
NMR-CHALLENGE FOR LLMs: EVALUATING CHEMICAL REASONING IN HUMANS AND AI

A PREPRINT

Samiha Sharlin¹, Fariha Agbere¹, Kevin Ishimwe³, Zuzana Osifová^{2,4}, Ondřej Socha², Martin Dračínský², and Tyler R. Josephson^{1,3}

¹Department of Chemical, Biochemical, & Environmental Engineering,
University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250

²Institute of Organic Chemistry and Biochemistry,
Czech Academy of Science, 160 00 Prague, Czech Republic

³Department of Computer Science & Electrical Engineering,
University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250

⁴Department of Chemical and Biological Physics, Weizmann Institute of Science, Rehovot, Israel

Abstract

Nuclear Magnetic Resonance (NMR) structure determination is an important problem in education, industry, and research. Solving NMR spectra requires expert knowledge, critical thinking, and careful evaluation of multiple features of spectral data. This study explores the capabilities of large language models (LLMs) for solving NMR spectral tasks. We selected 115 problems from NMR-Challenge.com, which has been used by students practicing NMR structure elucidation, collecting > 1 million human responses, and developed a plain text problem format for evaluating LLM reasoning in this domain. We evaluated 7 LLMs (GPT-4o, GPT-4o-mini, o1, o1-mini, o3-mini, Claude-3.5 Sonnet, and Gemini-2.0-Flash), comparing 5 prompts to spur chain-of-thought reasoning in different ways, especially comparing the influence of providing background NMR chemistry knowledge, reasoning strategy, or both. Newer models trained to emphasize reasoning performed better, and increasing reasoning effort led to modest improvements, but prompting and varying temperature didn't have an effect. We also evaluated undergraduate organic chemistry students in a controlled setting, and analyzed answer submission statistics from global submissions to NMR-Challenge.com, to characterize human performance on these problems. The top-performing students surpassed smaller models like GPT-4o by 24%, 33%, and 29% on the Easy, Moderate, and Hard sets. However, reasoning models like o1 exceeded student performance by 13%, 14%, and 19%, respectively. Patterns in mistakes made by humans and LLMs reveal that errors made by LLMs are similar to those typically made by humans, for instance, incorrect positioning of substituents on benzene and incorrect orientation of carboxyl groups in esters. However, LLMs still "think" differently from humans, in some cases, providing answers which no human submitted via the website. This work also illustrates how NMR spectral problems can be used to benchmark LLMs on reasoning-heavy tasks in chemistry, though for this particular set of problems, current LLMs already exceed undergraduate student performance.

1 Introduction

AI-augmented systems have grown significantly with the increasing capabilities of language models. The improvement results from two key factors – scaling model parameters or scaling inference times. The research landscape surrounding foundational language models is rapidly evolving. While larger models generally retain a performance edge, recent work with rstar-math [1], DeepSeek R1 [2], and others [3, 4] shows that smaller models can rival and occasionally exceed the performance of much larger ones. As more powerful models become available at lower costs, especially open-source ones, we can expect yet more increase in their

applications. The scientific research community will also leverage these advancements in AI-assisted systems across the fields of chemistry, biology, and materials science.

Large Reasoning Models (LRMs) are language models that “reason” using more tokens to analyze and explore different aspects of a problem in depth [5]. The early demonstration of this approach is seen in the basic Chain of Thought (CoT) prompting technique [6, 7]. Since then, effective strategies have been outlined in the literature, including reinforcement learning [2], supervised fine-tuning [3], and preference modeling [1]. These systems are designed to think deliberately [8] for problem-solving — reflection, backtracking, and trial-and-error — and some can even outperform humans [9].

Most studies on the reasoning capabilities of LLMs focus on their abilities to solve problems in math or planning. For instance, when employing in-context learning for mathematical problems like division, we present the model with several input-output examples, and the observed improvements are attributed to the model’s inductive reasoning abilities, but the problem itself more closely aligns with deductive reasoning [10]. This overlap of prompting technique and problem type makes it difficult to identify areas where LLMs need improvement in their reasoning abilities. This work provides benchmark data for complex, non-linear problem-solving in organic chemistry, while also addressing a practical challenge: identifying chemical structures that can explain observations from Nuclear Magnetic Resonance (NMR) spectroscopy experiments.

NMR uses oscillating magnetic fields to analyze molecule structures from their chemical shifts [11, 12]. Atoms resonate at different frequencies based on their local unique environments, and the resulting NMR spectrum provides insights into interactions among specific atoms within a molecule, ultimately deducing the molecule’s whole structure. NMR spectroscopy is an example of abductive reasoning used in scientific problem-solving. As this process involves using observations (NMR spectrum) to infer the most plausible explanation for a phenomenon (molecule identity), it is a form of “inference to the best explanation” where the goal is to identify the hypothesis that most likely explains the observed data [13]. Solving NMR spectral tasks “increases students’ critical thinking abilities because they have to evaluate multiple aspects at the same time, such as the number, intensity, and shape of signals, chemical shifts, and J-coupling values” [14], making these tasks a mainstay in undergraduate training in organic chemistry.

Modern tools for NMR structure determination most commonly include methods based on deep neural networks [15–23] and Transformers [24–28]. These approaches frame the problem as a supervised learning task, with spectra as features and molecular assignments as labels. After the model has been trained, inferences about new spectra can be obtained as model predictions. However, such methods typically rely on black-box architectures and hence do not offer explanations for their results. They also do not handle corner cases and instances not encountered during training well and require curated, large-scale labelled datasets, which are expensive to assemble for training. Automated tools that use formal logic have also been explored to identify molecules from spectral data [29–32]; however, they demand considerable effort from human experts to define the rules of the system accurately.

We recognize solving an NMR spectral problem as fundamentally a multi-step reasoning task [33], and aim to explore the reasoning capabilities of LLMs. Previous studies have focused on predicting observable signals from SMILES strings or molecular structures [34], or on extracting the NMR data itself [35], while this work highlights the reasoning capabilities of LLMs by providing the necessary data for predicting the molecule. The concept was developed during our 2024 LLM Hackathon for Applications in Materials and Chemistry project, where we manually curated NMR datasets to test our approach [36]. We found GPT-4 to be successful in some cases, however the scratchpad showed interesting patterns of (apparent) reasoning in its outputs. Building on this preliminary work, this paper presents an improved benchmark dataset and a systematic exploration of prompting strategies, temperatures, reasoning efforts, and models in identifying molecules from NMR spectra.

This work focuses on a single task in chemistry to evaluate the performance of LLMs, in contrast to other benchmarks that combine various sub-fields [37–39] or tasks together [34, 40]. Similar to the recently proposed ChemIQ benchmark [41], our prompting approach is open-ended rather than employing multiple-choice formats. However, we specifically instruct the LLM to identify chemicals from spectral data by their names and then convert the answer to SMILES ourselves during the grading process. In contrast, ChemIQ requires the generation of SMILES strings from NMR data, which introduces an additional reasoning component. Although an LLM may accurately identify the name of a molecule, it could still produce an incorrect SMILES string, making it difficult to identify its reasoning errors. Converting a molecule name to a SMILES string representation involves additional reasoning, which could be explored in future research.

