Bridging AI and Carbon Capture: A Dataset for LLMs in Ionic Liquids and CBE Research

Gaurab Sarkar[1*] **and Sougata Saha**[2*]
[1]State University of New York at Buffalo
[2]Mohamed bin Zayed University of Artificial Intelligence
[1]gaurabsa@buffalo.edu, [2]sougata.saha@mbzuai.ac.ae

Abstract

Large Language Models (LLMs) have demonstrated exceptional performance in general knowledge and reasoning tasks across various domains. However, their effectiveness in specialized scientific fields like Chemical and Biological Engineering (CBE) remains underexplored. Addressing this gap requires robust evaluation benchmarks that assess both knowledge and reasoning capabilities in these niche areas, which are currently lacking. To bridge this divide, we present a comprehensive empirical analysis of LLM reasoning capabilities in CBE, with a focus on Ionic Liquids (ILs) for carbon sequestration—an emerging solution for mitigating global warming. We develop and release an expert-curated dataset of 5,920 examples designed to benchmark LLMs' reasoning in this domain. The dataset incorporates varying levels of difficulty, balancing linguistic complexity and domain-specific knowledge. Using this dataset, we evaluate three open-source LLMs with fewer than 10 billion parameters. Our findings reveal that while smaller general-purpose LLMs exhibit basic knowledge of ILs, they lack the specialized reasoning skills necessary for advanced applications. Building on these results, we discuss strategies to enhance the utility of LLMs for carbon capture research, particularly using ILs. Given the significant carbon footprint of LLMs, aligning their development with IL research presents a unique opportunity to foster mutual progress in both fields and advance global efforts toward achieving carbon neutrality by 2050. Dataset link: https://github.com/sougata-ub/llms_for_ionic_liquids

1 Introduction

Despite notable advancements in modeling and simulation methods (van Gunsteren and Mark, 1998; van Gunsteren et al., 2018; Frenkel and Smit, 2023), fundamental research in CBE continues to rely heavily on experimental results. As computational models (Zhao et al., 2023), LLMs are predominantly advantageous in computation-intensive fields, making their precise role in enabling progress within experiment-driven domains like CBE unclear. Nonetheless, recent breakthroughs in material discovery (Lu et al., 2023; Luu and Buehler, 2024; Lu et al., 2024; Buehler, 2023b) and protein engineering (Jumper et al., 2021; Liu et al., 2022; Yu et al., 2022b,a; Hu and Buehler, 2022; Khare et al., 2022) demonstrate the potential of AI technologies to contribute meaningfully to such fields. To unlock the potential applications of LLMs in CBE, it is critical to assess their knowledge and reasoning capabilities. However, this requires robust, domain-specific evaluation benchmarks, which are currently lacking in CBE.

While evaluation frameworks exist in related fields, they predominantly rely on cloze-style tasks to assess LLMs' knowledge capacity or focus on narrow, task-specific evaluations (Zhao et al., 2024; Murakumo et al., 2023; Zhang et al., 2024; Guo et al., 2023; Bran et al., 2023). Such approaches are often insufficiently general and may not adequately capture the complexities of CBE. Given that LLMs have been trained on a vast corpus of publicly available online data (Villalobos et al., 2022, 2024), studies (Chu et al., 2025) have shown that these models can easily memorize and regurgitate information during cloze-style factual assessments. This limitation provides only a superficial understanding of LLM capabilities across domains. Furthermore, the concept of knowledge extends beyond factual recall to include its application (*p-knowledge*) (Fierro et al., 2024). Therefore, evaluating knowledge capacity alone fails to capture reasoning ability, hindering the practical deployment of LLMs, particularly in fields like CBE, where their utility remains uncertain. To address this gap, we introduce a reasoning evaluation test-bed de-

[*]Both authors contributed equally to this paper.

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 32–48

signed to more effectively estimate LLMs' applicability in such domains.

Global warming caused by greenhouse gas emissions remains a critical challenge (Wang et al., 2016; Sanz-Pérez et al., 2016), necessitating accelerated research into effective carbon capture solutions (Sheridan et al., 2018). Meeting the ambitious carbon-neutral target of the 2015 Paris Agreement by 2050 (Rhodes, 2016) requires not only reducing carbon emissions but also investing in technologies to remove CO_2 from the atmosphere. Among potential solutions, *Ionic Liquids* (ILs) (Zanco et al., 2021) stand out as promising candidates for CO_2 separation processes due to their non-volatile, non-toxic nature (*"green solvents"*), ease of regeneration, and high CO_2 absorption efficiency. However, experimentation with ILs and achieving industrial scalability are resource-intensive and costly, a challenge that AI technologies like LLMs could help address. In this paper, we take a foundational step toward exploring the role of LLMs in supporting carbon capture research using ILs. Specifically, we assess the potential of general-purpose LLMs in domain-specific scenarios by constructing a test bed of 5,920 expert-curated examples, spanning varying levels of difficulty, to evaluate the factual knowledge and reasoning capabilities of these models in the context of ILs. We benchmark three open-weight LLMs—Llama 3.1-8B (Dubey et al., 2024), Mistral-7B (Jiang et al., 2023), and Gemma-9B (Team et al., 2024)—on this dataset. Given the absence of prior research in this area, our work represents a critical step toward identifying the potential applications of LLMs in IL research. Furthermore, leveraging LLMs for CO_2 capture research offers an opportunity to indirectly address concerns about their environmental impact (Patterson et al., 2021; Strubell et al., 2019; Faiz et al., 2024; Li et al., 2023; Rillig et al., 2023) by aligning their use with climate solutions. Our contributions are as follows:

- **Dataset Creation:** Using ILs for carbon capture as a use case, we construct and publicly share[1] a textual entailment test bed containing 5,920 expert-curated samples designed to evaluate LLM reasoning capabilities in CBE.
- **Benchmarking:** We systematically benchmark three open-weight LLMs—Llama 3.1-8B, Mistral-7B, and Gemma-9B—on the test bed and share the resulting insights.
- **Analysis:** We discuss the implications of our results and the broader potential for LLMs to advance IL research and CO_2 capture technologies.

[1]Dataset available at: https://github.com/sougata-ub/llms_for_ionic_liquids

2 Related Work

2.1 Ionic Liquids for Carbon Capture

COP21 showed that amongst 196 participating countries, China, the United States, and India comprise the top three nations by share of worldwide CO_2 emissions. While the United States has pledged to reach "net-zero" by 2050, the deadlines set by China and India (the two most populous countries) are 2060 and 2070 respectively (Rhodes, 2016; Guiot and Cramer, 2016; Dimitrov, 2016; Robbins, 2016). In an attempt to offset the rising atmospheric carbon dioxide levels, carbon sequestration has emerged as an effective field of research and the timely development of materials and methods is pivotal for the efficient capture of CO_2 (Wang et al., 2016; Sanz-Pérez et al., 2016). Ionic Liquids have presented themselves as an excellent solution for CO_2 capture due to their environmentally friendly nature (Blanchard et al., 1999, 2001; Pérez-Salado Kamps et al., 2003; Anthony et al., 2002; Zeng et al., 2017; Husson-Borg et al., 2003; Aghaie et al., 2018; Ramdin et al., 2012). Thorough experimentation, with ILs, to provide a practical solution is time-conducive and entails high cost (Sheridan et al., 2018; Maginn, 2009). In that regard, various machine learning methods have found use to alleviate dependence on experiments (Cao et al., 2018; Baskin et al., 2022; Dhakal and Shah, 2022; Feng et al., 2022; Paduszynski, 2016).

