the prompt, compared to 76% without directions—an improvement of 14% when instructions are omitted on HAZMAT chemicals. In contrast, Phi-4 shows little difference on HAZMAT data (74% with vs. 75% without), but exhibits large variation on non-HAZMAT data, sometimes surpassing ChemLLM on HAZMAT. Empirically, this suggests that providing explicit chemical directions can worsen performance in some cases. We hypothesize that when prompts include detailed instructions, models may overfit to the structure of the examples, producing responses that mirror the prompt rather than generalizing more effectively. Interestingly, models such as `Llama-3.1`, `Med42`, and `Phi-3` exhibit moderate variation in performance on HAZMAT data, suggesting that structured prompting can enhance accuracy , but the extent of this benefit varies by model. These findings underscore the importance of tailoring prompt strategies to the specific architecture and reasoning capabilities of each model.

This choice was motivated by the substantial variation between Common Name and Synonyms, where prior examples offer little insight into representation patterns, unlike in the structured translation setting.

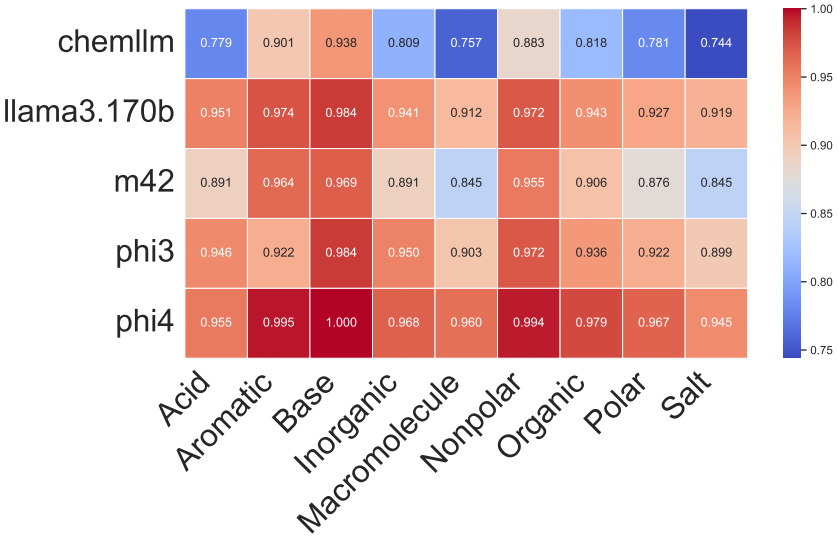


Figure 6: **Exact Match Scores for Model Performance on `hazmat` Data Across Chemical Categories.** This heatmap illustrates how different models perform when translating hazardous material data, highlighting variations in accuracy across chemical types.

G.1.2 Chemical Properties

The provided heatmaps in Figure 6 compare the performance of various models across nine distinct chemical categories: Acid, Aromatic, Base, Inorganic, Macromolecule, Nonpolar, Organic, Polar, and Salt.

Phi-4 consistently outperforms the other models, achieving nearly perfect performance in the Base category and maintaining high accuracy across the remaining categories. ChemLLM lags behind the other models, particularly in the Inorganic, Nonpolar, and Salt categories, with scores generally below 0.81. While it performs well for simpler categories like Aromatic and Base, it struggles with more complex chemical types.

Certain categories display clear performance patterns across models. The **Base** and **Aromatic** categories achieve consistently high exact match (EM) scores, with five-model averages of 0.975 and 0.951, respectively, indicating that these chemical types are reliably easier for models to translate. In contrast, the **Inorganic** and **Nonpolar** categories present greater challenges, with lower average EM scores of 0.912 and 0.876, and notable drops in models like ChemLLM (0.809 and 0.757, respectively). Categories such as **Macromolecule** and **Polar** demonstrate moderate variability, with EM scores ranging from 0.757 to 0.960 for Macromolecule and 0.818 to 0.979 for Polar. These variations highlight that while some chemical representations are universally well-handled, others reveal weaknesses in certain models, particularly lower-performing ones like ChemLLM and Med42. Overall, this suggests that model architecture and training significantly influence performance on more complex or diverse chemical categories.

