ChEmREF: Evaluating Language Model Readiness for Chemical Emergency Response

Anonymous ACL submission

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 four tasks: (i) domain knowledge question answering from chemical safety and certification exams, (ii) identification of chemical names between Common Name and Synonyms (*e.g.*, "ethanol" to "C_2H_6O"), (iii) fill-in-the-blank emergency response guide generation (*e.g.*, recommending appropriate evacuation distances), and (iv) classifying risk features from real-world HAZMAT incident response reports. Across all four tasks, the strongest models achieve 97.0% exact match on incident response recommendations with retrieval support, but only 63.9% accuracy on HAZMAT examination questions, falling short of the reliability required for safety-critical use. These results indicate that while language models can assist with retrieving information and high-level reasoning, they require human oversight before deployment in real-world HAZMAT emergency response.

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: are language models ready to 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 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 the **Chemical Emergency Response Evaluation Framework** (ChEmREF), designed to systematically assess the capabilities of language models in assisting with a HAZMAT incident. ChEmREF is organized into four tasks which reflect the key steps needed in HAZMAT emergency response: (i) HAZMAT domain knowledge question answering, (ii) chemical

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

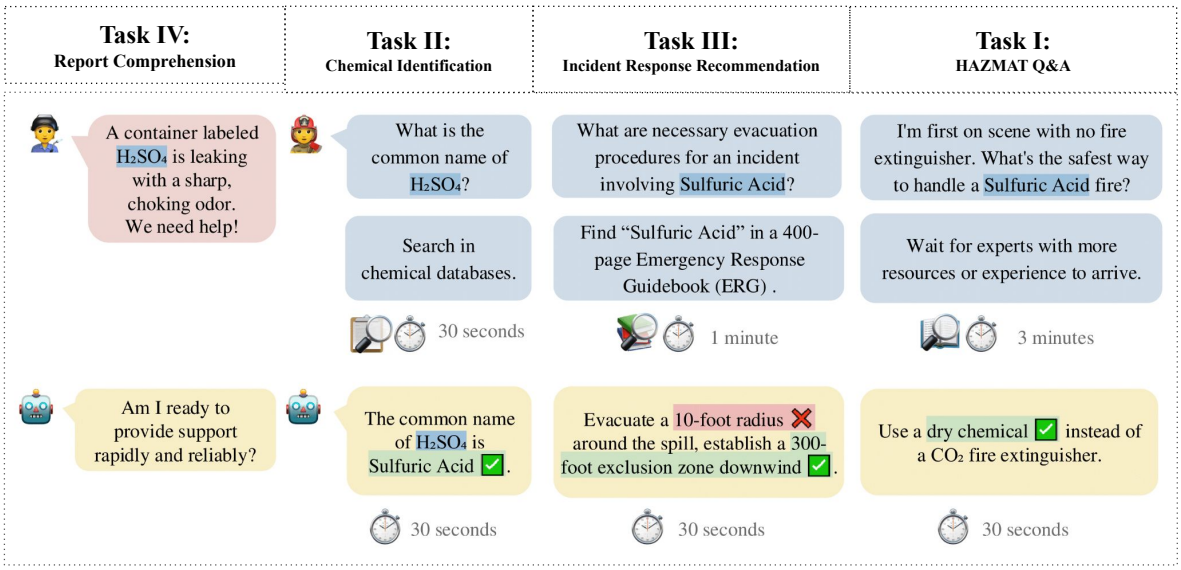


Figure 1: **LLM Assistants in Emergency Response.** In HAZMAT emergencies, timely decisions during scene size-up are critical (EMT Training Base, 2019). This figure compares the traditional chemical response timeline with one aided by an LM. While LMs may accelerate decision-making, they can also produce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

identification of Common Name representations, (iii) decision-making via emergency response guide generation, and (iv) comprehension of risk features in real-world emergency response reports. Figure 1 provides an overview of the tasks and their roles in the emergency response workflow. By covering 1,035 HAZMAT chemicals, 6 emergency response dimensions and 6 HAZMAT exam categories, ChEmREF provides a comprehensive evaluation of LMs in this high-stakes domain.

We evaluate four general-purpose and two domain-specific language models on ChEmREF. Overall, GPT-4O (Hurst et al., 2024) is the top-performing model among evaluated models, followed closely by MED42 (70B) (Christophe et al., 2024) and LLAMA-3.1 (70B) (Grattafiori et al., 2024). In addition to these model-only evaluations, we also experiment with a retrieval-augmented generation (RAG) setup that combines GPT-4O with a sparse lexical retriever based on BM25 (Robertson and Zaragoza, 2009).

Our results demonstrate that models exhibit distinct strengths and weaknesses on ChEmREF. Incorporating retrieval assistance can up to double performance for the strongest out-of-the-box models on Tasks II and III; however, performance on examination-style and comprehension-based questions in Tasks I and IV that require specialized domain knowledge remains limited. This highlights the interdisciplinary nature of chemical emergency response and suggests that task-specific adaptation may be necessary to further improve performance. Across all tasks, performance varies substantially by model scale and access to external knowledge, revealing persistent gaps that limit reliable deployment in real-world HAZMAT settings. 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, 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 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. Complementing these, our work offers a distinct perspective by focusing on the critical and specific domain of HAZMAT chemicals. Furthermore, we move be-

yond 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.

Appendix B describes other related work.

3 ChEmREF: Chemical Emergency Response Evaluation Framework

We introduce ChEmREF: a framework for evaluating whether LMs have the capabilities to assist first responders in chemical emergencies. Our framework consists of four tasks: (i) HAZMAT Examination, which assesses performance on general-knowledge multiple-choice questions from real-world certification exams for first responders and chemical safety tests, (ii) Chemical Representation Identification, which tests whether models can accurately convert between Common Name and Synonyms; (iii) Incident Response Recommendation, which measures the ability of LMs to generate relevant safety measures, evacuation distances, and containment strategies based on incident context; and (iv) HAZMAT Report Comprehension, which assesses how well a model can identify risk features and hazard classes from a real-world incident report. We present abbreviated examples of each task in Table 1. Full prompt examples can be found in Appendix C.