2 Method

2.1 Data Processing

We used the NMR spectral data that was used to create [NMR-Challenge.com](https://www.nmr-challenge.com), an interactive website featuring exercises categorized into Easy, Moderate, and Hard levels [14, 42]. For each problem, the website provides images of 1D spectra (^1H , ^{13}C , ^{19}F) at the basic level and additionally 2D spectra (COSY, HSQC, HMBC) at the advanced level, along with the corresponding chemical formula. We selected 53 Easy, 38 Moderate, and 24 Hard problems, and used MNOVA [43] multiplet analysis to extract and summarize numerical data about peaks. The multiplet analysis compresses the raw spectra data, making the input text more tractable for the LLMs, which was not done in the Hackathon version of our project [36]. This also prevents direct feeding of website data to the LLMs, reducing the risk of triggering memorization. For humans with access to the visual spectrum, peak-picking may be useful, but not likely as critical. We organized these data in JSON format where each entry contains molecular formula (if applicable), a unique problem ID, and an array of information about each peak: label with multiplicity, chemical shift, integration range, hydrogen count, integral value, multiplicity class, and J-coupling values as shown in Figure 1.

In the runs in which we did not provide the molecular formula, we normalized the peak integrals to give the smallest peak an area of 1, to prevent leakage of information about the number of hydrogens associated with each peak. For example, $\text{C}_4\text{H}_8\text{O}_2$ has peaks A, B, and C with integrals of 3.0, 2.05, and 3.02, respectively. Normalization yielded values of approximately 1.46 for A ($3.0/2.05$), 1 for B, and approximately 1.47 for C ($3.02/2.05$). When analyzing the data alongside the molecular formula, the normalized integrals correspond to the number of hydrogen atoms present. However, without the formula, the normalized values only provide relative ratios without specific hydrogen counts. Additionally, we conducted another set of experiments that included unnormalized ^1H -NMR and ^{13}C -NMR data, along with the molecular formula in the prompts, to replicate the conditions for solving these problems as you would on the website.

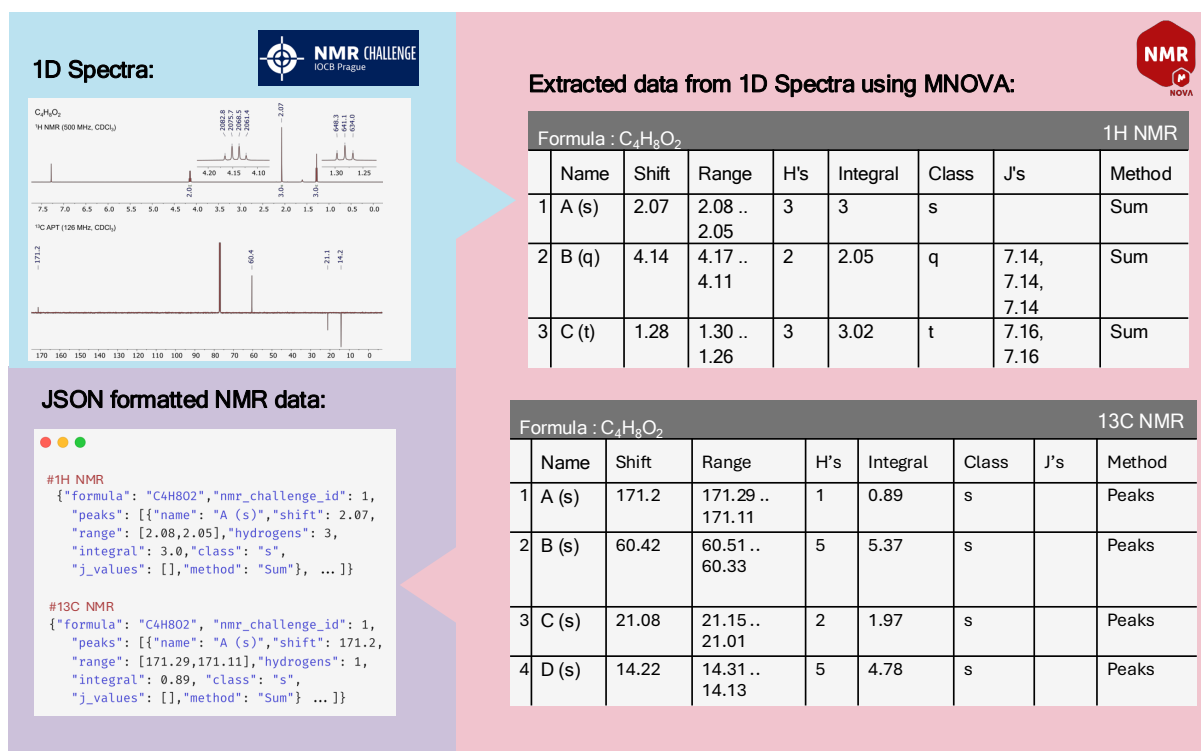


Figure 1: **Dataset processing for input into the LLMs.** The table above (in red box) illustrates the raw data extracted from MNOVA software for $\text{C}_4\text{H}_8\text{O}_2$ from the NMR challenge website. The information is then organized in JSON format, with selected information sent to the LLMs/LRMs, depending on the testing condition.

2.2 Experiment Details

We evaluated seven large language models (LLMs) on the NMR benchmark, including GPT-4o, GPT-4o mini, o1, o1-mini, o3-mini from OpenAI [44], Claude-3.5 Sonnet from Anthropic [45], and Gemini-2.0-Flash from Google DeepMind [46]. We varied the model temperature when allowed to assess the effect of deterministic versus stochastic token sampling on performance. Aside from OpenAI’s GPT-4o and GPT-4o mini, other models are described as reasoning models that employ some form of scaled test-time computation [47]. However, since the models are not open source, we do not know exactly how reasoning works for the majority of them.

Each model was tested with five prompt styles to establish baseline performance and evaluated with and without chemical formulae in the spectral information. The prompts were designed to assess how varying levels of guidance influence the reasoning process and overall model performance. In particular, we considered the domain of NMR structure elucidation as requiring domain-specific knowledge as well as domain-specific problem-solving strategy. For instance, the characteristic shifts, peak intensities, and splitting patterns associated with important functional groups would be learned knowledge. In contrast, “iteratively hypothesize fragments before synthesizing a whole structure (and repeat until the solution is consistent)” would be logical strategy.

P1. Base prompt: This prompt provides minimal user input, allowing for non-guided generation. It was used as a control condition to observe the inherent reasoning abilities of LLMs for solving NMR spectra problems.

P2. Standard CoT prompt: This prompt uses the standard CoT prompt introduced by Wei et al. [7] which encourages step-by-step reasoning to solve complex problems.

P3. CoT prompt with logical tips: Building on the standard CoT, this prompt variation includes domain-specific logical tips. It encourages the model to identify fragments, rank hypotheses, verify stoichiometry, and propose plausible molecular fragments, while ensuring consistency in connectivity.

P4. CoT prompt with knowledge: A human expert on NMR data interpretation provided approximately 1 page of guidance on domain-specific knowledge involved in solving NMR spectra, for instance, the characteristic shifts, peak intensities, and splitting patterns associated with important functional groups.

P5. CoT prompt with knowledge and logical tips: This is the most comprehensive of our modalities and combines the CoT prompt with logical tips and human expert insights.

2.3 Evaluation Metrics

To evaluate how well large language models (LLMs) predict molecular structures from NMR spectral data, we needed a reliable method to compare their output to the correct answer, given that a molecule can have many different names, and we could not anticipate which name the LLM would generate. We retrieved Simplified Molecular Input Line Entry System, i.e., SMILES [48] representations of chemical names for the correct answer and the LLM’s response using a Python library called CIRpy. CIRpy converts any chemical identifier to another chemical representation developed by the CADD Group at the NCI/NIH [49].