G.1.3 Input-Output Types

Translation performance rarely varies with source chemical representation type for structured representation types. Figure 7 provides a overall summary of all structured input-output combinations. For CHEMLLM, we see that that an input-output combination of SMILES to InChI is the most difficult translation.

Translation performance significantly varies with source chemical representation type for unstructured representation types. Although Common Name and UN-Number are the only representations provided in the ERG, we observe that most models struggle to accurately identify a UN-Number and translate from a Common Name. Comparing the worst and best performing models, PHI-3 and GPT-4O, respectively, we see that both models perform better when tasked with translating to a Common Name, which is encouraging for potential use by first responders. However, when translating away from a Common Name, models tend to perform better when translating to structured forms like IUPAC Name or Molecular Formula rather than to a UN-Number. Despite this, performance when translating away from Common Name remains significantly lower, achieving only about half the accuracy observed when translating between structured forms, where models exhibit near-perfect performance.

Interestingly, the high translation scores on Synonyms suggest that models are more adept at recognizing alternative names for the same chemical, providing additional motivation for their use by first responders. Synonym recognition can be critical in high-stakes situations where rapid identification of hazardous materials is essential. By leveraging this strength, models can assist emergency personnel in identifying chemicals accurately, even when encountering less familiar terms, ultimately enhancing response efficiency and safety.

G.1.4 HAZMAT Chemicals v.s. Non-HAZMAT chemicals.

Translation of more common HAZMAT chemicals is easier than non-HAZMAT chemicals. In this section, we compare Chemical Familiarity by HAZMAT vs Non-HAZMAT data. Non-HAZMAT Data was acquired from the PubChem database, and a subset of 100 chemicals were selected.

Overall, in Figure 10, the Translation Task on HAZMAT data received a 90.9%, whereas the Translation task on Non-HAZMAT data scored a 77.7%. The significant improvement in chemical comprehension and translation accuracy are encouraging towards the use of LLMs for HAZMAT scenarios, because they suggest that general-models have more knowledge on HAZMAT data than that of the non-HAZMAT corpus. The improved performance on HAZMAT data suggests that even general-purpose

models can effectively capture the complexities of hazardous materials—an essential capability for safety-critical applications like emergency response, chemical spill mitigation, and regulatory oversight.

G.2 Analyzing Task II: Incident Response

G.2.1 Chemical Guides

Observing ERG Guide numbers allow us to pinpoint performance by chemical properties. Each HAZMAT chemical is linked to one of hundreds of guides in the ERG, explaining what to do in case of a chemical emergency. There can be more than one chemical linked to the same ERG guide, as many substances share similar hazards and response protocols. These guides provide critical information on fire and explosion risks, health hazards, protective measures, spill containment, and first aid procedures. Emergency responders rely on them to determine safe evacuation distances, recommended PPE, and appropriate mitigation strategies. Given the vast number of chemicals and their potential dangers, efficiently associating a chemical with the correct ERG guide is essential for rapid and effective decision-making in emergency situations. In Figure 13, we can see that there is a fairly even distribution across guides (between 77.5% - 80.5%) in terms of most answered correctly and incorrectly, suggesting that models may not be partial to solely flammable or toxic chemicals.

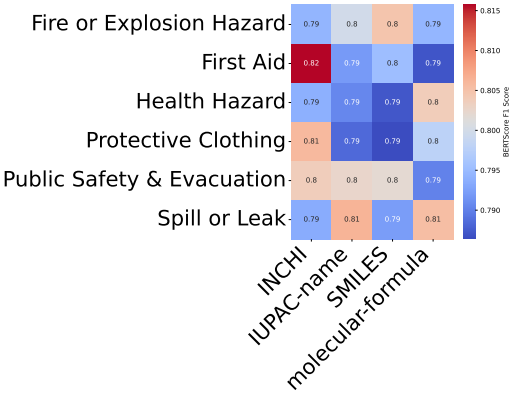


Figure 12: **Incident Response Task Model Performance:** Evaluation Using BERTScore-F1, average scores across all models.

G.2.2 Models Often Provide Additional Information Beyond ERG Guidance

As noted in the main text, models frequently generate responses that go beyond the explicit guidance in the Emergency Response Guidebook (ERG), often including supplementary safety recommendations. Upon manually analyzing a subset of 20 model outputs from GPT-4O, we observed that the models tend to generate extended responses that incorporate multiple safety protocols, regardless of the specific prompt category.