3.1 Task I: HAZMAT Examination

A fundamental understanding of HAZMAT concepts is essential for first responders, who are required to pass HAZMAT certification exams. These exams could also offer a valuable benchmark for evaluating a language model's understanding of HAZMAT chemicals. In Task I, we collect a total of 865 multiple-choice questions from 46 publicly available quizzes on ProProfs,[2] covering a wide range of topics, such as HAZMAT Awareness, Workplace Safety, and Transportation Safety. 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.

Task I involves answering multiple-choice questions with inputs presented in a standardized format: *Question: ... Answer Choices: ... Answer:*. We evaluate LMs via zero-shot prompting without examples or intermediate reasoning.

3.2 Task II: Chemical Identification

On receiving a dispatch call, first responders must rapidly and accurately identify the emergency chemicals involved, however, this process can be complicated by the existence of multiple representations. For example, Table 7 provides several different common 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 II investigates whether LLMs can assist in this critical task efficiently. To build this task, we first obtain all chemicals from the 2024 Emergency Response Guidebook (ERG)(U.S. DoT, 2024) and collect their corresponding chemical representations by cross-referencing the publicly available PubChem(Kim et al., 2025) database. Across all 1,035 chemicals, we select the Common Name, as well as the top 3 most commonly used Synonyms.

Task II evaluates whether models can identify the standard chemical name given alternative Synonyms or chemical identifiers. Inputs include various synonym formats such as chemical IDs, systematic names (*e.g.*, trioxochromium"), and trade names. We evaluate LMs via zero-shot prompting, instructing models to respond with only the standard chemical name without additional explanation.

3.3 Task III: Incident Response Recommendation

Once the chemical involved in an incident has been identified, emergency responders must quickly ac-

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

Task I: HAZMAT Examination (§3.1)
Q: When transporting hazardous materials, when must the driver check the vehicle's tires? A) Only at the start of the trip B) Every 150 miles or every three hours C) Only if the warning light turns on D) After reaching the destination *Answer:* **B**
Task II: Chemical Identification (§3.2)
Q: Given the chemical synonym `trioxochromium`, identify the common/standard name. *Answer:* **Chromium Trioxide**
Task III: Incident Response (§3.3)
Q: Complete the ERG evacuation guidance: "Immediately evacuate area for at least _____ meters." *Answer:* **100**
Task IV: Report Comprehension (§3.4)
Q: <INCIDENT REPORT> Based on the incident report for FLAMMABLE LIQUIDS, N.O.S. (UN1993), which response action is explicitly described? *Answer:* **Apply water spray/fog to cool.**

Table 1: **Representative examples for each task in the Hazmat Evaluation Framework (Section 3).** Models are instructed to return *only* the final answer without additional explanation.

cess critical safety information from regulatory sources such as the publicaly available Emergency Response Guidebook (ERG). This information includes evacuation distances, protective equipment requirements, containment procedures, and first aid measures. This task evaluates both quantitative knowledge (*e.g.*, evacuation distances, time limits) and qualitative safety information (*e.g.*, protective measures, first aid guidance) extracted from the official 2024 ERG.

Rather than generating free-form recommendations, Task III assesses model comprehension of the ERG via a fill-in-the-blank approach. Specifically, we require LMs to complete a short masked passage from an actual ERG entry. For example, given a partial safety directive such as Immediately evacuate the area for at least [BLANK] meters," the model must identify the correct numerical value (e.g., 100") or qualitative phrase (e.g., "Wear protective equipment").

We construct this task by extracting passages from ERG entries across five key safety dimensions: fire or explosion, health hazards, public safety, spill or leak procedures, and first aid., and masking selected measurements and important words/action verbs from a predefined list (please see Table 6 for a list of masked words and phrases, averaging 6 characters, with 75% single-word answers) from these passages. This results in 2,400 fill-in-the-blank queries across all sampled chemicals.

We evaluate LMs via zero-shot prompting, providing the chemical name, the masked passage with context on both sides of the blank, and requesting a concise completion. This task evaluates both factual recall and comprehension of regulatory safety guidance.

3.4 Task IV: HAZMAT Report Comprehension

The HAZMAT Report Comprehension task comprises of five distinct sub-tasks designed to assess different aspects of hazardous material incident comprehension and classification, as observed in Table 2. Models are provided with real-world event narratives and publicaly available HAZMAT incident data from the PHMSA (Pipeline and Hazardous Materials Safety Administration, 2025). Together, these four classification and one identification subtasks provide a comprehensive evaluation in the hazardous materials domain.

Task IV assesses deeper incident investigation capabilities by requiring models to analyze hazmat incidents and classify them along multiple dimensions, such as identifying the component that failed, the failure mode, or the root cause. Inputs include

Task	Description	Type	Potential Labels
IV.A Hazard Class	Determine DOT hazard class from incident context	Multi-class	10 classes
IV.B Component Failure	Identify failed packaging component	Multi-class	21 components
IV.C Failure Mode	Identify how the component failed	Multi-class	12 modes
IV.D Root Cause	Infer underlying cause beyond immediate failure mode	Multi-class	19 causes
IV.E UN Number	Identify correct UN hazardous material number	Open-Ended	3000+ UN IDs

Table 2: **Task IV: Report Comprehension sub-task definitions and evaluation characteristics.** Models are provided real-world HAZMAT report event details such as event narratives and chemical name. We observe performance on 4 classification sub-tasks, as well as one open-ended identification task.

detailed incident descriptions with structured information (*e.g.*, material type, location, event narrative), as well as a selection of possible responses. We evaluate LMs via zero-shot prompting without intermediate reasoning steps, as we observed that models perform better with minimal guidance on open-ended analysis tasks. This task evaluates both analytical reasoning and domain knowledge of hazardous materials handling.

4 Evaluating LMs on ChEmREF

Models. We evaluate six different open-weight LMs on ChEmREF. Three are general-purpose models covering a range of scales: PHI-3 (3.8B) (Abdin et al., 2024a), PHI-4 (14B) (Abdin et al., 2024b), and LLAMA-3.1 (70B) (Grattafiori et al., 2024). We also include two domain-specialized models: CHEMLLM (7B) (Zhang et al., 2024) for chemistry and MED42 (70B) (Christophe et al., 2024) for medicine. Both models are adapted via supervised fine-tuning (SFT), with CHEMLLM fine-tuned from INTERN-LM-2 (7B) (Cai et al., 2024) and MED42 fine-tuned from LLAMA-2 (70B) (Touvron et al., 2023). We evaluate GPT-4O (Hurst et al., 2024) as a representative closed-source model. We also combine BM25 (Robertson and Zaragoza, 2009) lexical retrieval with GPT-4O to establish a strong retrieval-augmented baseline, where BM25 retrieves relevant incident reports and GPT-4O generates answers conditioned on the retrieved context.