We graded the prediction of the models using multiple criteria: compound name matching, SMILES string match, and Tanimoto similarity using `rdFingerprintGenerator` in RDKit [50]. A prediction received a score of 1 if any of the following conditions were met: (1) the predicted compound name matched the ground truth name; (2) the predicted and ground truth SMILES matched exactly; (3) the Tanimoto similarity exceeded similarity exceeded 0.99, confirming that the molecules were truly identical. All models produced responses with varying adherence to the template provided in the context. Consequently, different functions were written to extract the name of the molecule, with some answers requiring manual extraction. We did not penalize models in assessing their chemical reasoning abilities if they failed to follow our specified output template. However, we observed that models that scored higher in NMR structure determination also tended to follow instructions correctly more often. We then verified instances where CIRpy failed to generate SMILES strings against PubChem to distinguish between cases where the LLM predicted chemically invalid compound names versus actual CIRpy limitations.

3 Results and Discussion

We organize our findings into three sections. First, we analyze the performance of all models under varying temperatures, prompting styles, and reasoning levels in solving NMR spectra problems. Next, we show

results of human performance on the same problems. Finally, we present additional experiment results where we align LLM conditions more closely with those used in human testing.

3.1 Tuning Prompts, Temperature, and Reasoning Depths of LLMs and LRMs

3.1.1 Effect of Prompting Style

Easy problems consistently scored the highest in both experiment settings – with and without formula included. Moderate and Hard tasks have similar accuracy when formula is included. However, there is a significant drop in performance for Hard tasks, which fell well below Moderate tasks when no formula is included.

We hypothesized that incorporating domain knowledge and logical tips into our prompts would lead to systematic improvements in model performance. However, in practice, neither LLMs nor the LRMs showed any benefit from these advanced prompt templates. In fact, we sometimes observed a decline in scores for example, Gemini 2.0 Flash scored dropped from P4 to P5 on Easy problems (with formula), and Claude Sonnet 3.5 score fell from P1 to P2 on Moderate problems (with formula). The changes are minimal for most models, with some modest improvements; however, no consistent upward or downward trend is observed, contrary to our expectations. OpenAI’s guide to using reasoning models dictates that these models perform better on tasks with high-level guidance alone, and that CoT-type prompts may impede performance, which seems to be observed in some of the cases here [51].

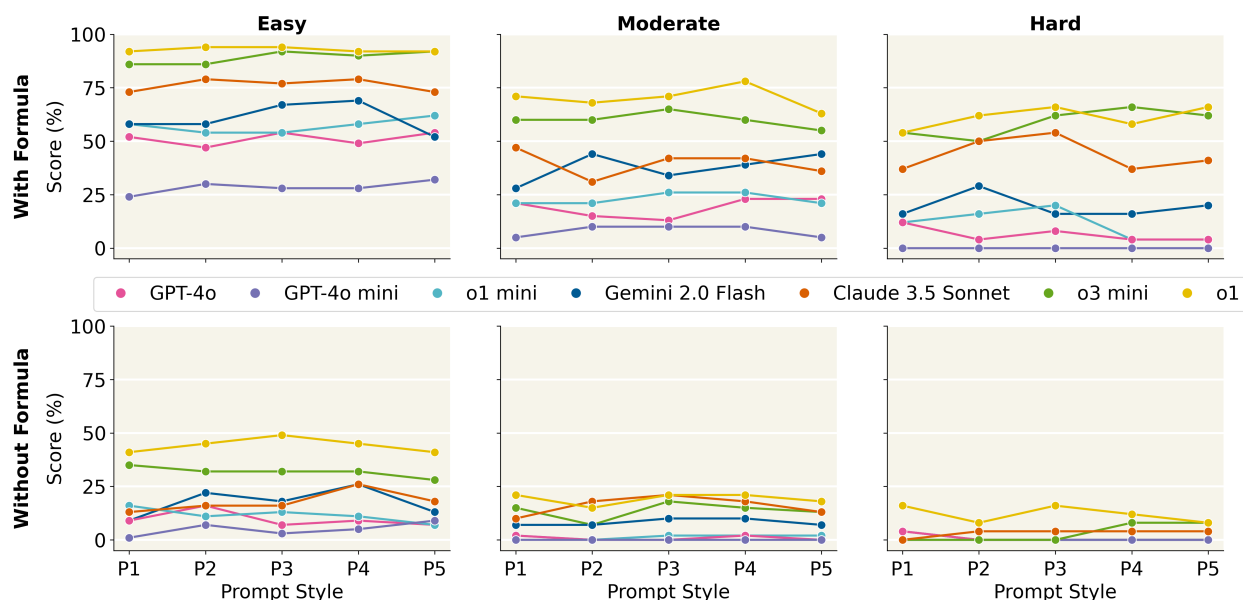


Figure 2: **Effect of prompting style across NMR problem difficulties.** We experimented with five prompting styles. Here, P1 represents the base prompt, P2 is the standard CoT prompt, P3 is the CoT prompt with logic, P4 is CoT with knowledge, and P5 is CoT with both logic and knowledge. All model data presented here is for temperature T1 and the default reasoning effort (medium) is used for OpenAI models.

3.1.2 Effect of Temperature

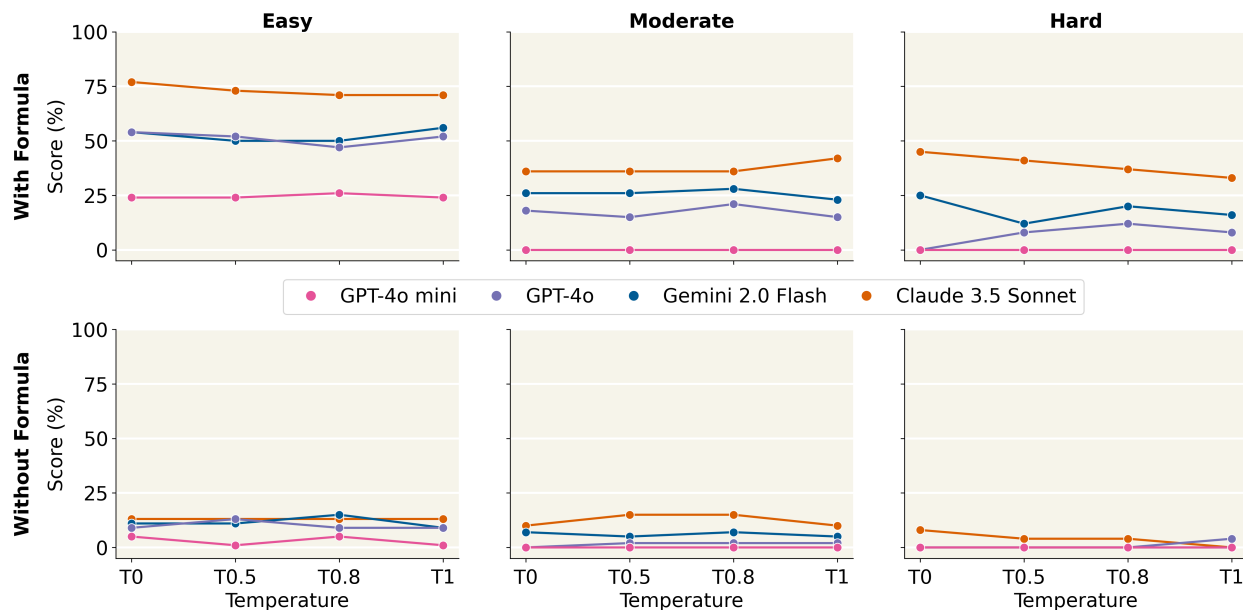


Figure 3: **Effect of temperature across NMR problem difficulties for base prompt (P1).** Four temperature levels were experimented: T0, T0.5, T0.8, and T1.

We tested the effect of temperature on four models: GPT-4o, GPT-4o mini, Claude 3.5 Sonnet, and Gemini 2.0 Flash. The other OpenAI models (o1 mini, o3 mini and o1) have their temperature fixed at 1 as part of model design.