For instance, The following example illustrates how when prompted with a query about *Fire or Explosion Hazard* for Sodium superoxide, GPT-4o not only provided fire suppression guidance but also included additional considerations such as evacuation distances, personal protective equipment (PPE), and hazard assessment protocols. Despite the prompt explicitly asking for information relevant to fire or explosion hazards, the model integrated relevant details from other ERG guide subcategories, demonstrating an implicit understanding of the intersectional nature of hazardous material incidents.

```
In dealing with a fire involving Sodium superoxide (NaO2), it is essential
    to consider its chemical properties and associated hazards. Sodium
    superoxide is a powerful oxidizer and can react vigorously with
    combustible materials, leading to potential fire and explosion risks.
    Here are my recommendations:

### Fire Control
1. **Use of Appropriate Extinguishing Agents:**
   - **Do Not Use Water:** Sodium superoxide reacts with water to produce
     sodium hydroxide and oxygen gas, which can exacerbate the fire.
     Moreover, the heat generated during this reaction can cause further
     thermal decomposition, releasing more oxygen.
   - **Avoid Carbon Dioxide and Foam:** These are ineffective because
     sodium superoxide can decompose them, potentially leading to further
     reactions.
```

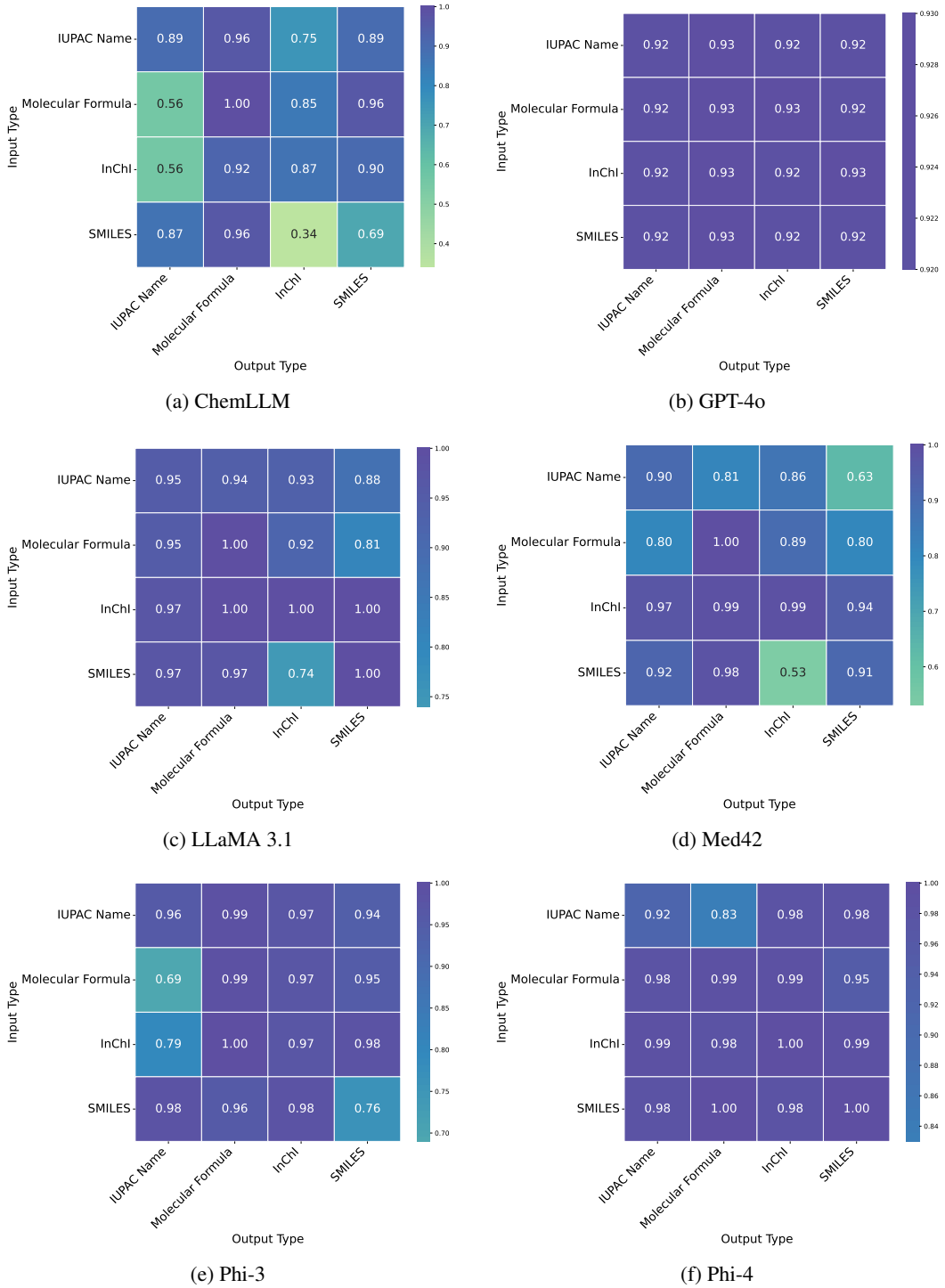
```
   - **Use Dry Powder Extinguishers:** Class D fire extinguishers (
     specifically designed for metal fires) containing dry powder agents
     such as sodium chloride or graphite are recommended. These agents can
     smother the fire without providing additional fuel (oxygen or moisture).


2. **Isolation of Fuel Sources:**
   - Remove or isolate any nearby combustible materials to prevent further
     reaction and fire spread.

### Explosion Risks
1. **Reactivity with Combustibles:** Sodium superoxide can react
     explosively with organic materials (e.g., paper, wood) and certain
     metals. Ensure that such materials are not present in the vicinity or
     are removed promptly.
2. **Controlled Ventilation:**
   - Ensure adequate ventilation to prevent the accumulation of oxygen-rich
      atmospheres that can increase the risk of explosion.
3. **Temperature Control:**
   - Cool surrounding areas with inert gases (if available) or non-reactive
      cooling methods to prevent the decomposition of sodium superoxide,
     which releases oxygen.

### Evacuation Procedures
1. **Initial Evacuation Radius:**
   - Establish an initial evacuation perimeter of at least 150 meters (500
     feet) to account for potential explosions and toxic smoke.
2. **Hazard Assessment:**
   - Conduct a continuous hazard assessment to determine the need for
     expanding the evacuation zone. Factors include the fire's intensity,
     wind direction, and presence of hazardous materials.
3. **Personnel Protection:**
   - Ensure that all responders and personnel in the vicinity are equipped
     with appropriate personal protective equipment (PPE), including self-
     contained breathing apparatus (SCBA) to protect against inhalation of
     toxic fumes.
4. **Communication:**
   - Maintain clear and constant communication with emergency personnel and
      responders about the status of the fire.
```