4.1 Evaluation Metrics.

All models and tasks are evaluated on accuracy (for multiple choice questions), or on exact match with the ground truth, leveraging the fact that ground truth answers are explicitly defined in hazardous materials regulatory documents and incident reports. Please see Appendix C for full prompt examples.

To assess reliability on Tasks II and III, we report the mean and standard deviation of exact match (EM) accuracy under stochastic decoding. **Consistency** is measured by running each model three times at a fixed temperature ($T = 0.7$), capturing variability under identical sampling conditions. **Robustness** is measured by evaluating EM accuracy across multiple temperatures ($T \in \{0.3, 0.7, 1.2\}$), capturing sensitivity to changes in decoding stochasticity. In both cases, higher mean EM indicates stronger performance, while lower standard deviation reflects greater stability.

4.2 Overall Results

GPT-4O is the best overall, with retrieval-augmented methods and LLAMA-3.1 following closely. We present the ChEmREF evaluation results in Table 3. In addition to performance on individual tasks, we compute an overall average score over all accuracy and exact match metrics for a high-level comparison of different models. **GPT-4o + BM25** achieves the highest overall average score of 80.1%, followed by GPT-4O (68.1%) and LLAMA-3.1 (54.3%). This demonstrates the effectiveness of combining sparse retrieval with dense language models for HAZMAT-related tasks, particularly for synonym resolution (Task I) and incident recommendation (Task II) where textbook context is critical.

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 as one of the worst in synonym resolution (1.2% EM)

Model	I. Exam	II. Identification		III. Incident Response		IV. Report	Avg. (↑)
	Acc.	EM Con.	EM Rob.	EM Con.	EM Rob.	Acc.	
Phi-3 (3.8B)	49.0	0.5 ± 0.3	0.5 ± 0.3	10.8 ± 11.7	12.3 ± 10.4	32.3	23.2
ChemLLM-7B	47.3	1.2 ± 1.5	1.3 ± 1.4	20.1 ± 7.6	18.2 ± 9.1	47.1	28.9
Phi-4 (14B)	60.0	0.0 ± 0.0	0.0 ± 0.0	5.3 ± 2.5	5.2 ± 2.4	32.3	24.4
Med42-70B	58.0	4.5 ± 2.6	4.8 ± 2.4	31.7 ± 2.0	32.4 ± 1.7	49.4	35.9
Llama-3.1-70B	60.0	14.8 ± 8.4	15.6 ± 6.9	64.7 ± 1.1	63.6 ± 5.1	77.6	54.3
GPT-4o	63.9	38.2 ± 6.5	33.0 ± 11.7	83.7 ± 1.2	84.7 ± 2.3	**86.7**	68.1
BM25 + GPT-4o	†**100.0**	$\mathbf{85.2 \pm 2.1}$	$\mathbf{82.4 \pm 6.1}$	$\mathbf{97.0 \pm 0.0}$	$\mathbf{95.9 \pm 2.7}$	85.1	†**80.1**

Table 3: **ChEmREF Evaluation Results Across All Four Tasks.** Task I (Exam) and Task IV (Report) report accuracy. Tasks II (Identification) and III (Incident Response) report Exact Match consistency (mean $\pm$ std. over repeated runs at $T = 0.7$) and robustness (mean $\pm$ std. across temperatures $T \in \{0.3, 0.7, 1.2\}$). † BM25+GPT-4o achieves 100% accuracy on Task I when external reference materials are provided, rendering the closed-book exam trivial under retrieval augmentation. Aggregate averages exclude robustness and are computed using Task I accuracy (except for BM25 + GPT-4o), Task II EM consistency, Task III EM consistency, and Task IV accuracy.

and remains among the lowest performers on ERG comprehension (1.3% EM).

Notably, general-purpose models including GPT-4O, LLAMA-3.1, and PHI-4 outperform domain-specialized models on Task I: 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 their 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 techniques. We leave the exploration of more effective domain or task-specific training strategies as future work.

Retrieval augmentation excels at synonym resolution but provides limited gains on analytical tasks. Comparing **BM25 + GPT-4o** against GPT-4O alone reveals significant task-dependent benefits. On Task II (Chemical Synonym Resolution), retrieval augmentation achieves 85.2% EM versus GPT-4o's 38.2%—a 47.0 percentage point improvement. Similarly, on Task III (Incident Response), we see up to a 13.3% increase. However, on Task IV, there is no visible improvement from retrieval (-1.6%). This suggests retrieval augmentation excels at factual lookups where external knowledge is directly applicable, but provides diminishing returns for analytical and comprehension tasks requiring deeper reasoning, despite being provided resources such as the ERG.

5 Analyzing LLM performance on ChEmREF

We now further break down model performance on the four core tasks and analyze common failure cases and limitations. We highlight several key observations, with additional discussion deferred to Appendix D.

5.1 Analyzing Task I: HAZMAT Examination

Category	Mean Acc.	# 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 4: **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 4, we report the accuracy on different exam categories. Models perform well on instructional content, such as Hazard Communication and Workplace Safety (Accuracy: 75.6%), 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.

5.2 Analyzing Task II: Chemical Representation Identification

Robustness and consistency tradeoffs in retrieval-based systems. While retrieval augmentation improves identification accuracy, it does not introduce notable tradeoffs in robustness and consistency. BM25 + GPT-4o exhibits high accuracy, consistent across multiple runs of the same and variating temperature, at around 80-90% accuracy.

To Common Name vs. From Common Name. Performance varies significantly depending on whether the task involves mapping *to* or *from* a Common Name (see Appendix B). This asymmetry underscores the challenges posed by alias resolution and name normalization in real-world emergency reporting contexts.

5.3 Analyzing Task III: Incident Response Recommendation

Overall incident response quality and reliability. We observe a pronounced separation in Task III performance between large general-purpose models and smaller or domain-specific models. As shown in Table 3, GPT-4o achieves a consistent (83.7 $\pm$ 1.2) and robust (84.7 $\pm$ 2.3) EM, indicating both accurate and stable incident response recommendations. Llama-3.1-70B follows with moderate performance (64.7 $\pm$ 1.1), while smaller models such as Phi-3 and ChemLLM-7B frequently fail to generate complete or operationally correct response actions, despite occasionally exhibiting high consistency. These results suggest that incident response recommendation requires not only domain familiarity but also sufficient model capacity to integrate multiple response constraints.