2.2 LLMs for Scientific Research

Recently, LLMs (Brown, 2020; Chowdhery et al., 2023; Taylor et al., 2022; OpenAI et al., 2024) have gained significant popularity with a wide range of possibilities (Ge et al., 2024; Bubeck et al., 2023; Nadkarni et al., 2021; Beltagy et al., 2019; Schick et al., 2024; Buehler, 2023a; Luu et al., 2023; Mialon et al., 2023; Wei et al., 2023), and the integration of these transformer-based models into the fields of materials science and discovery has yielded tremendous results. Leveraging the abilities of LLMs has been beneficial in various downstream tasks such as protein design and folding (Jumper et al., 2021; Liu et al., 2022; Yu et al., 2022b,a; Hu and Buehler, 2022; Khare et al., 2022), material discovery (Lu et al., 2023;

Luu and Buehler, 2024; Lu et al., 2024; Buehler, 2023b), educational tasks (Lim et al., 2023; Milano et al., 2023; Inguva et al., 2021) and chemistry-related tasks (Castro Nascimento and Pimentel, 2023; White, 2023; Jablonka et al., 2023). The reliability of LLMs is still a massive topic of discussion, and their accuracy is often determined by the size and complexity of the model. Despite their promises, present pitfalls include the issues of hallucinations and fact recall, which warrants a careful validation of the model's output and its eventual ramifications (Hu and Buehler, 2023; Azamfirei et al., 2023; Kandpal et al., 2023; Varshney et al., 2023; Ji et al., 2023; McKenna et al., 2023; Harrer, 2023). Invariably, training and using such networks comes at a huge environmental cost, largely in terms of carbon emissions (Li et al., 2023; Patterson et al., 2021; Strubell et al., 2019; Faiz et al., 2024; Rillig et al., 2023).

The power of LLMs can aid carbon capture by helping researchers with their advances to address the growing problem of global warming and offset the model's carbon footprint to reach the end goal of 'net-zero' carbon emissions.

3 A Practical Test for Knowledge

Although there are several standard definitions of knowledge in Philosophy (Sartwell, 1992; Nozick, 2016; Williamson, 2005; Zagzebski, 2017; Austin, 1961), the most prevalent ones for non-human entities like LLMs are *tb* and *p-knowledge* (Fierro et al., 2024). Most knowledge probing tasks test for *tb-knowledge*, where the model passes the test if it can recall an answer. For example, probing for factual questions like "What is the capital of Germany?" Such tests are weak estimates of knowledge and hold little pragmatic significance, especially in domains like CBE, where the intended use of LLMs is still unclear. LLMs as reasoners can be of better practical use in such domains. Although some methods estimate the model's uncertainty (Huang et al., 2024, 2023; Ye et al., 2024; Geng et al., 2023), they still pertain to *tb-knowledge*. However, a more complete measure of knowledge is *p-knowledge*, which tests a model's capability to use knowledge in practical tasks. For example, sociodemographic prompting (Saha et al., 2025; Pandey et al., 2025; Li et al., 2024b; AlKhamissi et al., 2024; Nadeem et al., 2021; Nangia et al., 2020; Wan et al., 2023; Jha et al., 2023; Li et al., 2024a; Cao et al., 2023; Tanmay et al., 2023; Rao et al., 2023) such as "What would a German find difficult to understand from a text X?" necessitates a model to reason from a group's perspective, which requires prerequisite knowledge. Motivated to create stronger test beds, we set up an *entailment task* to benchmark LLMs' factual capacity in CBE, where the model is provided a claim and a list of propositions and is tasked with determining all propositions that entail the claim or none. Thus, testing the model's reasoning capabilities in a practical setting is warranted.

3.1 Argument Structures

A claim constitutes one or more facts (propositions), where some are evident (explicit) from the text, and some are assumed (implicit) to be known by the reader (enthymemes) (Walton, 1996; Besnard and Hunter, 2008; Walton et al., 2008; Bitzer, 2020). Within a field (such as CBE), the degree of knowledge of the assumed propositions is subjective and varies by person, which impacts the understanding of the claim. For example, the claim "Ionic Liquids are low-melting, non-volatile salts which categorize them within the green solvents category" explicitly informs that (i) Ionic Liquids are low-melting, non-volatile salts. (ii) Ionic Liquids are categorized as green solvents. It also entails that low-melting and non-volatile salts are green solvents, which might be unknown (or partially known) to someone from CBE[2]. The degree of knowledge about the implicit assumption is subjective and varies within the domain[3]. We aim to test this domain-specific knowledge in LLMs via an entailment task.

3.2 The Entailment Task

Hypothesizing that **knowledgeable agents should perform consistently, irrespective of the adversaries**, we create an *entailment task* with the following setups to benchmark LLMs' reasoning capacity, where the model is provided a claim and a list of propositions and tasked to determine all propositions that entail the claim, if applicable.

1. Change the number of adversaries: (i) Keeping the number of entailing propositions constant for a claim, the number of non-entailing propositions should not affect the model's entailment

[2]This is different from general knowledge. For example, understanding the claim also requires knowledge of "low-melting, non-volatile salts" and "green solvents", which is an assumed prerequisite for a domain expert.

[3]We are only interested in domain-specific knowledge. An outsider might not possess such knowledge.

performance. (ii) When provided with only non-entailing options and an additional *"none of the above"* option, a consistent agent should always choose the *"none"* option. A drop in performance indicates a lack of knowledge and supposedly more reliance on linguistic cues for entailment.

2. Introduce linguistic perturbations: A knowledgeable agent should be invariant to paraphrased options. Failure to do so indicates reliance on linguistic cues instead of factual cues for entailment.

2. Apply common sense: Knowledgeable agents should not be derailed by incorrect facts that can be discerned by common sense.

3.3 Dataset Creation

The dataset is created in multiple phases, employing two expert annotators, one with a background in CBE and another from Computer Science and Linguistics (CSL). The CBE expert has domain knowledge of ILs for carbon capture, while the CSL expert is generally unaware of the domain. Figure 1 illustrates the data creation pipeline with an actual example. We detail the pipeline below:

Phase 1 encompassed knowledge creation, where the CBE expert constructed paragraphs capturing the different aspects of carbon capture using ionic liquids. The aspects encompassed the need for carbon capture, ionic liquids, their physical and chemical characteristics, and their advantages. Next, the annotator extracted claims from the paragraphs, which are sentences containing salient knowledge pertaining to ionic liquids for carbon capture, yielding 74 in total.