Evaluation Summary The model output provides a detailed and accurate description of the fire and explosion hazards associated with Sodium superoxide, including its reactivity with combustibles and the need for appropriate extinguishing agents. It also covers the potential for explosive reactions and the importance of controlling ventilation and temperature. However, it does not explicitly mention the possibility of containers exploding when heated or the risk of runoff creating a fire or explosion hazard, which are key points in the ERG guidance. This highlights how models tend to extrapolate relevant safety protocols even when not explicitly requested. The LLM Judge, scored this a Partial (0.5).



(a) ChemLLM (b) GPT-4o

(c) LLaMA 3.1 (d) Med42

(e) Phi-3 (f) Phi-4

Figure 7: **Average Translation Task EM** `hazmat` **Scores Across Input/Output Representation Type.** This figure presents Exact Match (EM) scores for the Translation Task, grouped by model.

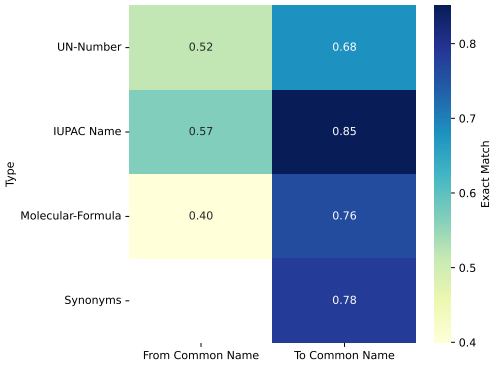


Figure 8: Exact Match Heatmap: Best Performing Model (GPT-4o).