Subtask-specific performance disparities. A more granular breakdown of Task III in Figure 2 (left) reveals systematic variation across incident response categories. Models generally perform better on response types that follow well-defined procedural templates, such as *spill or leak mitigation* and *first aid*, while struggling on categories that require spatial reasoning or public safety judgment, such as *public safety and evacuation*. This disparity is especially pronounced for smaller models, which often omit critical response components or provide incomplete guidance, indicating limitations in multi-step causal and situational reasoning.

5.4 Analyzing Task IV: Incident Response Comprehension

End-to-end incident comprehension. Task IV evaluates models' ability to synthesize incident descriptions into correct high-level interpretations, including hazard classification, component failure identification, failure mode analysis, root cause analysis, and UN number identification. As shown in Table 3, GPT-4o substantially outperforms all other models (86.7% accuracy), with Llama-3.1-70B achieving the second-best performance (77.6%). Smaller models exhibit markedly lower accuracy, indicating persistent difficulty in integrating multiple facets of incident information into a coherent global understanding.

Error patterns across comprehension subtasks. The Task IV breakdown in Figure 2 (right) highlights distinct error modes across comprehension categories. While hazard class identification is generally well-handled by larger models, performance degrades for component failure, failure mode, and root cause identification. These subtasks require causal inference beyond surface-level hazard recognition, and errors frequently involve misattributing failures to incorrect components or conflating immediate failure modes with underlying root causes. These results indicate that while models can recognize hazardous contexts, deeper causal reasoning and attribution remain open challenges for HAZMAT incident understanding.

UN number identification as a knowledge bottleneck. UN number identification remains one of the most challenging subtasks across all models. Even high-performing systems exhibit reduced accuracy relative to other Task IV categories, reflecting the need for precise domain-specific knowledge rather than general hazard reasoning. The strong performance of retrieval-augmented systems (BM25 + GPT-4o) on related identification tasks further suggests that access to external structured knowledge may be critical for reliable UN number inference.

6 Conclusion

We introduced ChEmREF, an evaluation framework for LLMs covering four tasks related to chemical emergencies. Our results show that incorporating retrieval assistance can significantly improve performance for the strongest out-of-the-box models on Tasks II and III; however, performance on examination-style and comprehension-based ques-

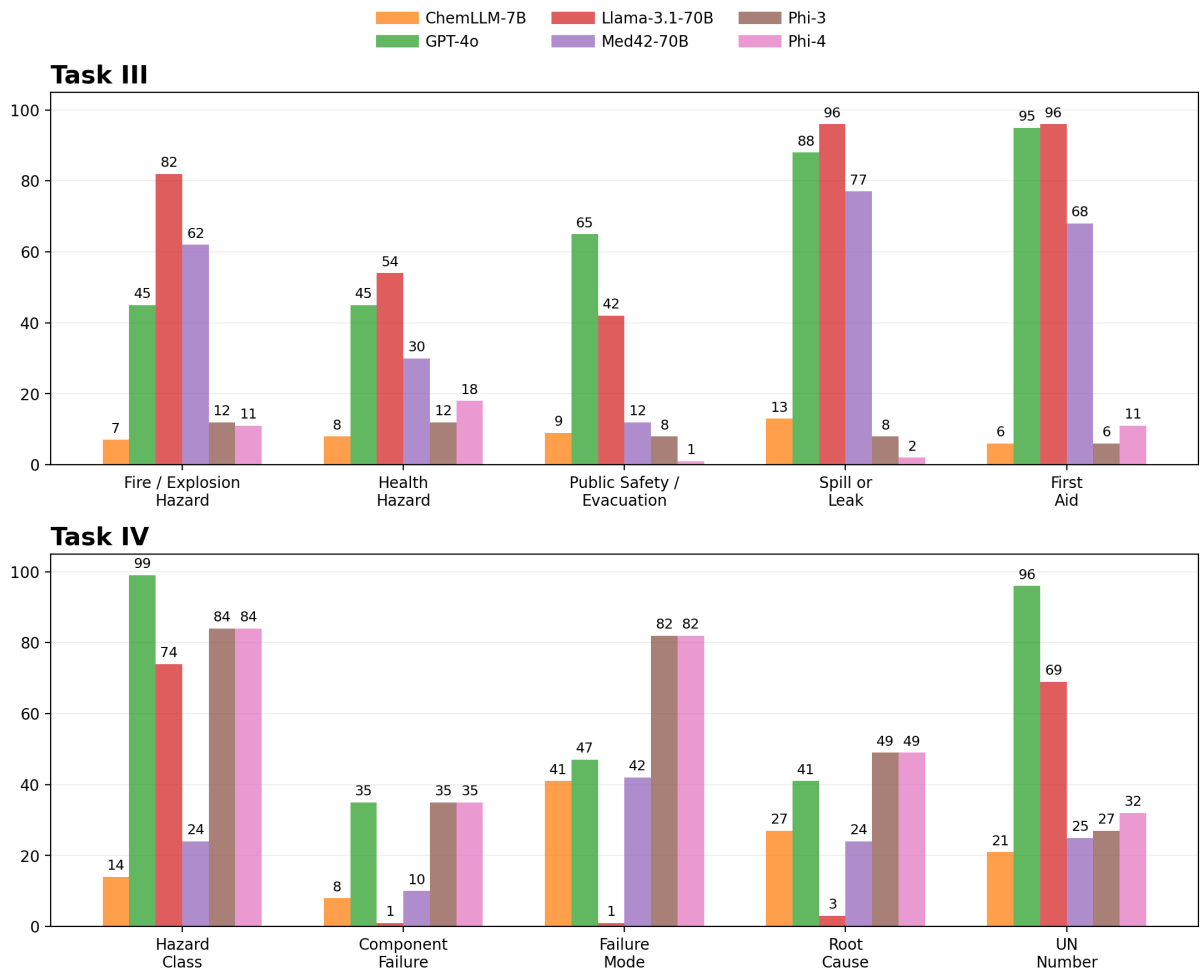


Figure 2: **Task III (left) and Task IV (right) results by prompt type** For Task III, we observe of incident response recommendation performance across sections of an ERG. For Task IV, we observe results across report comprehension sub-task.

tions in Tasks I and IV that require specialized domain knowledge remains limited. Our results 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. Addressing this human-AI communication gap should be a key consideration in the design of future AI-assisted emergency response systems. More broadly, we believe LLMs have the potential to assist with chemical identification and incident response, but should be used with caution and with human oversight. LLMs thus could be positioned as a complement to, rather than a replacement for, human experts in emergency response.