Phase 2 encompassed identifying the explicit and implicit propositions from each claim and implemented in two stages: (i) **LLM-based annotation:** We prompted Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) to identify the explicit and implicit propositions from a claim. As depicted in Figure 5 (Appendix A), the prompt comprised a short task description and three examples of how to perform the task, followed by the actual claim for annotation. (ii) **Expert evaluations:** The CBE expert extensively evaluated the model response by editing, deleting, or unchanging each model-identified proposition. Additionally, for each claim, the expert added propositions that were missed by the model (if any). Overall, 48 (65%) of the 74 LLM-based annotations were deemed correct by the expert and were unmodified, yielding 164 propositions across 74 claims.

Phase 3 encompassed data standardization. The propositions, being fundamental pieces of knowledge, are universal. Hence, in this phase, we standardized the propositions across all claims. Using sentence transformers (Reimers and Gurevych, 2019), we clustered the propositions by their embedding cosine-similarity[4] and computationally marked propositions belonging to the same cluster as equivalent. The CBE expert evaluated the clustering results, which were accurate in only 28% of cases. The expert annotated and rectified the incorrect cluster assignments, yielding 125 universal propositions across all 74 claims.

Phase 4 involved constructing false variants of the propositions at three difficulty levels: (i) **Low:** Invalid version of a proposition, and can be discerned using common sense reasoning. For example, the proposition *"Ionic liquids can be categorized as conventional or task-specific"* was augmented to *"Ionic liquids can be categorized as conventional or task-specific only while recharging batteries."* (ii) **Medium:** Invalid version of a proposition that might need a mix of common sense and knowledge of science for discerning. For example, *"Ionic liquids can be categorized as conventional or task-specific due to specific environmental conditions and chemical habitability."* (iii) **High:** Determining invalidity requires considerable knowledge about ILs. For example, *"Ionic liquids can be categorized as conventional or task-specific based on molecular weight, isotope atom count, and hydrogen bonding capabilities."* All variants were manually constructed by the CBE expert and evaluated by the CSL expert, who does not know ILs. The CSL expert evaluated 60 random propositions (15 original and 15 from each level of difficulty) by determining if the proposition was correct or assigning a level of difficulty if they thought it was incorrect. Comparing their response with the original labels, the expert attained an F1 score of 67% in discerning factual correctness. For the incorrect propositions, the expert attained F1 scores of 80%, 15%, and 42% for levels 1, 2, and 3, indicating the difficulty of the options for a non-expert.

In **phase 5**, we introduced linguistic variations in the original and all three incorrect variants of each proposition by paraphrasing. We prompted the Llama-3.1-8B instruction-tuned variant (Dubey et al., 2024) using the prompt *"Paraphrase the following text without changing the meaning of the*

[4]We used the 'all-MiniLM-L6-v2' model for computing embeddings.

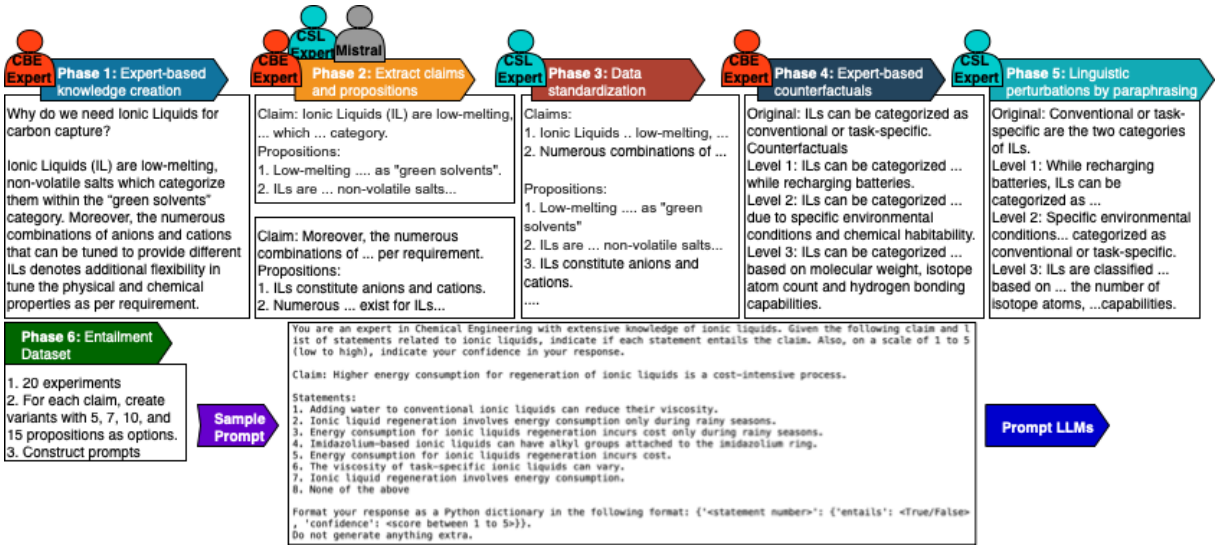

Figure 1: Dataset Creation Pipeline

Group	Description	Id	Experiment	Correct options		Incorrect options		Median F1			Std Dev F1		
				Present	Para-phrased	Difficulty	Para-phrased	Gemma	Llama	Mistral	Gemma	Llama	Mistral
0	Baselines	1	orig+random	Yes	No	Random	No	49.0	**66.0**	55.0	21.7	**16.1**	18.9
		2	para+random		Yes			49.5	**63.0**	57.0	21.8	**14.0**	18.5
1	Only providing incorrect options	3	none+level1	No	No	Level 1	No	9.0	**30.0**	1.5	5.5	6.8	**3.1**
		4	none+level2			Level 2		2.0	**29.0**	0.0	4.9	12.3	**1.0**
		5	none+level3			Level 3		3.0	**21.5**	0.0	4.3	8.6	**0.5**
		6	none+level1-para			Level 1	Yes	3.0	**17.5**	0.0	1.7	7.3	**0.5**
		7	none+level2-para			Level 2		0.0	**17.5**	0.0	2.0	6.4	**0.5**
		8	none+level3-para			Level 3		1.0	**18.5**	0.0	5.3	2.4	**0.0**
2	Difficulty level of incorrect options	9	orig+level1	Yes	No	Level 1	No	35.0	**73.5**	62.5	26.6	**14.6**	19.6
		10	orig+level2			Level 2		32.0	**68.5**	60.5	19.9	**12.3**	18.7
		11	orig+level3			Level 3		29.5	**67.5**	58.0	18.0	**12.1**	16.9
3	Paraphrasing the correct options	12	para+level1	Yes	Yes	Level 1	No	40.5	**66.5**	62.0	24.1	**14.4**	20.1
		13	para+level2			Level 2		34.5	**66.0**	60.0	23.3	**13.5**	18.1
		14	para+level3			Level 3		31.5	**64.0**	60.0	18.8	**13.0**	17.0
4	Paraphrasing the incorrect options	15	orig+level1-para	Yes	No	Level 1	Yes	39.5	**66.0**	57.0	22.1	**11.7**	17.3
		16	orig+level2-para			Level 2		35.5	**64.5**	57.5	17.0	**11.1**	17.1
		17	orig+level3-para			Level 3		34.0	**63.0**	58.5	13.5	**10.5**	16.6
5	Paraphrasing all options	18	para+level1-para	Yes	Yes	Level 1	Yes	33.5	**63.5**	58.0	**10.7**	11.2	16.9
		19	para+level2-para			Level 2		35.5	**64.5**	59.5	15.6	**11.8**	17.6
		20	para+level3-para			Level 3		36.5	**60.5**	58.5	11.9	**10.8**	17.8

Table 1: Definitions of experiments and aggregated model results (median F1 and standard deviation) across experiments with 5, 7, 10, and 15 options. The best scores are highlighted in bold.

text. Text: <text>" and resorted to greedy decoding for paraphrasing.