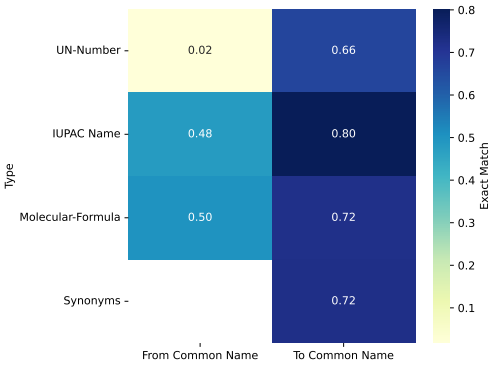


Figure 9: Exact Match Heatmap: Worst Performing Model (Phi-3).

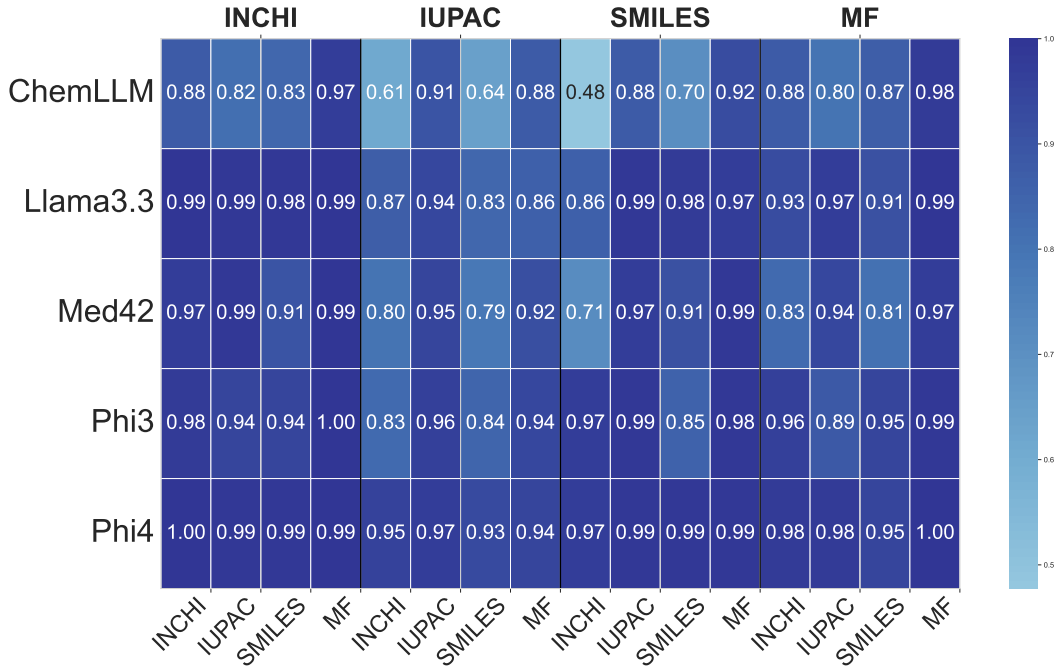


Figure 10: **EM Averages for HAZMAT Data across Input and Output Representation Types.** Each heatmap cell corresponds to a translation task from an **input type** (top row) to an **output type** (bottom column). Note that these results are based on a different evaluation cohort (100 distinct chemicals) compared to Fig. 7, leading to variation in EM scores.