Future work could extend our ChEmREF benchmark in several important directions. First, future extensions could evaluate models with vision capabilities for interpreting cues such as HAZMAT placards and container labels, and audio capabilities for analyzing and engaging with real-time dispatch communications. Second, conducting real-world studies with domain experts and emergency responders would provide valuable validation of model outputs, help identify critical failure modes, and support future research on human–AI collaboration in high-stakes environments.

Limitations

Understanding how well LLMs perform on chemistry tasks and emergency response management is crucial to ensure their proper and safe deployment as decision-support tools. Much of the information evaluated in this benchmark is publicly available online and may have been encountered during model pretraining; while this does not invalidate the benchmark's ability to measure task performance, it limits our ability to distinguish memorization from genuine reasoning or generalization. Further, as emergency scenarios become

more complex and multifaceted, automated evaluations may fall short of capturing the full scope of required decision-making capabilities. In such cases, human evaluation from domain experts—such as HAZMAT professionals, chemical safety officers, and emergency responders—is essential to oversee the value of the benchmark. While one of the authors is a certified EMT by the NREMT and HAZMAT certified, this perspective provides limited coverage and cannot fully substitute for broader expert review across various emergency response domains.

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, and 1 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv: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, and 8 others. 2024b. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Tatiana Anikina. 2023. Towards efficient dialogue processing in the emergency response domain. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 212–225, Toronto, Canada. Association for Computational Linguistics.

Jianxiong Cai, Jiawei Hou, Yiren Lu, Hongyu Chen, Laurent Kneip, and Soren Schwertfeger. 2020. 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.

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, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. *Preprint*, arXiv: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. 2024. Med42 – evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. *Preprint*, arXiv:2404.14779.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

EMT Training Base. 2019. Scene size-up: The first step of ems calls. Accessed: 2025-02-14.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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. *Preprint*, arXiv:2305.18365.

Yang Han, Ziping Wan, Lu Chen, Kai Yu, and Xin Chen. 2025. From generalist to specialist: A survey of large language models for chemistry. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1106–1123, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Miltiadis Marios Katsakioris, Ioannis Konstas, Pierre Yves Mignotte, and Helen Hastie. 2021. Learning to read maps: Understanding natural language instructions from unseen maps. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 11–21, Online. Association for Computational Linguistics.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. 2025.

Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

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

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

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.

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, and 9 others. 2024. Are large language models superhuman chemists? *Preprint*, arXiv:2404.01475.

Aryan Naim, Ryan Alimo, and Jay Braun. 2021. Ai agents in emergency response applications. *Preprint*, arXiv:2109.04646.

Pipeline and Hazardous Materials Safety Administration. 2025. Hazmat incident reports - data mining tool. Last updated July 30, 2025.

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Amir Sharifi, Ahmadreza Zibaei, and Mahdi Rezaei. 2020. Deephazmat: Hazardous materials sign detection and segmentation with restricted computational resources. *Preprint*, arXiv: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, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Camilo Thorne and Saber Akhondi. 2024. NLP for chemistry – introduction and recent advances. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 45–49, Torino, Italia. ELRA and ICCL.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

U.S. DoT. 2024. Emergency response guidebook (erg). Accessed: 2025-02-14.

U.S. EPA. 2024. What is the cameo software suite? Accessed: 2025-02-14.

U.S. HHS. 2024a. Chemical incident management: Chemm-ist and chemm-ist for first responders. Accessed: 2025-02-14.

U.S. HHS. 2024b. The golden first minutes — initial response to a chemical hazardous materials incident. Accessed: 2025-02-14.

Yotaro Watanabe, Kentaro Inui, Shingo Suzuki, Hiroko Koumoto, Mitsuhiro Higashida, Yuji Maeda, and Katsumi Iwatsuki. 2013. Computer-assisted structuring of emergency management information: A project note. In *Proceedings of the Workshop on Language Processing and Crisis Information 2013*, pages 10–18, Nagoya, Japan. Asian Federation of Natural Language Processing.

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. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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. 2024. Chemllm: A chemical large language model. *Preprint*, arXiv:2402.06852.

A Glossary

See Table 5.

B Extended Related Work

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.

C Example Prompts and Outputs

In this section, we provide examples of full inputs and expected outputs (marked in blue) for each of the four tasks in ChEmREF.

Please see Table 6 for examples of masked words and phrases for Task III.

C.1 Task I: 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.

C.2 Task II: Identification Task

C.2.1 Unstructured Identification: Common Name Identification from Synonyms

You are an expert chemist. Question: What is the most likely common or standard name for the chemical synonym *Sodium dioxide*?
Answer: Sodium superoxide.

C.2.2 Structured Identification: Molecular Formula from Common Name

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.

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 Name)	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. Regarding SMILES evaluation, we rely on the single SMILES string that PubChem now provides for every compound.
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 U.S. Department of Transportation (DOT) Regulations.
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 5: **Key Definitions.** A glossary of important terms related to HAZMAT and emergency response.

To balance the charges, one sodium ion (Na^+) will pair with one superoxide ion (O_2^-).
Therefore, the molecular formula is NaO_2.

C.3 Task III: Incident Response Task

Scenario: A fire has started involving Mercuric nitrate (IUPAC name).
Prompt: For the chemical *Mercuric nitrate*, complete this sentence from its HAZMAT Emergency Response Guide by filling in the blank:

> Fire involving this material: Use _____ to cool containers and prevent fire spread. Avoid direct water streams that may spread contamination. Use dry chemical powder or CO_2 for fire suppression. Maintain an isolation distance of _____ and evacuate non-essential personnel.

Answer: water spray; 100 meters.