Using the 74 claims, the 125 original propositions, and their incorrect and paraphrased variants, we constructed the test set for the entailment task in **phase 6**. For each claim, we created variants with 5, 7, 10, and 15 propositions as options. Listed in Table 1, we constructed 20 experiments using different permutations of the original and paraphrased versions of the correct and incorrect propositions, yielding a dataset of 5,920 examples. On average, the claims contain 14 words, the original propositions contain 12, and the incorrect propositions contain 17.

3.4 Experiments

Listed in Table 1, we group the 20 experiments into five groups and test our three hypotheses in Section 3.2. Comprising two experiments, Group 0 serves as the baseline. By only providing incorrect propositions with their difficulty and stylistic variations, the experiments in Group 1 test the model's knowledgeability by measuring the propensity of selecting the "none" option. Group 2 quantifies the effect of varying the difficulty levels of the incorrect options while keeping the original propositions unchanged. Group 3 perturbs the original proposition by paraphrasing and measures the impact of changing the incorrect option difficulty levels. Groups 4 and 5 measure a model's invariance to linguistic

variations and the difficulty levels of the incorrect options. For all groups, we experiment with 5, 7, 10, and 15 propositions as options to test for the model's capability of being invariant to additional incorrect options. Except for group 1, the number of correct options varies from 1 to 5. As depicted in Figure 1 (Phase 6), we construct prompts from each example and probe the Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and gemma-2-9b-it, setting the temperature to 0. We process each model response and use Llama-3.1-8B-Instruct to rectify ill-formatted outputs. Figures 6 and 7 (Appendix A) illustrate the prompts for the entailment task and correcting the ill-formatted LLM responses.

4 Results and Observations

Table 1 shares the model-wise median F1 score and the standard deviation across all options (5, 7, 10, and 15). Figures 2, 3, and 4 plot the precision and recall scores for all groups of experiments. The baseline results (Group 0) in Table 1 indicate that LLMs are knowledgeable about ionic liquids and carbon capture. Llama performs the best, followed by Mistral and Gemma. However, paraphrasing the original propositions (Id 2) reduces Llama's performance, which contrasts with Mistral and Gemma, where the performance increases. This effect of stylistic perturbations on the model results shows a tendency to rely on linguistic cues.

Effect of the number of incorrect options
We observe a correlation between model performance and the number of incorrect options in Figures 2, 3, and 4. The precision scores for all models drop with more incorrect options, indicating an adverse effect of the number of adversaries on their reasoning capabilities. For Llama and Mistral, the recall scores remain mostly consistent, but drop for Gemma. Nonetheless, as depicted in Table 1, the standard deviation of Llama is the lowest, followed by Mistral and Gemma. For Llama and Mistral, this decline in precision but constant recall scores indicates a propensity to make more predictions as the number of options increases without changing the prediction for the correct propositions. On the contrary, increasing the number of adversaries causes Gemma to change the prediction for the correct propositions, indicating an unreliability of utilizing facts for reasoning.

Effect of the difficulty of incorrect options
Comparing experiment Id 1 with Group 2 and Id 2 with Group 3 in Table 1 and Figure 2, we observe that increasing the difficulty level of the adversarial facts hampers the model performance for Llama and Mistra, which is the opposite for Gemma. The comparisons indicate that the experiments comprising random adversaries (Orig/para+random) are more challenging test beds than the difficulty-controlled adversaries, especially for Llama and Mistral. We hypothesize that since we gradually balance between common sense and domain-specific knowledge across three difficulty levels, higher performance in level 1 can be due to the model's capability of common sense reasoning, which decreases as the difficulty increases, requiring more domain-specific knowledge. However, using random adversaries presents less scope for common-sense reasoning and requires domain-knowledge-based reasoning for entailment resolution. Gemma, on the other hand, is more reliant on syntactic cues than reasoning. Hence, it falters when provided with factually incorrect yet syntactically similar options to the claim. This is also evident from Gemma's decreasing recall scores in Figure 2, compared to Llama and Mistral, which are more consistent.

Effect of only incorrect propositions as options
Compared to the baseline (Group 0) in Table 1, in Group 1, the performance of all models drastically reduces when presented with only incorrect facts and a "none" option to choose from. Mistral and Gemma perform worse than Llama, with median F1 scores < 10 for all experiments and near zero for some. All models perform worse with paraphrased incorrect options. Figure 4 plots the **pr**ecision, **re**call, and **f1** scores for Group 1 experiments. Interestingly, for all three models, sometimes the precision increases with higher options in some experiments. For Gemma, the precision scores increase while the recall decreases with an increase in incorrect choices. On the contrary, for Llama and Mistral, the precision and recall scores increase for some experiments. For Llama, presenting 7 and 10 options yields higher F1 scores for most experiments compared to 5 options. Mistral yields higher F1 scores when prompted with 7 choices compared to other options. We hypothesize that for Llama and Mistral, increasing the choices provides more inter-option reasoning opportunities, resulting in higher F1 scores. We also think the position of the "none" option in the prompt might be a confounding variable, which we leave for future work. Nonetheless, when only presented with incorrect facts and a "none" option, the drastic reduction in