In all panels of Fig. 3, we see no consistent effect of temperature across model type or difficulty level. No model significantly outperforms or underperforms at any given temperature. However, the average performance consistently follows the order from Easy to Moderate to Hard across all models. Notably, including the formula significantly improves accuracy: the “with formula” conditions (top row) consistently achieve much higher scores than their “without formula” counterparts (bottom row). These findings are also consistent across other prompt types (see Figure 9 in SI). Another study [52] also noted that if a model performs reasonably well at problem-solving, temperature values between 0 and 1 do not significantly impact the results.

At its core, temperature impacts the softmax function used to generate output probabilities that create responses token by token. Lower temperatures sharpen the distribution, making the model more deterministic. Higher temperatures lead to more randomness and a uniform output distribution. This randomness can foster creativity for certain tasks, but it can also result in inaccuracies due to unpredictability. Additionally, we cannot pinpoint the effects of temperature across models, as they use different search algorithms, which may or may not be influenced by temperature. For example, a model using beam search behaves very differently from one using greedy search [53]. We anticipated that a higher temperature would encourage creative hypothesis formation and hinder precise reasoning, potentially leading to counterbalancing effects. However, we did not observe significant effects in any of the cases.

3.1.3 Effect of Reasoning Effort

Reasoning models provide both a response and a CoT, allowing them to learn not just “what” to answer but also “how” to answer. This improvement in performance, as they generate more tokens, is called inference-time scaling. This is achieved in various ways, such as selecting the best answer after sampling many generations (output-focused) or modifying the proposal distribution (input-focused) by assigning more weight to reasoning tokens.

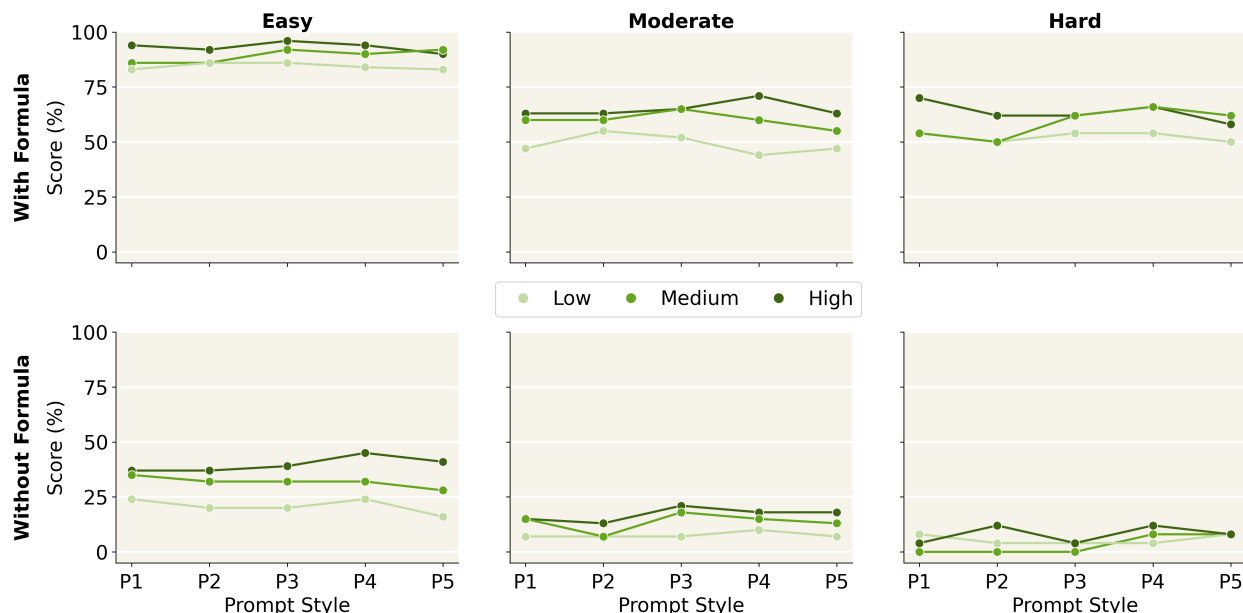


Figure 4: **Effect of reasoning effort across NMR problem difficulties and prompting styles.** OpenAI’s o3 mini model was tested in this study.

The recent OpenAI reasoning models provide fine-tuning through reasoning effort tuning as a hyperparameter. The `reasoning_effort` parameter has three settings: low, medium, and high. Low prioritizes speed and uses fewer tokens; high enables detailed reasoning. The default setting is medium, which balances speed and accuracy. We tested the o3-mini model with varying levels of “reasoning effort” to observe how it reasons and performs across the NMR problem categories. Across all prompt styles, higher reasoning effort leads to better scores (see Fig. 4) as expected. In the “with formula” condition, the gap between low, medium, and high effort is wider for Moderate and Hard problems. Conversely, the “without formula” condition shows a narrower gap with difficulty, and overall lower scores. We can apply high reasoning effort to difficult problems (Moderate/Hard, “with formula”) or to problems with unknown domains (Easy, “without formula”).

3.1.4 Overall Model Performance

When we aggregate results across various prompting styles, temperatures, reasoning efforts, and difficulty levels, we observe that reasoning models outperform regular LLMs in both experimental settings, with and without the molecular formula. The scores with the molecular formula are consistently three to six times higher than those without. The models analyzed in this study are closed-source, so we cannot assess their performance based on size or complexity. However, there is a correlation between higher output costs and improved performance (see Table 1).

Model	Release Date (MM/YY)	Output cost per 1M tokens (\$)	Score, with formula (%)
GPT-4o mini	07/24	0.6	12
GPT-4o	05/24	10	26
o1 mini	09/24	4.4	30
Gemini 2.0 Flash	02/25	0.4	40
Claude Sonnet 3.5	06/24	15	55
o3 mini	01/25	4.4	68
o1	12/24	60	75

Table 1: Comparison of models with their output costs and performance with formula. Prices were retrieved from their official documentation on 05/2025.

Among the models, Gemini 2.0 Flash was the least expensive on a per-token basis (including over lower-performing LLMs GPT-4o and GPT-4o-mini), while o1 was the most expensive.