C.4 Task IV: Incident Report Comprehension Task (5 Subtasks)

Scenario: A hazmat incident report states: *"A cargo tank transporting **Mercuric nitrate** was involved in a vehicle fire. Responders applied water spray to cool adjacent containers. The report notes an evacuation of **100 meters** downwind and recommends SCBA for responders."*

Subtask IV-A (Hazard Class Classification).
Prompt: Based on the report, select the most likely hazard class for the material involved.
Options: Class 1; Class 2; Class 3; Class 4; Class 5; Class 6; Class 7; Class 8; Class 9
Answer: Class 6.

Masking Level	Targets Masked in Prompt Excerpts
Level A (Numeric patterns)	Distances/units and ranges (e.g., `100 meters`, `0.5 miles`, `3-5 km`), time durations (e.g., `10 minutes`, `2 hours`), and percentages (e.g., `5%`).
Level B (Critical keywords)	Hazard and action-critical terms, including hazard classes (e.g., `flammable`, `toxic`, `corrosive`, `oxidizer`), imperative actions (e.g., `evacuate`, `avoid`, `ventilate`, `isolate`), PPE/medical response (e.g., `respiratory protection`, `gloves`, `seek medical attention`), and fire/explosion cues (e.g., `fire`, `ignition`, `explode`, `runoff`).

Table 6: **Task III: Masking targets used to construct fill-in-the-blank prompts.** Level A masks quantitative recommendations (numbers/units). Level B masks hazard- and action-critical keywords.

Subtask IV-B (Primary Response Action Classification). Prompt: Which response action is explicitly described in the report?
Options: Evacuate area; Ventilate enclosure; Apply water spray/fog to cool; Apply alcohol-resistant foam; Neutralize spill
Answer: Apply water spray/fog to cool.

Subtask IV-C (PPE Recommendation Classification). Prompt: What PPE is recommended for responders in the report?
Options: Safety glasses only; N95 respirator; Self-contained breathing apparatus (SCBA); No PPE required
Answer: Self-contained breathing apparatus (SCBA).

Subtask IV-D (Evacuation Direction Classification). Prompt: What evacuation direction is specified?
Options: Upwind; Downwind; Upstream; No direction specified
Answer: Downwind.

Subtask IV-E (Evacuation Distance Extraction / Classification). Prompt: What evacuation distance is specified in the report?
Options: 30 meters; 50 meters; 100 meters; 300 meters
Answer: 100 meters.

D Extended Results

In this section, we provide supplementary figures and analysis for the Task I: HAZMAT Examination and Task II: Identification.

D.1 Analyzing Task I: HAZMAT Examination

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.

D.2 Analyzing Task II: Chemical Representation Identification

In Figure 3, we break down the performance on Task I by source and target chemical representations (defined in Table 7. A The reported performance is averaged with all six evaluated models.

Structured Identification vs. Unstructured Identification. We observe that models perform significantly better when translating between *structured* chemical formats such as IUPAC Name, Molecular Formula, InChI, and SMILES[3], compared to identification involving *unstructured* forms like Common Name, UN-Name, and Synonyms. For example, PHI-4 achieves a EM of 97.2% on structured identification, but the performance drops to 48.7% on unstructured identification. 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 identification 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

[3]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).

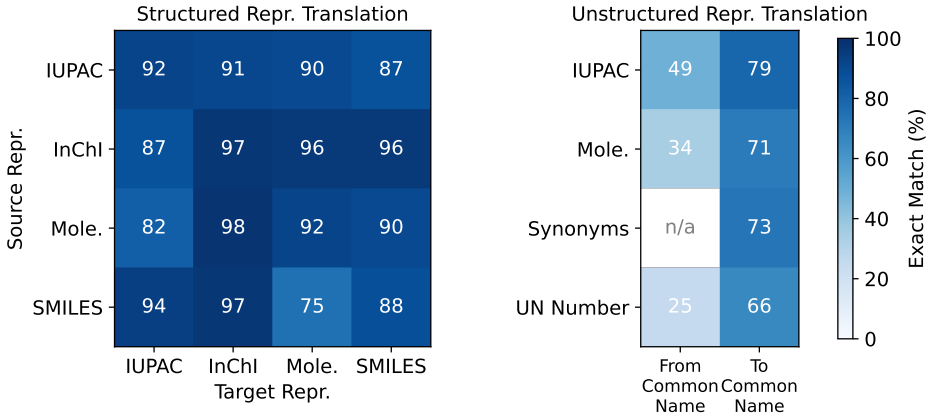


Figure 3: **Task I Performance Breakdown. Left/Right: Structured/Unstructured Identification.** For brevity, we use "Mole." to denote Molecular Formula. We report EM scores for each source-target pair, averaged across all evaluated models.

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 7: **Different Chemical Representations of tert-Butyl alcohol, a sample hazardous material.** In Task II: Chemical Identification (§3.2) 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 guidebooks.

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. The identification of a 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.

E Self-identification

Self-translations (*e.g.*, from SMILES to SMILES) is a simple extension from the identification task. We additionally evaluate models on self-identification on the four structured representations, as a sanity check. We report the detailed model-specific results in Fig. 9. 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.

E.1 Prompting Methods

For Task II, 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 identification 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.

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. As we provide the model with less examples, for a model like ChemLLM,

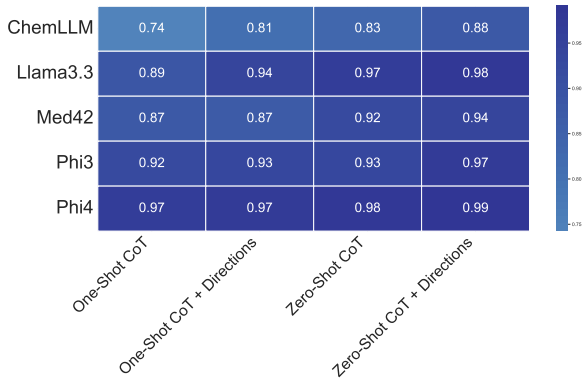


Figure 4: **Identification Task Model Performance Across Prompt Types:** Performance heatmap using HAZMAT data.

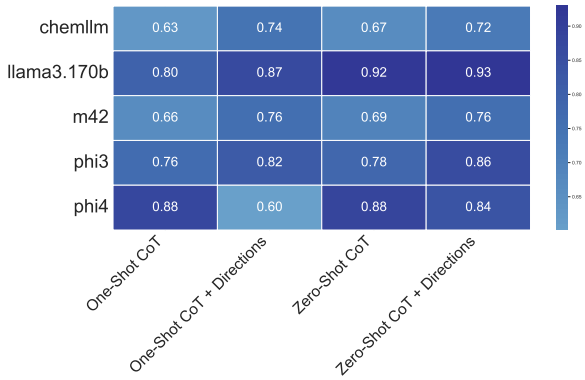


Figure 5: **Identification Task Model Performance Across Prompt Types:** Performance heatmap using Non-Hazmat chemical data.