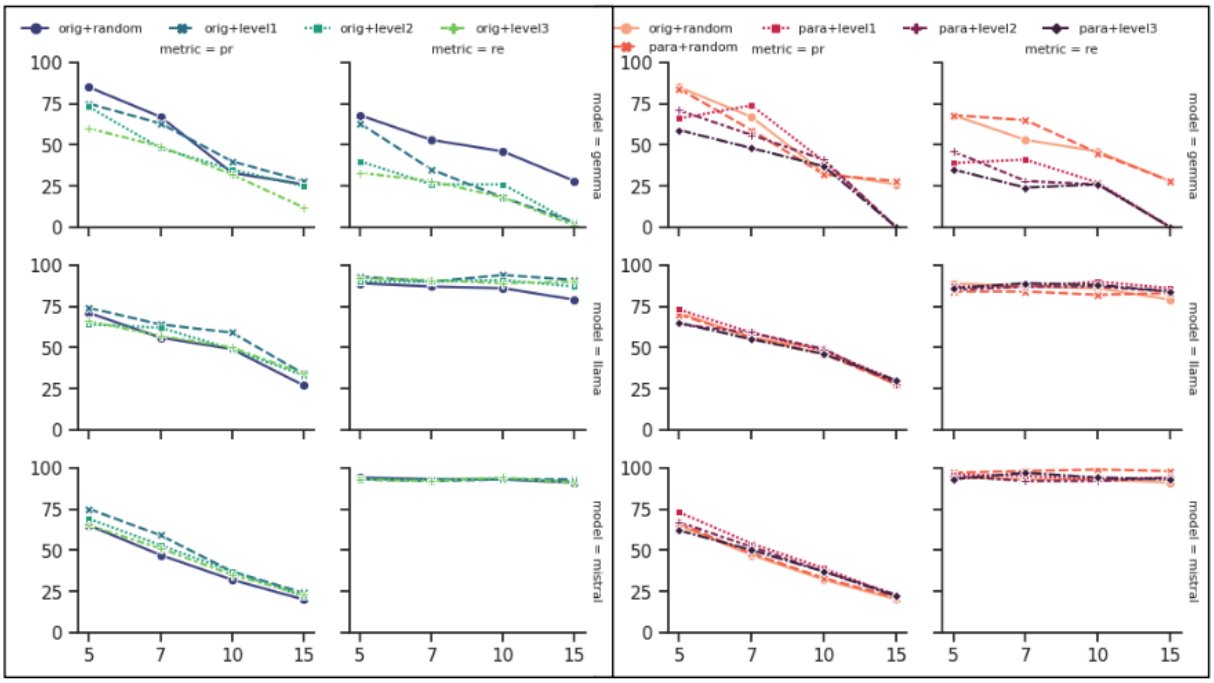

Figure 2: Model-wise precision and recall for experiments in Group 2 (left) and Group 3 (right).

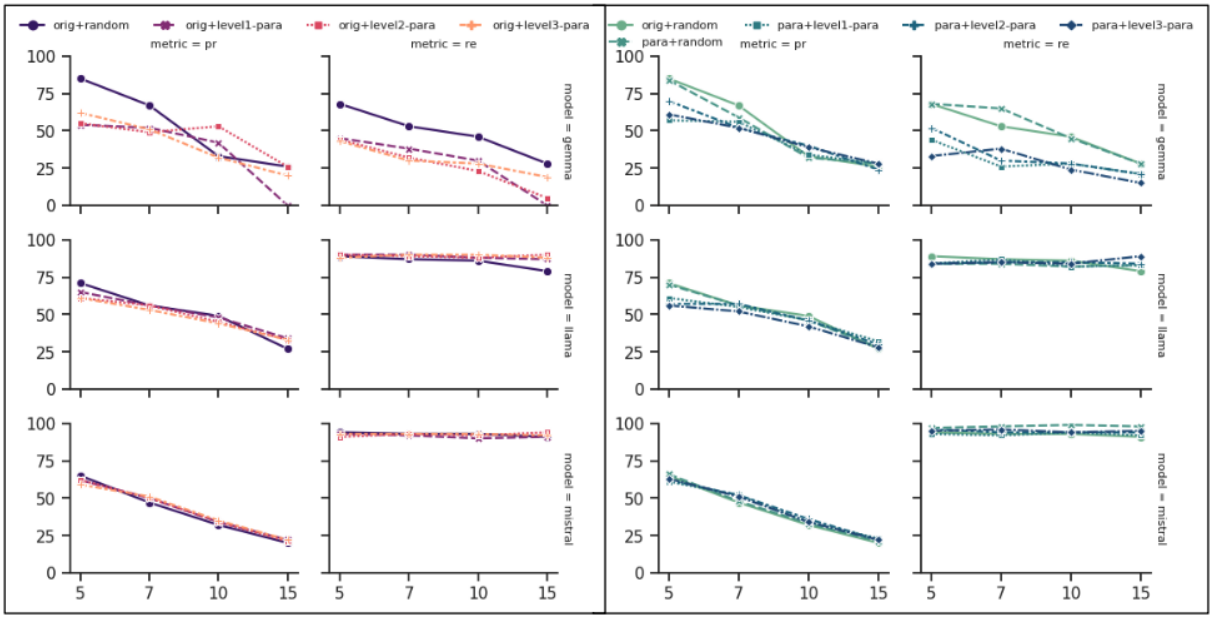

Figure 3: Model-wise precision and recall for experiments in Group 4 (left) and Group 5 (right).

performance for all models indicates that although LLMs contain facts about ionic liquids, they can't reliably utilize and reason with them for complex tasks.

Effect of paraphrasing

Comparing Groups 2 and 3 in Table 1, although paraphrasing the correct options reduces the F1 score across all difficulty levels for Llama and Mistral, paraphrasing the incorrect options in Group 4 has a higher diminishing effect on the model performance than Group 2, which is the opposite for Gemma. We hypothesize that this might be due to Gemma's reliance on linguistic cues for entailment compared to Llama and Mistral, where Gemma relies more on syntactic similarity than semantics.

Comparing Groups 3 and 5, paraphrasing the incorrect options reduces the F1 score across all difficulty levels for Llama and Mistral, which is the opposite for Gemma, except for experiment 15. Comparing Groups 4 and 5, paraphrasing the cor-

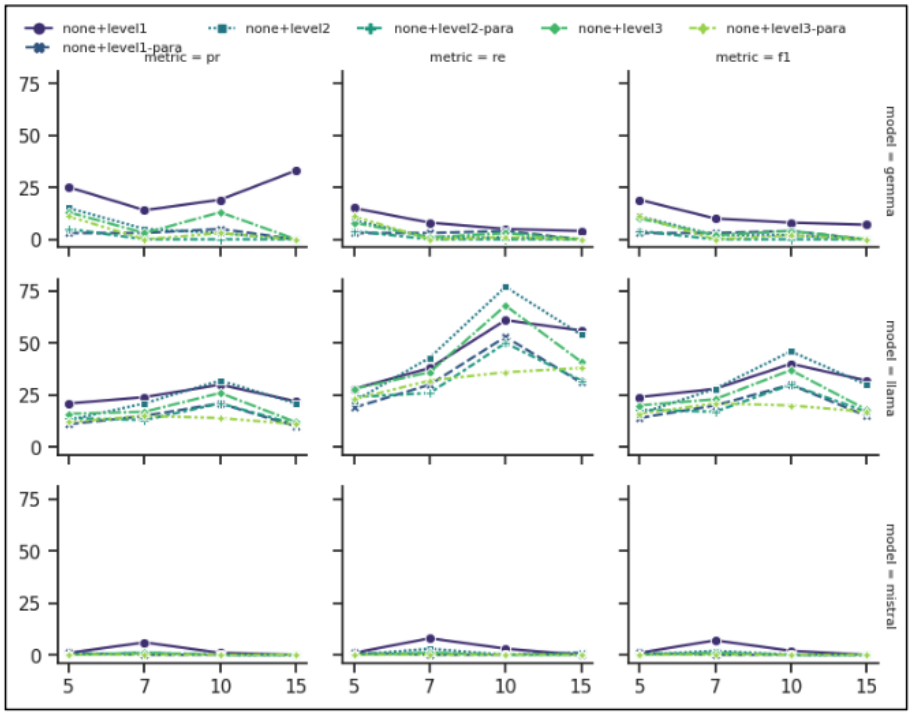

Figure 4: Model-wise precision, recall, and F1 for experiments in comparison suite 5.

rect options reduces the F1 score for Llama across all difficulty levels. On the contrary, the F1 score increases or remains the same for Gemma and Mistral, except for experiment 18. We hypothesize that since the correct and incorrect options share syntactic similarities, they get equally transformed while paraphrasing, causing their paraphrased versions to maintain syntactic similarity, which weaker reasoning models like Gemma exploit. We leave the testing out of this hypothesis as future work.