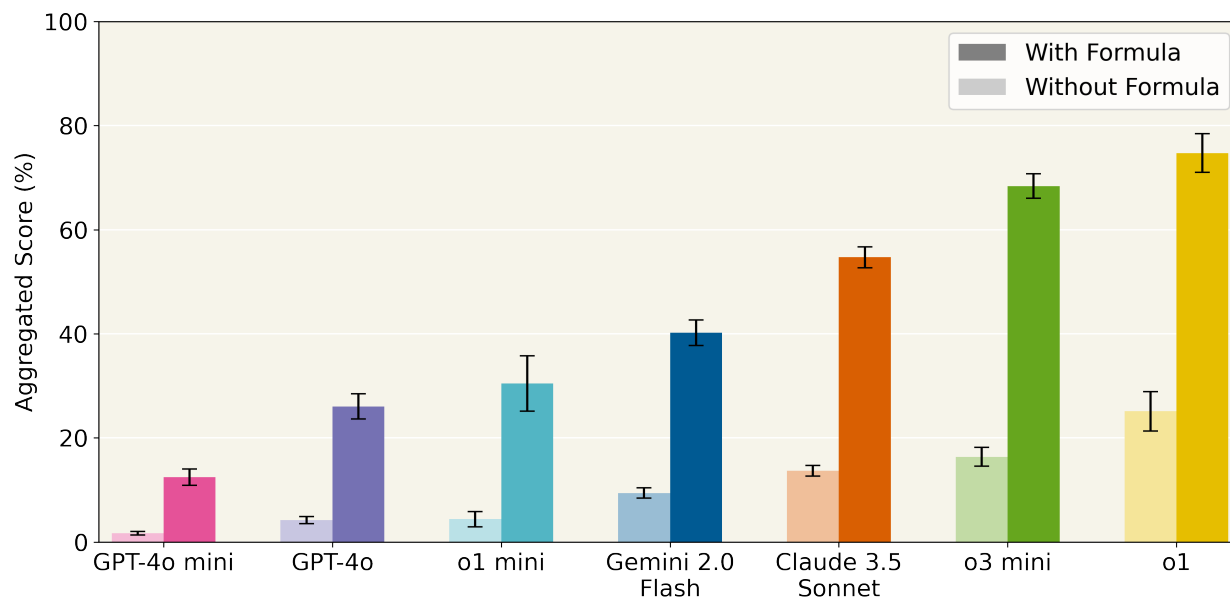


Figure 5: **Aggregated model performance.** From the right, model o1 leads with about 75% accuracy when formula is included, compared to only 25% without it. Next is o3-mini, scoring approximately 68% with formula and about 16% without. Claude 3.5 Sonnet follows with 55% accuracy when using formula and 14% without. Gemini 2.0 Flash scores around 40% with formula, but only 9% without it. Both o1-mini and GPT-4o perform in the 30s to mid-20s with formula, respectively, but their scores drop to around 4% without it. Finally, GPT-4o-mini has the lowest performance, achieving about 12% with the formula and only about 2% without.

3.2 Assessing Student Performance

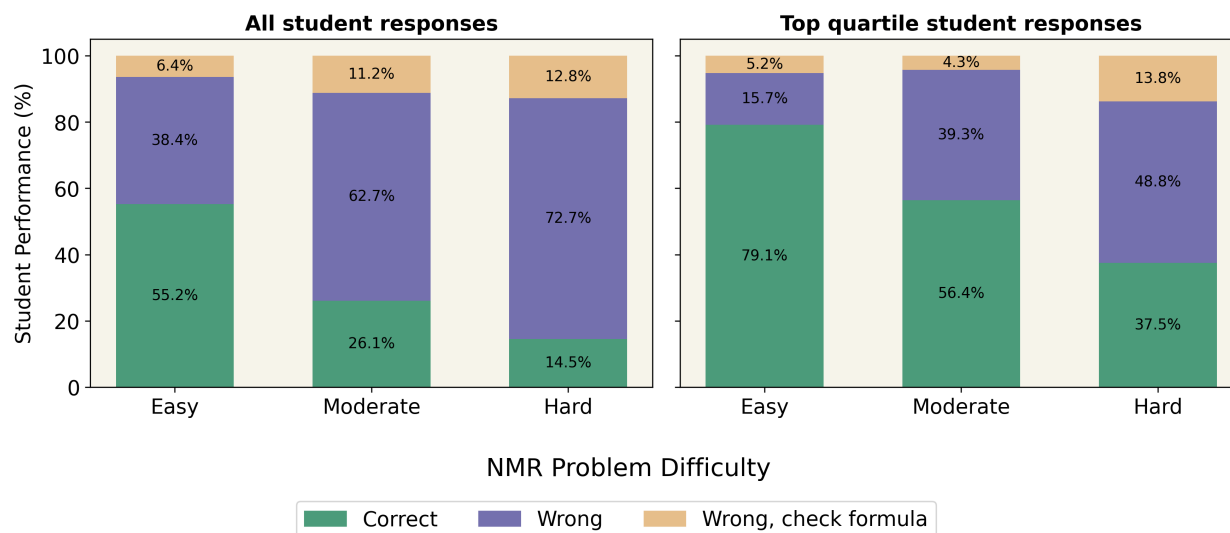


Figure 6: **Student performances across NMR problems.** Stacked bars show the percentage of correct (green), wrong (purple), and wrong with an incorrect molecular formula (beige) responses for all students (right) and the top-quartile students (left).

The study involved 140 undergraduate students enrolled in the Organic Chemistry Laboratory course at the University of Maryland, Baltimore County (UMBC). Before participating in this research, the students

had covered the topic of NMR spectroscopy and had approximately two months of experience with it. Each student was assigned six problems categorized as Easy, three as Moderate, and two as Hard, all delivered through the interactive testing environment on the [NMR-Challenge.com](https://www.nmr-challenge.com) website [14, 42]. Additionally, they were given six optional problems to solve, with two from each category. While they could take as long as necessary, they were encouraged to spend around 15 minutes on each problem.

NMR-Challenge.com provides molecular formulas, as well as images of both ^{13}C -NMR and ^1H -NMR spectra with annotated peak information, and tasks students with drawing a chemical structure in the app. Submitted answers are graded immediately, with additional feedback provided if the submitted answer had the wrong formula. To identify the top quartile, we grouped participants by “Participant ID” and counted the number of NMR problems each attempted and the number they answered correctly. We then calculated an individual accuracy rate by dividing correct answers by total attempts. We set the 75th percentile of these accuracy rates as the cut-off for the top quartile. 39 of the 140 participants met this threshold and were labeled top quartile, shown in the plot on the right. Section 5.2 in SI shows the distribution of problems attempted by students across NMR problem difficulty levels.

3.3 Comparing Students vs. LLMs and LRMs

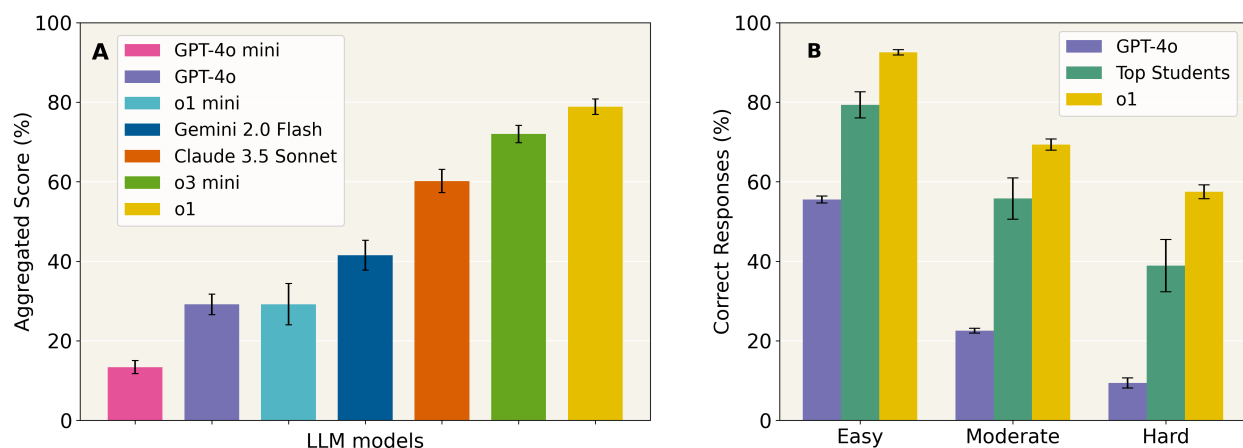


Figure 7: **Comparison of human vs. model accuracy across NMR difficulty.** The bar plot on the left (A) shows the performance of LLM models under conditions matching the human experiment conditions. Bar plots on the left (B) show the mean percentage of correct responses for top quartile human participants (in green), top reasoning model (o1, in yellow) and top non-reasoning model (GPT-4o in purple), with error bars indicating the standard error of the mean (SEM). For human data, we first computed each top quartile participant’s percentage of correct responses separately for Easy, Moderate, and Hard problems, i.e., (number correct ÷ number attempted × 100), and then averaged those percentages across the participants; the SEM reflects variability across participants. For the LLM, percentage correct was aggregated for three reasoning levels and five prompt types; the SEM reflects variability across experimental settings.