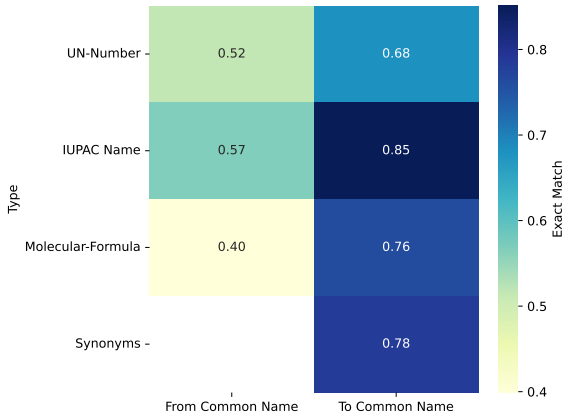


Figure 6: Exact Match Heatmap: Best Performing Model (GPT-4o).

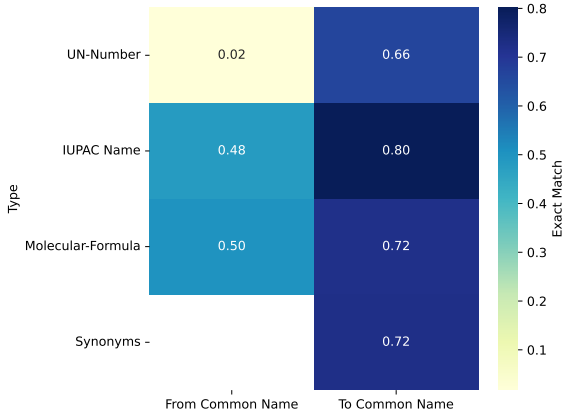


Figure 7: Exact Match Heatmap: Worst Performing Model (Phi-3).

prompt format can mean an improved score of up to 14% on HAZMAT chemicals. Alternatively, a model like Phi-4 shows minimal performance variation across prompt formats for HAZMAT chemicals, but exhibits significant variation on non-HAZMAT data, achieving performance comparable to ChemLLM on HAZMAT data. We hypothesize that including too much chemical instruction in the prompt can sometimes lead the model to produce outputs that closely mirror the examples provided, rather than generating more generalizable responses. 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.

Insights from this pilot study informed our decision in Section 4.1 to adopt one-shot Chain-of-Thought (CoT) prompting for the structured identification task. This approach served as a balanced middle ground among prompt types, providing sufficient contextual guidance without excessive complexity. Alternatively, for the unstructured identification task in our main experiments, we employed zero-shot prompting. 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 identification setting.

E.1.1 Chemical Properties

The provided heatmaps in Figure 8 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

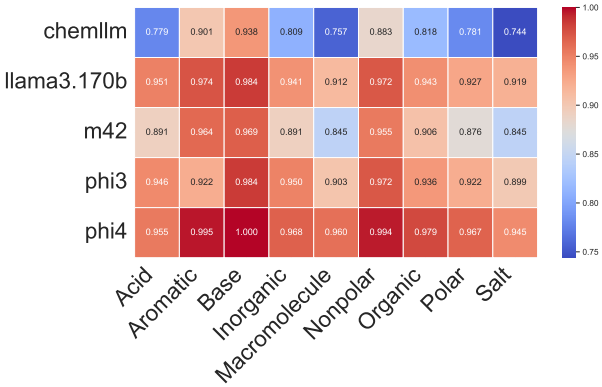


Figure 8: **Exact Match Scores for Model Performance on `hazmat` Data Across Chemical Categories.** This heatmap illustrates how different models perform when translating hazardous material data, highlighting variations in accuracy across chemical types.

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.

E.2 Input-Output Types

Identification performance rarely varies with source chemical representation type for structured representation types. Figure 9 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 identification.

Identification 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 identification 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.

E.2.1 HazMat Chemicals v.s. Non-HazMat chemicals.

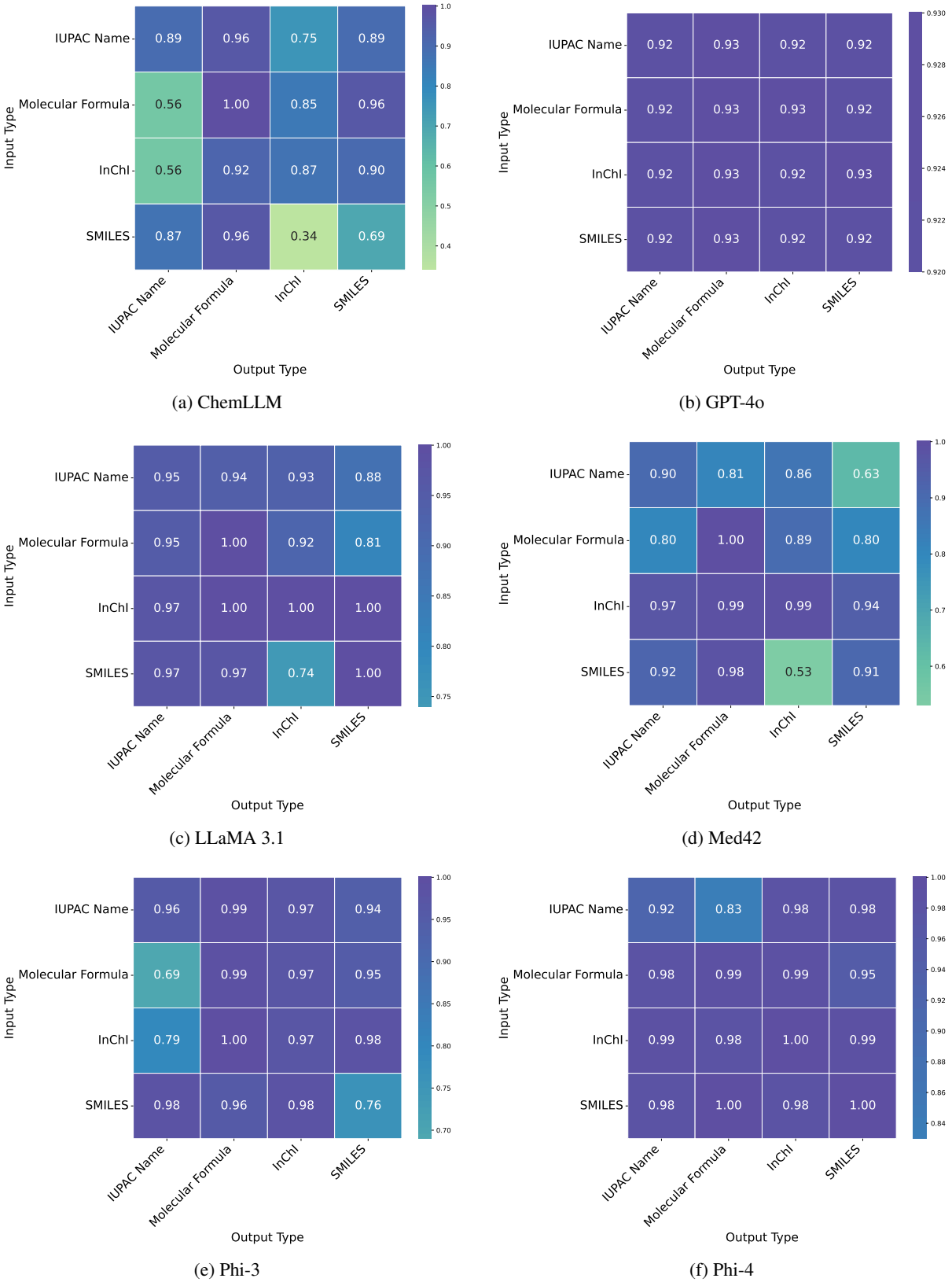
Identification of more common HazMat chemicals is easier than non-HazMat chemicals. In this section, we compare Chemical Familiarity by HazMat vs Non-HazMat data. Non-HazMat Data was acquired from the PubChem database, and a subset of 100 chemicals were selected.