Overall, Llama performs best across all experiments, followed by Mistral and Gemma. Our results indicate that although LLMs possess knowledge of ionic liquids and carbon capture, their domain-specific reasoning capabilities are limited. The performance drop in Group 1 experiments is drastic for all models and sometimes near zero for Mistral and Gemma, which questions their reasoning capabilities.

5 Discussion

Our experiments indicate that smaller LLMs struggle to coherently reason within the domain-specific constraints and choose non-probable options in the entailment task. This is likely because LLMs are general-purpose and not geared to niche domains such as ILs. We propose that LLMs should be fine-tuned for CBE using curated datasets. Pre-training the models on domain-specific data, fine-tuning using PEFT (Mangrulkar et al., 2022) methods like LoRA (Hu et al., 2021), or in-context learning and efficient methods such as RAG (Lewis et al., 2020; Gao et al., 2024) should help impart the domain-specific knowledge and constraints, which requires collaborative advancements in the intersection of LLMs and CBE. Such domain-specific LLMs can scale IL research by assisting researchers in the bottlenecked areas of data analysis, experiment design, and property predictions. Furthermore, they can serve as educational guides to researchers willing to gain familiarity with the field. This work should be a valuable resource for researchers eager to evaluate LLMs for varied fields and collaboratively help attain the sustainability goals of the UN[5].

6 Conclusion

Global warming remains a pressing challenge, necessitating scalable and interdisciplinary solutions such as carbon capture. To address this need, we propose leveraging LLMs to support research on Ionic Liquids, a promising avenue for carbon capture. As a foundational step, we construct and publicly share an expert-curated dataset designed to evaluate LLMs' knowledge and reasoning capabilities within the specialized domain of Ionic Liquids. Our benchmarking of three open-weight

[5]https://sdgs.un.org/goals

LLMs—Llama, Gemma, and Mistral—reveals that while general-purpose models, particularly Llama, demonstrate a strong grasp of Ionic Liquid-related knowledge, they fall short in domain-specific reasoning tasks. Building on these findings, we outline potential pathways for LLMs to advance Ionic Liquid research, including their use as agents in simulations, reasoners for material discovery and design, and educational tools to help researchers familiarize themselves with the field. Moreover, optimizing LLMs for climate research not only advances carbon capture efforts but also offers a dual benefit by mitigating the models' own carbon footprint. This alignment between AI innovation and environmental goals supports the broader aim of achieving carbon neutrality by 2050.

Limitations

This study has some notable limitations. Firstly, we only evaluate three open-weight models with less than 10B parameters for their knowledge and reasoning ability with ILs. Although extraneous experiments with larger and open-API models indicate a similar trend, they are not quantified and non-generalizable. Secondly, our entailment test set is not an exhaustive resource for IL research. It contains limited facts and only tests reasoning capabilities through entailment. We need more diverse datasets that probe the reasoning capabilities of LLMs from multiple aspects. Thirdly, we do not experiment with fine-tuning the models on our dataset and measure their impact on reasoning, which we intend as future work. Also, our work is limited to two expert evaluators and might benefit from multiple experts. Despite these limitations, our research takes a foundational step in the interdisciplinary field of LLMs for ionic liquid research, which is very nascent.

Ethics Statement

We confirm that all conducted experiments are solely for academic purposes and adhere to ethical standards. The expert evaluators were appropriately compensated for their tasks, following all administrative and regulatory policies. The shared dataset strictly pertains to ionic liquids. It does not contain potentially explicit and sensitive content that might exhibit bias, be hurtful, or offend anyone.

References

Mahsa Aghaie, Nima Rezaei, and Sohrab Zendehboudi. 2018. A systematic review on co2 capture with ionic liquids: Current status and future prospects. *Renewable and sustainable energy reviews*, 96:502–525.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Jennifer L Anthony, Edward J Maginn, and Joan F Brennecke. 2002. Solubilities and thermodynamic properties of gases in the ionic liquid 1-n-butyl-3-methylimidazolium hexafluorophosphate. *The Journal of Physical Chemistry B*, 106(29):7315–7320.

John L Austin. 1961. Other minds.

Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):120.

Igor Baskin, Alon Epshtein, and Yair Ein-Eli. 2022. Benchmarking machine learning methods for modeling physical properties of ionic liquids. *Journal of Molecular Liquids*, 351:118616.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Lloyd F Bitzer. 2020. Aristotle's enthymeme revisited. In *Landmark Essays on Aristotelian Rhetoric*, pages 179–191. Routledge.

Lynnette A Blanchard, Zhiyong Gu, and Joan F Brennecke. 2001. High-pressure phase behavior of ionic liquid/co2 systems. *The Journal of Physical Chemistry B*, 105(12):2437–2444.

Lynnette A Blanchard, Dan Hancu, Eric J Beckman, and Joan F Brennecke. 1999. Green processing using ionic liquids and co2. *Nature*, 399(6731):28–29.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Markus J Buehler. 2023a. Generative pretrained autoregressive transformer graph neural network applied to the analysis and discovery of novel proteins. *Journal of Applied Physics*, 134(8).

Markus J Buehler. 2023b. Melm, a generative pretrained language modeling framework that solves forward and inverse mechanics problems. *Journal of the Mechanics and Physics of Solids*, 181:105454.

Lingdi Cao, Peng Zhu, Yongsheng Zhao, and Jihong Zhao. 2018. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids. *Journal of hazardous materials*, 352:17–26.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *Preprint*, arXiv:2501.17161.

Pratik Dhakal and Jindal K Shah. 2022. A generalized machine learning model for predicting ionic conductivity of ionic liquids. *Molecular Systems Design & Engineering*, 7(10):1344–1353.

Radoslav S Dimitrov. 2016. The paris agreement on climate change: Behind closed doors. *Global environmental politics*, 16(3):1–11.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,

Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Prateek Sharma, Fan Chen, and Lei Jiang. 2024. Llmcarbon: Modeling the end-to-end carbon footprint of large language models. *arXiv preprint arXiv:2309.14393*.

Haijun Feng, Pingan Zhang, Wen Qin, Weiming Wang, and Huijing Wang. 2022. Estimation of solubility of acid gases in ionic liquids using different machine learning methods. *Journal of Molecular Liquids*, 349:118413.

Constanza Fierro, Ruchira Dhar, Filippos Stamatiou, Nicolas Garneau, and Anders Søgaard. 2024. Defining knowledge: Bridging epistemology and large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16096–16111, Miami, Florida, USA. Association for Computational Linguistics.

Daan Frenkel and Berend Smit. 2023. *Understanding molecular simulation: from algorithms to applications*. Elsevier.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2024. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*.