Since [NMR-Challenge.com](https://www.nmr-challenge.com) [14, 42] provides molecular formulas as well as images of both ^{13}C -NMR and ^1H -NMR spectra, we conducted an additional experiment with the LLMs. The main difference from our previous experimental runs was the inclusion of ^{13}C NMR data alongside un-normalized ^1H NMR data (formatted in JSON as illustrated in Figure 1), so that human participants and LLMs received the same information. That said, the human participants had access to visual representations of the spectra, while the LLMs were provided with a text input after multiplet peak analysis. This text-based input would have been challenging for human participants, in any case. While researchers have explored multimodal approaches and visual language models for interpreting NMR spectra [23, 54–56], our work emphasizes *reasoning* capabilities, which are better evaluated when decoupled from perception-related issues associated with figure interpretation. The aggregated model performance across temperatures, prompts, and reasoning levels with molecular formula included is shown in the left panel of Figure 7A.

The performance trends for the models are consistent with our previous runs, with o1 leading with an accuracy of 79%. o3-mini follows at 72%, then Claude 3.5 Sonnet at 60%. Gemini 2.0 Flash scores around 42%, similar to its performance in the previous run under similar conditions. o1-mini and GPT-4o score about

29%, while GPT-4o mini has the lowest performance at about 13%. Overall, performance improved compared to previous runs, indicating the benefit of incorporating ^{13}C NMR data (and illustrating that these models are able to effectively make use of ^{13}C NMR data). In the right panel of Figure 7B, we categorically compare the performance of top quartile students with the leading reasoning and non-reasoning language models. The results show that students outperform GPT-4o (LLM); however, o1 (LRM) outperforms students across all categories of problems.

These findings do not detract from NMR's special position in chemical education, and we recommend that instructors continue teaching and assigning NMR spectral tasks. NMR spectral tasks have long been a mainstay of chemical education, specifically because they cultivate critical thinking alongside understanding of molecular structure and characterization. After all arithmetic has been essential for math education, both before and after calculators could perform this task more easily and accurately than humans. The presence of new AI tools for solving chemistry problems doesn't necessarily change what problems are best for building up human skills.

3.4 Error Pattern Analysis of LLMs and Humans

Both humans and LLMs must generate and navigate an implicit “space” of possible structural hypotheses to evaluate and rank. We presume that when a human or an LLM produces a formula-consistent yet inaccurate answer, this reveals characteristics of the hypothesis spaces generated and navigated by humans and LLMs. If an LLM's answer is provided in a direct response (no chain of thought), we could enumerate possible answers, and measure log probabilities of the tokens to decipher the model's relative ranking of hypotheses. However, the scratchpad renders this approach invalid, and we also don't have access to what the humans were thinking; so we instead analyze the distribution of right and wrong answers as indicative of this distribution of hypotheses.

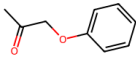
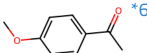
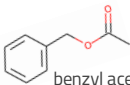
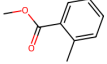

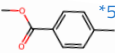
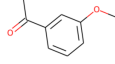

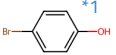
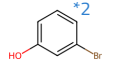

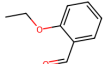
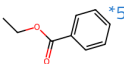
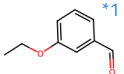
True Answer	GPT-4o	o1	NMR-Challenge
 Phenoxyl-2-propanone	 4'-Methoxyacetophenone *6		 benzyl acetate
 Methyl o-methylbenzoate	 Methyl benzoate *Incorrect formula	 Methyl 4-methylbenzoate *5	 2-Methoxyacetophenone
 2-bromophenol	 4-bromophenol *1	 3-bromophenol *2	 4-bromophenol
 2-ethoxy benzaldehyde	 ethyl benzoate *5	 3-ethoxybenzaldehyde *1	

Figure 8: **Analyzing patterns of errors.** The numbers in blue asterisk (in GPT-4o and o1 answers) indicate its rank among the most frequently submitted incorrect structures by global human participants on NMR-Challenge.com. Structures in the last column represent the most common incorrect submissions from human participants, all of whom are students, not experts in NMR data interpretation. For a complete list, refer to 5.3. A previous study [14] identified common mistakes made by humans, which include the recognition of isomeric esters (top 2 rows) and the identification of substitution positions in disubstituted benzenes (bottom 2 rows).

The most common mistakes made by language models in comparison to human participants across all 115 benchmark problems are listed in SI section 5.3, with a subset illustrated in Figure 8. The human participant data for comparison is from global submissions to NMR-Challenge.com, since it has a large database (of over 1 million submissions) and is more statistically significant. In contrast, the study involving UMBC students in Section 3.2, while gathering many fewer responses, provides a controlled environment for assessing accuracy, since website submissions allow humans multiple tries. We extracted the most common mistakes made from language models by combining the answers from all experiments conducted with the o1 and GPT-4o models in section 3.3. For o1 model, each problem has 5 (prompt variations) \times 3 (reasoning effort) possible answers, and GPT-4o model has 5 (prompt variations) \times 4 (temperature variations) possible answers.

In our opinion, one of the most fundamental reasoning aspects of these problems (when the formula is provided) is ensuring that the hypothesized molecule has the correct formula. By simply checking the formula (and not giving up!), one can be certain to prevent wrong answers of this kind. Yet for GPT-4o, this was the most common failure mode: 60 out of 115 problems had molecules with an incorrect formula as the most popular wrong answer. This revealed an astounding *inability* of GPT-4o to perform in-context reasoning about the most basic aspect of its answer, the formula. In comparison, the o1 model showed a markedly lower incidence of incorrect molecular formulas (7 out of 115) in its most common erroneous output. This is much better, but far from perfect reasoning behavior: keep in mind that we specifically prompted these models to “be mindful of stoichiometry and ensure consistency with the given formula”, and models failed to follow this instruction if they didn’t continue to perform this check until they generated a consistent guess. NMR-Challenge.com records answer submissions, but doesn’t store those with incorrect formulas, so this particular error type could not be directly compared to human submissions.

The NMR-Challenge.com website data does allow us to evaluate more advanced errors, arising when the formula is correct, yet the answer is incorrect. Comparing these errors from LLMs and human participants reveals overlapping reasoning patterns and shared failure modes. Across the benchmark set of 115 problems, GPT-4o scored 100% in 17 cases (14 classified as Easy and 3 as Moderate), never generating an incorrect answer for us to consider. For these same tasks, human participants on the NMR-Challenge.com platform achieved success rates ranging from 41% to 100%, with an average of 80%. In contrast, GPT-4o scored 0% on 36 tasks (8 Easy, 15 Moderate, and 13 Hard). Human performance on these tasks was also significantly lower, with success rates between 8% and 64% (mean: 39%). Among the 38 tasks where GPT-4o did not succeed in 100% of the trials but generated structures with correct formulas—reveals further insight into the model’s behavior. In 8 of these cases, the GPT-4o output matched the most frequently submitted incorrect structure by human users. An additional 6 and 12 cases matched the second and third-to-fifth most common incorrect human submissions, respectively. In 10 cases, GPT-4o’s proposed structure was observed less frequently among human responses, and in 2 cases, GPT-4o’s answer was unique, and never submitted to NMR-Challenge.com. In contrast, o1 scored 100% in 68 tasks (43 Easy, 17 Moderate, 8 Hard), for which the average human success rate was 57%. It failed entirely on 7 tasks (1 Easy, 3 Moderate, 3 Hard), scoring 0% where the average human success rate was 35%. Among tasks not consistently solved by o1, only 7 featured incorrect molecular formulas in its most common erroneous output. For o1’s incorrect but formula-valid predictions (in 40 tasks), alignment with human errors was again notable: 13 matched the most common incorrect human structure, 3 matched the second most common, and 8 matched the third-to-fifth most common. In 12 cases, o1’s outputs were less frequently observed among human submissions. Two incorrect structures were not observed in the human dataset, and two were chemically invalid due to violations of basic valency rules (e.g., pentavalent carbon, trivalent oxygen). However, we cannot compare these to invalid molecule responses from humans, as the NMR-Challenge website provides strong feedback via the drawing tool, which is not provided in the LLM’s text-based environment. All incorrect structures proposed by the models are detailed in 5.3.