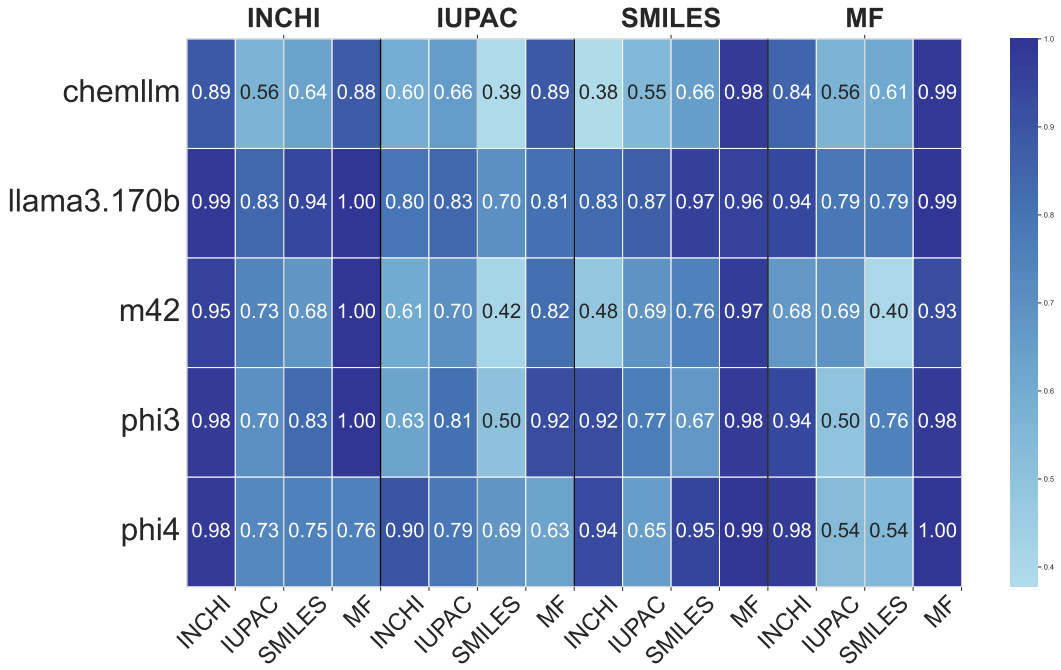


Figure 11: **EM Averages for Non-HAZMAT Data across Input and Output Representation Types.** Similar to Fig. 10, this heatmap reflects translations between input (top row) and output (bottom column) types using a separate cohort of PubChem chemicals.

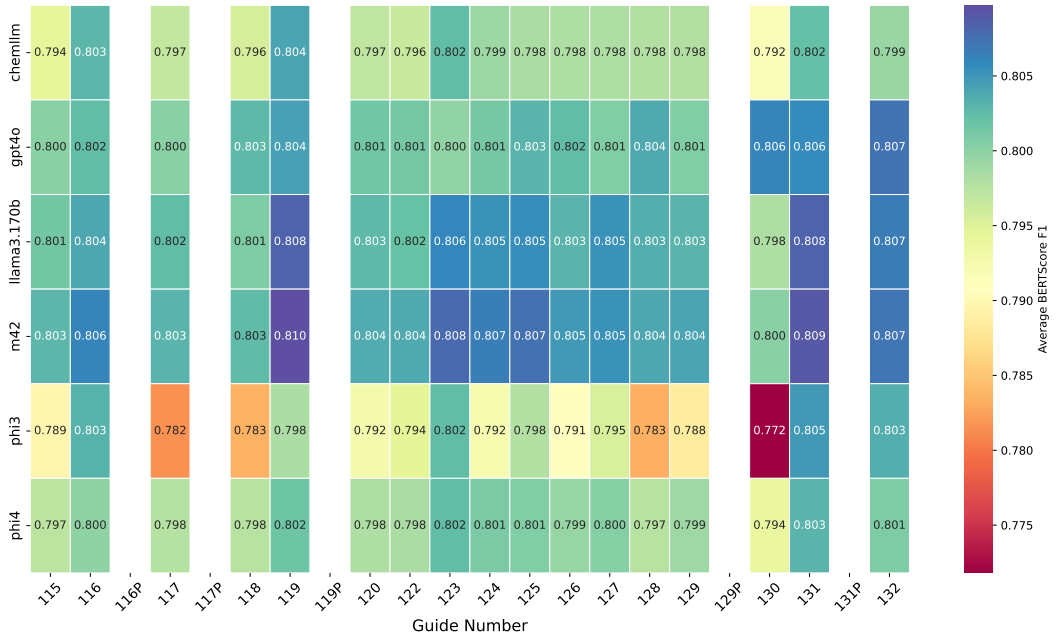


Figure 13: **Translation Task Model Performance:** Evaluation Across Molecule Types. This figure illustrates how model performance varies when translating chemical representations for different molecule types. Each HAZMAT molecule is associated with a specific guide number, which categorizes chemicals based on shared safety characteristics such as flammability, toxicity, and reactivity. Since multiple chemicals can correspond to the same guide number, these guides serve as a standardized way to group similar substances, providing a structured framework for understanding performance variations across molecular categories.