Joel Guiot and Wolfgang Cramer. 2016. Climate change: The 2015 paris agreement thresholds

and mediterranean basin ecosystems. *Science*, 354(6311):465–468.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yiwen Hu and Markus J Buehler. 2022. End-to-end protein normal mode frequency predictions using language and graph models and application to sonification. *ACS nano*, 16(12):20656–20670.

Yiwen Hu and Markus J Buehler. 2023. Deep language models for interpretative and predictive materials science. *APL Machine Learning*, 1(1).

Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Pascale Husson-Borg, Vladimir Majer, and Margarida F Costa Gomes. 2003. Solubilities of oxygen and carbon dioxide in butyl methyl imidazolium tetrafluoroborate as a function of temperature and at pressures close to atmospheric pressure. *Journal of Chemical & Engineering Data*, 48(3):480–485.

Pavan Inguva, Vijesh J Bhute, Thomas NH Cheng, and Pierre J Walker. 2021. Introducing students to research codes: A short course on solving partial differential equations in python. *Education for Chemical Engineers*, 36:1–11.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 2023. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

Eesha Khare, Constancio Gonzalez-Obeso, David L Kaplan, and Markus J Buehler. 2022. Collagentransformer: end-to-end transformer model to predict thermal stability of collagen triple helices using an nlp approach. *ACS Biomaterials Science & Engineering*, 8(10):4301–4310.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024a. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. 2023. Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv preprint arXiv:2304.03271*.

Weng Marc Lim, Asanka Gunasekara, Jessica Leigh Pallant, Jason Ian Pallant, and Ekaterina Pechenkina. 2023. Generative ai and the future of education: Ragnarök or reformation? a paradoxical perspective from management educators. *The international journal of management education*, 21(2):100790.

Frank YC Liu, Bo Ni, and Markus J Buehler. 2022. Presto: Rapid protein mechanical strength prediction with an end-to-end deep learning model. *Extreme Mechanics Letters*, 55:101803.

Wei Lu, David L Kaplan, and Markus J Buehler. 2024. Generative modeling, design, and analysis of spider silk protein sequences for enhanced mechanical properties. *Advanced Functional Materials*, 34(11):2311324.

Wei Lu, Nic A Lee, and Markus J Buehler. 2023. Modeling and design of heterogeneous hierarchical bioinspired spider web structures using deep learning and additive manufacturing. *Proceedings of the National Academy of Sciences*, 120(31):e2305273120.

Rachel K Luu and Markus J Buehler. 2024. Bioinspiredllm: Conversational large language model for the mechanics of biological and bio-inspired materials. *Advanced Science*, 11(10):2306724.

Rachel K Luu, Marcin Wysokowski, and Markus J Buehler. 2023. Generative discovery of de novo chemical designs using diffusion modeling and transformer deep neural networks with application to deep eutectic solvents. *Applied Physics Letters*, 122(23).

Edward J Maginn. 2009. Molecular simulation of ionic liquids: current status and future opportunities. *Journal of Physics: Condensed Matter*, 21(37):373101.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334.

Kusuri Murakumo, Naruki Yoshikawa, Kentaro Rikimaru, Shogo Nakamura, Kairi Furui, Takamasa Suzuki, Hiroyuki Yamasaki, Yuki Nishigaya, Yuzo Takagi, and Masahito Ohue. 2023. Llm drug discovery challenge: A contest as a feasibility study on the utilization of large language models in medicinal chemistry. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A Smith, Hannaneh Hajishirzi, and Tom Hope. 2021. Scientific language models for biomedical knowledge base completion: an empirical study. *arXiv preprint arXiv:2106.09700*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Robert Nozick. 2016. Knowledge and scepticism. In *Readings in Formal Epistemology: Sourcebook*, pages 587–603. Springer.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun

Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kamil Paduszynski. 2016. In silico calculation of infinite dilution activity coefficients of molecular solutes in ionic liquids: critical review of current methods and new models based on three machine learning algorithms. *Journal of chemical information and modeling*, 56(8):1420–1437.

Saurabh Kumar Pandey, Harshit Budhiraja, Sougata Saha, and Monojit Choudhury. 2025. CULTURALLY YOURS: A reading assistant for cross-cultural content. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 208–216, Abu Dhabi, UAE. Association for Computational Linguistics.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Álvaro Pérez-Salado Kamps, Dirk Tuma, Jianzhong Xia, and Gerd Maurer. 2003. Solubility of co2 in the ionic liquid [bmim][pf6]. *Journal of Chemical & Engineering Data*, 48(3):746–749.

Mahinder Ramdin, Theo W de Loos, and Thijs JH Vlugt. 2012. State-of-the-art of co2 capture with ionic liquids. *Industrial & Engineering Chemistry Research*, 51(24):8149–8177.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Christopher J. Rhodes. 2016. The 2015 paris climate change conference: Cop21. *Science Progress*, 99(1):97–104. PMID: 27120818.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Anthony Robbins. 2016. How to understand the results of the climate change summit: Conference of parties21 (cop21) paris 2015. *Journal of public health policy*, 37(2):129–132.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can llms identify cross-cultural communication gaps? *Preprint*, arXiv:2502.09636.

Eloy S Sanz-Pérez, Christopher R Murdock, Stephanie A Didas, and Christopher W Jones. 2016. Direct capture of co2 from ambient air. *Chemical reviews*, 116(19):11840–11876.

Crispin Sartwell. 1992. Why knowledge is merely true belief. *The Journal of Philosophy*, 89(4):167–180.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Quintin R Sheridan, William F Schneider, and Edward J Maginn. 2018. Role of molecular modeling in the development of co2–reactive ionic liquids. *Chemical reviews*, 118(10):5242–5260.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. 2023. Probing the moral development of large language models through defining issues test. *Preprint*, arXiv:2309.13356.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Wilfred F. van Gunsteren, Xavier Daura, Niels Hansen, Alan E. Mark, Chris Oostenbrink, Sereina Riniker, and Lorna J. Smith. 2018. Validation of molecular simulation: An overview of issues. *Angewandte Chemie International Edition*, 57(4):884–902.

Wilfred F. van Gunsteren and Alan E. Mark. 1998. Validation of molecular dynamics simulation. *The Journal of Chemical Physics*, 108(15):6109–6116.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by actively validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Douglas N Walton. 1996. *Argument structure: A pragmatic theory*. University of Toronto Press Toronto.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

Shaofei Wang, Xueqin Li, Hong Wu, Zhizhang Tian, Qingping Xin, Guangwei He, Dongdong Peng, Silu Chen, Yan Yin, Zhongyi Jiang, et al. 2016. Advances in high permeability polymer-based membrane materials for co2 separations. *Energy & Environmental Science*, 9(6):1863–1890.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Andrew D White. 2023. The future of chemistry is language. *Nature Reviews Chemistry*, 7(7):457–458.

Timothy Williamson. 2005. Knowledge, context, and the agent's. *Contextualism in philosophy*, page 91.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Chi-Hua Yu, Wei Chen, Yu-Hsuan Chiang, Kai Guo, Zaira Martin Moldes, David L Kaplan, and Markus J Buehler. 2022a. End-to-end deep learning model to predict and design secondary structure content of structural proteins. *ACS biomaterials science & engineering*, 8(3):1156–1165.