The most common structural errors made by the LLMs share many features with the most frequent incorrect submissions by human participants on NMR-Challenge.com [57]. Previous work has characterized the common failure modes of humans in these problems [14]; we find that LLMs likewise suffer from similar failure modes, especially misassigning the positions of substituents on aromatic rings and incorrectly orienting ester functional groups. These recurring mistakes highlight specific challenges in interpreting NMR spectra that are common to both human and machine reasoning. These results point to two possible conclusions about the nature of human and LLM reasoning in NMR structure determination. First, it may suggest a convergence between the reasoning pathways of LLMs and humans when interpreting NMR data, especially in the nature and distribution of incorrect structural hypotheses. However, even if LLMs and humans reason rather differently about these problems (such as when incorrect LLM-generated answers were never proposed

by humans), their answer distributions can still look similar when a handful of high-probability hypotheses dominate, while rare alternatives will remain hidden with standard sampling.

3.5 Suitability of NMR-Challenge.com as an LLM benchmark

Our benchmark includes 115 problems with experimental spectra, for which human performance data is available. With o1 already achieving 60% accuracy on Hard problems and outperforming top students (Figure 7), this suggests our benchmark is already or nearly saturated by existing LLMs (though they have some room for improvement). Generating synthetic NMR spectra of increasing difficulty could be considered next, though one would need to be careful in the problem design to avoid indeterminacy. We aimed to focus this benchmark on reasoning and non-linear problem-solving capabilities, and thus our pre-processing pipeline (multiplet analysis) transformed the raw spectra into a more abstract representation. If more than one structure could be a faithful explanation of the collection of multiplets, then grading would become problematic. The current dataset avoids ambiguity, having been carefully designed by human educators to have a single correct answer.

In terms of practical implications, we think these frontier chatbot models would not be the best choice for automating NMR structure determination in laboratory settings – that task would be better handled by specialized ML systems. Specifically purpose-built transformer models trained on millions of synthetic spectra [25, 27, 54] would almost certainly excel on our benchmark; such tools are reported to demonstrate effectiveness on molecules with 19 heavy atoms and trillions of isomers [27] (our dataset does include a few molecules with up to 23 heavy atoms (ID 174), but these have substantial symmetry that make them still tractable for human problem-solving). Such models might exploit subtle signal patterns in their large training datasets that wouldn't be reflected in human reasoning. General-purpose AI can and should be capable of understanding molecular structure through chemical reasoning, but don't necessarily need to be equipped for such intense tasks. We found anecdotally that chatbots can address highly related tasks, such as "What parts of this molecule correspond to this peak?" Such questions cannot even be posed to systems take raw spectra as input and output SMILES. Just as a chatbot should ideally be able to add 2-digit numbers, while switching to a calculator for evaluating 8-digit multiplication, a chemical reasoning model should ideally know when to switch from chain of thought reasoning in a domain it understands, to calling domain-specific tools (like a fine-tuned ML model for raw spectra-to-SMILES) when faced with a scenario in which reasoning is unlikely to be sufficient.

4 Conclusions

This work is an initial reasoning-focused LLM benchmark for evaluating chemical problem-solving in the NMR domain, in which chemistry knowledge is intertwined with complex, multistep reasoning. The study demonstrates that language models can interpret chemical structures from NMR spectroscopy data, with older, non-reasoning models like GPT-4o demonstrating significantly poorer performance than more recent reasoning-focused models like o1. o1 outperformed top undergraduate students in these tasks, but is not a perfect reasoner, continuing to make mistakes such as answering with a molecule whose formula is inconsistent with that given in the prompt. Language models implicitly use chain-of-thought reasoning, and neither temperature nor prompt engineering consistently affect their performance. Humans and LLMs generally found the same sets of problems to be easy and hard, and share common failure modes (e.g. incorrectly orienting ester functional groups), yet some LLM-generated answers had never been guessed by humans.

Acknowledgements

This work was supported by the Department of Energy's Bioenergy Technologies Office through the MSI STEM Research & Development Consortium, under contract number W911SR22F0104 (OR#64), and by the National Science Foundation (NSF) CAREER grant #2236769. We appreciate OpenAI for funding experiments with their GPT models, and Dr. Aaron Jaesch for advice. We especially thank Dr. Marcin Ptaszek for his assistance in setting up the experiments and integrating this work into his Organic Chemistry Laboratory class activities, as well as the students who volunteered and participated in the study, providing valuable data on human performance.

Conflicts of Interest

Although OpenAI provided credits for experiments with their models, they were not involved in the study design or evaluation, and were not provided with answers. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The code, prompts, and data supporting the findings of this work are available at the Atoms Lab GitHub repository: github.com/ATOMSLab/LLM-NMR.

References

- [1] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. *arXiv preprint arXiv:2501.04519*, 2025.
- [2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [3] N Lambert, J Morrison, V Pyatkin, S Huang, H Ivison, F Brahman, LJV Miranda, A Liu, N Dziri, S Lyu, et al. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. URL <https://arxiv.org/abs/2411.15124>, 2025.
- [4] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process Reinforcement through Implicit Rewards. *arXiv preprint arXiv:2502.01456*, 2025.
- [5] Maarten Grootendorst. A Visual Guide to Reasoning LLMs. <https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms>. Accessed: 2025-04-29.
- [6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022.
- [7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [8] Daniel Kahneman. *Thinking, Fast and Slow*. MacMillan, 2011.
- [9] Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnampati, Samuel G Rodrigues, and Andrew D White. Language Agents Achieve Superhuman Synthesis of Scientific Knowledge. *arXiv preprint arXiv:2409.13740*, 2024.
- [10] Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, et al. Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs. *arXiv preprint arXiv:2408.00114*, 2024.
- [11] David R Klein. *Organic Chemistry*. John Wiley & Sons, 2020.
- [12] Martin Dračinský. *NMR Spectroscopy for Chemists*. IOCB Prague, 2025. Accessed: 2025-07-01.
- [13] John Bigelow. Quine, Mereology, and Inference to the Best Explanation. *Logique et Analyse*, pages 465–482, 2010.
- [14] Zuzana Osifová, Ondřej Socha, and Martin Dračinský. NMR-Challenge. com: Exploring the Most Common Mistakes in NMR Assignments. *Journal of Chemical Education*, 101(6):2561–2569, 2024.
- [15] Jens Meiler and Martin Will. Automated Structure Elucidation of Organic Molecules from ^{13}C NMR Spectra using Genetic Algorithms and Neural Networks. *Journal of Chemical Information and Computer Sciences*, 41(6):1535–1546, 2001.
- [16] Zhaorui Huang, Michael S Chen, Cristian P Woroch, Thomas E Markland, and Matthew W Kanan. A Framework for Automated Structure Elucidation from Routine NMR Spectra. *Chemical Science*, 12(46):15329–15338, 2021.