Overall, in Figure 10, the Identification Task on HazMat data received a 90.9%, whereas the Identification task on Non-HazMat data scored a 77.7%. The significant improvement in chemical comprehension and identification 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.

F AI Statement

During the preparation of this work, the author(s) used AI assistants to improve the readability and language of the manuscript at a small, formatting and sentence-level scale. After using these tools, the author(s) carefully reviewed, edited, and adapted all content.



(a) ChemLLM

(b) GPT-4o

(c) LLaMA 3.1

(d) Med42

(e) Phi-3

(f) Phi-4

Figure 9: **Average Identification Task EM `hazmat` Scores Across Input/Output Representation Type.** This figure presents Exact Match (EM) scores for the Identification Task, grouped by model.

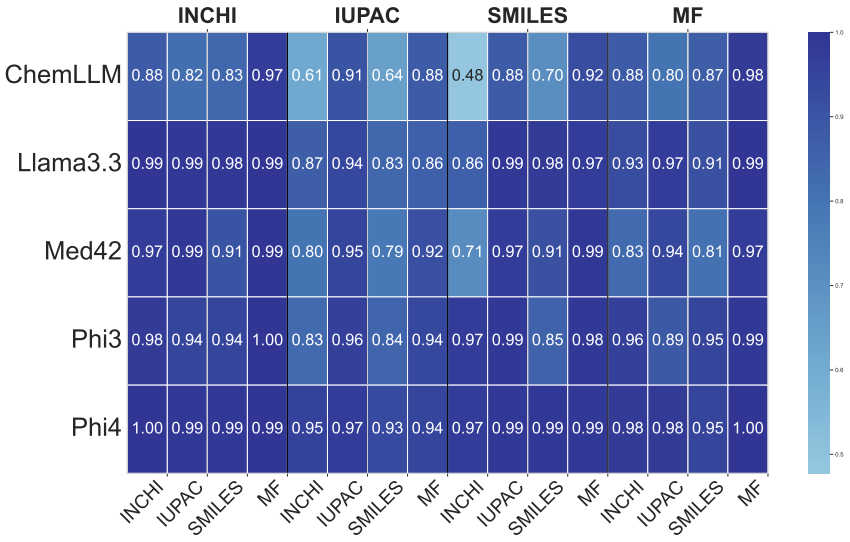


Figure 10: **EM Averages for HAZMAT Data across Input and Output Representation Types.** Each heatmap cell corresponds to a identification task from an **input type** (top row) to an **output type** (bottom column). Note that these results are based on a different evaluation cohort (100 distinct chemicals) compared to Fig. 9, leading to variation in EM scores.

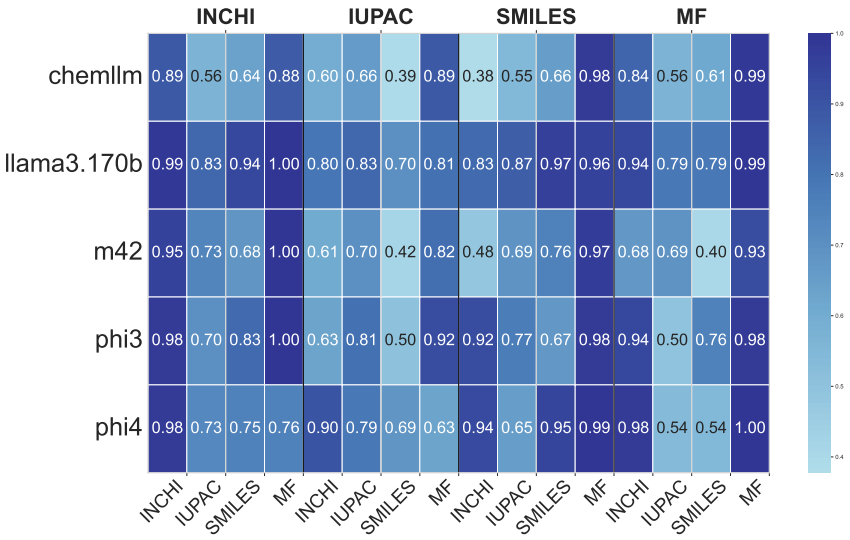


Figure 11: **EM Averages for Non-HAZMAT Data across Input and Output Representation Types.** Similar to Fig. 10, this heatmap reflects translations between input (top row) and output (bottom column) types using a separate cohort of PubChem chemicals.