Chi-Hua Yu, Eesha Khare, Om Prakash Narayan, Rachael Parker, David L Kaplan, and Markus J Buehler. 2022b. Colgen: An end-to-end deep learning model to predict thermal stability of de novo collagen sequences. *Journal of the mechanical behavior of biomedical materials*, 125:104921.

Linda Zagzebski. 2017. What is knowledge? *The Blackwell guide to epistemology*, pages 92–116.

Stefano E Zanco, José-Francisco Pérez-Calvo, Antonio Gasós, Beatrice Cordiano, Viola Becattini, and Marco Mazzotti. 2021. Postcombustion co2 capture: a comparative techno-economic assessment of three technologies using a solvent, an adsorbent, and a membrane. *ACS Engineering Au*, 1(1):50–72.

Shaojuan Zeng, Xiangping Zhang, Lu Bai, Xiaochun Zhang, Hui Wang, Jianji Wang, Di Bao, Mengdie Li, Xinyan Liu, and Suojiang Zhang. 2017. Ionic-liquid-based co2 capture systems: structure, interaction and process. *Chemical reviews*, 117(14):9625–9673.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Haochen Zhao, Xiangru Tang, Ziran Yang, Xiao Han, Xuanzhi Feng, Yueqing Fan, Senhao Cheng, Di Jin, Yilun Zhao, Arman Cohan, et al. 2024. Chemsafetybench: Benchmarking llm safety on chemistry domain. *arXiv preprint arXiv:2411.16736*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Appendix

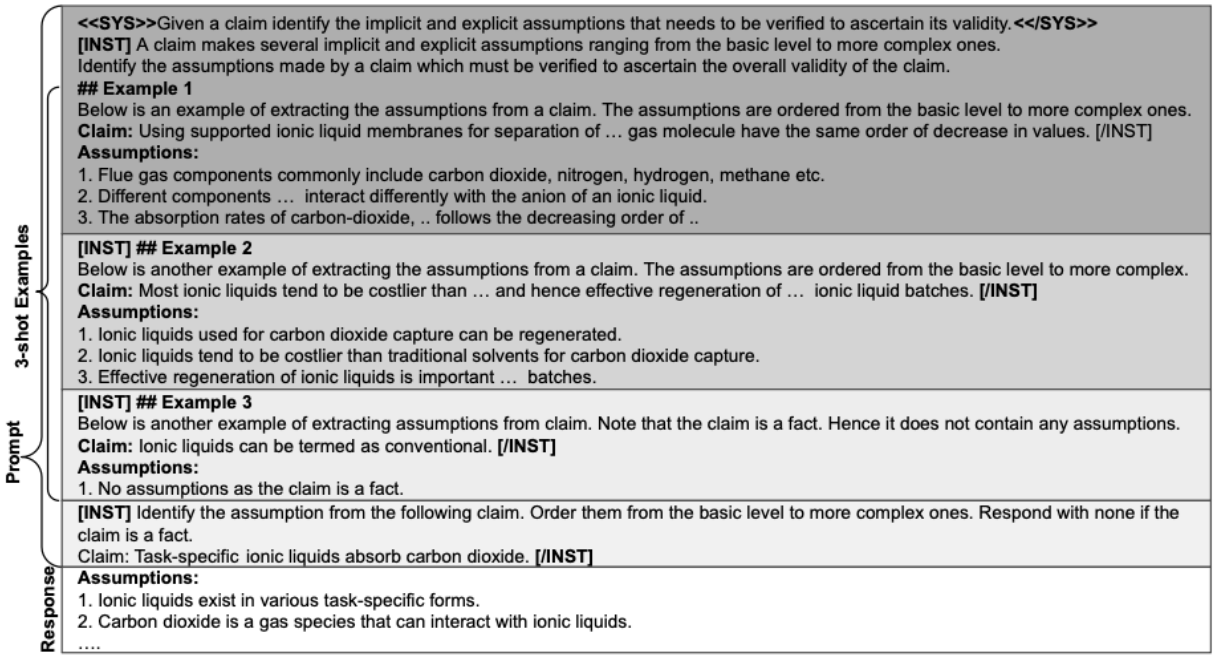

Figure 5: Mistral 3-shot prompt to automatically extract and generate the missing assumptions from claims.

You are an expert in Chemical Engineering with extensive knowledge of ionic liquids. Given the following claim and list of statements related to ionic liquids, indicate if each statement entails the claim. Also, on a scale of 1 to 5 (low to high), indicate your confidence in your response.

Claim: Higher energy consumption for regeneration of ionic liquids is a cost-intensive process.

Statements:
1. Adding water to conventional ionic liquids can reduce their viscosity.
2. Ionic liquid regeneration involves energy consumption only during rainy seasons.
3. Energy consumption for ionic liquids regeneration incurs cost only during rainy seasons.
4. Imidazolium-based ionic liquids can have alkyl groups attached to the imidazolium ring.
5. Energy consumption for ionic liquids regeneration incurs cost.
6. None of the above

Prompt

Format your response as a Python dictionary in the following format: {'<statement number>': {'entails': <True/False>, 'confidence': <score between 1 to 5>}}.
Do not generate anything extra.

{'1': {'entails': False, 'confidence': 5}, '2': {'entails': True, 'confidence': 5}, '3': {'entails': True, 'confidence': 5}, '4': {'entails': False, 'confidence': 5}, '5': {'entails': True, 'confidence': 5}, '6': {'entails': False, 'confidence': 5}}

Response

Figure 6: Sample prompt for the entailment task.

Correct the incorrect input as a Python dictionary in the following format: {{'<number between 1 to {n_statements}>': {{'entails': <Boolean True/False>, 'confidence': <Integer score between 1 to 5>}}}}
Please keep the following points in mind:
1. The size of the output dictionary must be {n_statements} and the dictionary elements must be numbered between 1 to {n_statements}.
2. If there are missing dictionary elements, add them as {{'<number>': {{'entails': False, 'confidence': 1}}}}
3. If the 'entails' key is missing in a dictionary element, add 'entails': False
4. If the 'confidence' key is missing in a dictionary element, add 'confidence': 1
5. If there are additional outside numbers 1 to {n_statements}, remove them.
6. Do not generate the code.

Prompt

Incorrect input: {'0': {'entails': False}, '2': {'entails': False, 'confidence': 5}, '3': {'entail': False, 'confidence': 5}, '4': {'entails': False, 'confidence': 5}, '5': {'entails': False, 'confidence': 5} 5: {'entails': True, 'confidence': 5},}

{'1': {'entails': False, 'confidence': 1}, '2': {'entails': False, 'confidence': 5}, '3': {'entails': False, 'confidence': 5}, '4': {'entails': False, 'confidence': 5}, '5': {'entails': False, 'confidence': 5}, 6: {'entails': True, 'confidence': 5}}

Response

Figure 7: Sample prompt for correcting the LLM response using Llama.