- [17] Stefan Kuhn, Eda Tumer, Simon Colreavy-Donnelly, and Ricardo Moreira Borges. A Pilot Study for Fragment Identification using 2D NMR and Deep Learning. *Magnetic Resonance in Chemistry*, 60(11):1052–1060, 2022.
- [18] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, et al. Graph Neural Networks for Materials Science and Chemistry. *Communications Materials*, 3(1):93, 2022.
- [19] Chongcan Li, Yong Cong, and Weihua Deng. Identifying Molecular Functional Groups of Organic Compounds by Deep Learning of NMR Data. *Magnetic Resonance in Chemistry*, 60(11):1061–1069, 2022.
- [20] David Williamson, Santiago Ponte, Isaac Iglesias, Nicola Tonge, Carlos Cobas, and E Kate Kemsley. Chemical Shift Prediction in ¹³C NMR Spectroscopy using Ensembles of Message Passing Neural Networks (MPNNs). *Journal of Magnetic Resonance*, 368:107795, 2024.
- [21] Xin-Yu Lu, Hao-Ping Wu, Hao Ma, Hui Li, Jia Li, Yan-Ti Liu, Zheng-Yan Pan, Yi Xie, Lei Wang, Bin Ren, et al. Deep Learning-Assisted Spectrum-Structure Correlation: State-of-the-Art and Perspectives. *Analytical Chemistry*, 96(20):7959–7975, 2024.
- [22] Sriram Devata, Bhuvanesh Sridharan, Sarvesh Mehta, Yashaswi Pathak, Siddhartha Laghuvarapu, Girish Varma, and U Deva Priyakumar. DeepSPInN - Deep Reinforcement Learning for Molecular Structure Prediction from Infrared and ¹³-C NMR Spectra. *Digital Discovery*, 3(4):818–829, 2024.
- [23] Adrian Mirza and Kevin Maik Jablonka. Elucidating Structures from Spectra using Multimodal Embeddings and Discrete Optimization. *ChemRxiv*, 2024.
- [24] Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, and William S Noble. De novo Mass Spectrometry Peptide Sequencing with a Transformer Model. In *International Conference on Machine Learning*, pages 25514–25522. PMLR, 2022.
- [25] Marvin Alberts, Federico Zipoli, and Alain Vaucher. Learning the Language of NMR: Structure Elucidation from NMR Spectra using Transformer Models. In *Annual Conference on Neural Information Processing Systems*, 2023.
- [26] Lin Yao, Minjian Yang, Jianfei Song, Zhuo Yang, Hanyu Sun, Hui Shi, Xue Liu, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Conditional Molecular Generation Net Enables Automated Structure Elucidation based on ¹³C NMR Spectra and Prior Knowledge. *Analytical chemistry*, 95(12):5393–5401, 2023.
- [27] Frank Hu, Michael S Chen, Grant M Rotskoff, Matthew W Kanan, and Thomas E Markland. Accurate and Efficient Structure Elucidation from Routine One-Dimensional NMR Spectra using Multitask Machine Learning. *ACS Central Science*, 10(11):2162–2170, 2024.
- [28] Kehan Guo, Yili Shen, Gisela Abigail Gonzalez-Montiel, Yue Huang, Yujun Zhou, Mihir Surve, Zhichun Guo, Prayel Das, Nitesh V Chawla, Olaf Wiest, et al. Artificial Intelligence in Spectroscopy: Advancing Chemistry from Prediction to Generation and Beyond. *arXiv preprint arXiv:2502.09897*, 2025.
- [29] Wolfgang Bremser. Structure Elucidation and Artificial Intelligence. *Angewandte Chemie International Edition in English*, 27(2):247–260, 1988.
- [30] Robert K Lindsay, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. *Artificial intelligence*, 61(2):209–261, 1993.
- [31] Martin Will, Winfried Fachinger, and Joachim R Richert. Fully Automated Structure Elucidation: A Spectroscopist’s Dream Comes True. *Journal of Chemical Information and Computer Sciences*, 36(2):221–227, 1996.
- [32] Jean-Marc Nuzillard and Vicente de Paulo Emerenciano. Automatic structure elucidation through data base search and 2d nmr spectral analysis. *Natural Product Communications*, 1(1):1934578X0600100111, 2006.
- [33] Ryan L Stowe and Melanie M Cooper. Arguing from Spectroscopic Evidence. *Journal of Chemical Education*, 96(10):2072–2085, 2019.
- [34] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. Are Large Language Models Superhuman Chemists? *arXiv preprint arXiv:2404.01475*, 2024.

- [35] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From Text to Insight: Large Language Models for Chemical Data Extraction. *Chemical Society Reviews*, 2025.
- [36] Yoel Zimmermann, Adib Bazgir, Zartashia Afzal, Fariha Agbere, Qianxiang Ai, Nawaf Alampara, Alexander Al-Feghali, Mehrad Ansari, Dmytro Antypov, Amro Aswad, et al. Reflections from the 2024 Large Language Model (LLM) Hackathon for Applications in Materials Science and Chemistry. *arXiv preprint arXiv:2411.15221*, 2024.
- [37] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [38] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*, 2024.
- [39] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity’s Last Exam. *arXiv preprint arXiv:2501.14249*, 2025.
- [40] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can Large Language Models do in Chemistry? A Comprehensive Benchmark on Eight Tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.
- [41] Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. Assessing the Chemical Intelligence of Large Language Models. *arXiv preprint arXiv:2505.07735*, 2025.
- [42] Ondřej Socha, Zuzana Osifová, and Martin Dračinský. NMR-Challenge.com: An interactive website with exercises in solving structures from NMR spectra. *Journal of Chemical Education*, 2023.
- [43] Mnova Software Suite - Mestrelab. <https://mestrelab.com/main-product/mnova>. Accessed: 2025-04-29.
- [44] OpenAI. <https://platform.openai.com/docs/models>. Accessed: 2025-04-29.
- [45] Anthropic. <https://www.anthropic.com/news/claude-3-5-sonnet>. Accessed: 2025-04-29.
- [46] Gemini. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash>. Accessed: 2025-04-29.
- [47] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- [48] David Weininger. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.
- [49] CIRpy: Python interface for the Chemical Identifier Resolver. <https://github.com/mcs07/CIRpy>. Accessed: 2025-06-261.
- [50] RDKit: Open-source Cheminformatics. <http://www.rdkit.org>. Accessed: 2025-06-261.
- [51] Reasoning Best Practices. <https://platform.openai.com/docs/guides/reasoning-best-practices>. Accessed: 2025-04-29.
- [52] Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, 2024.
- [53] Mike Cvet. Temperature Scaling and Beam Search Text Generation in LLMs, for the ML-Adjacent. <https://towardsdatascience.com/temperature-scaling-and-beam-search-text-generation-in-llms-for-the-ml-adjacent-21212cc5dddb/#4e6e>. Accessed: 2025-04-29.
- [54] Martin Priessner, Richard Lewis, Jon Paul Janet, Isak Lemurell, Magnus Johansson, Jonathan Goodman, and Anna Tomberg. Enhancing molecular structure elucidation: Multimodaltransformer for both simulated and experimental spectra. *ChemRxiv*, 2024.
- [55] Xiaofeng Tan. A transformer based generative chemical language AI model for structural elucidation of organic compounds. *Journal of Cheminformatics*, 17(1):103, 2025.

-
- [56] Martin Priessner, Anna Tomberg, Jon Paul Janet, Richard Lewis, Magnus Johansson, and Jonathan Goodman. Enhancing Molecular Structure Elucidation with Reasoning-Capable LLMs. *ChemRxiv*, 2025.
- [57] Zuzana Osifová, Ondřej Socha, and Martin Dračinský. NMR-Challenge.com. <https://nmr-challenge.uochb.cas.cz/>. Accessed: 2025-07-01.