A MolPuzzle Benchmark Details

This section complements Section 3 with a fine-grained summary of the dataset collection, results validation, and evaluation procedure, along with a fuller characterization of the task instances and the corresponding prompts.

A.1 Data Collection

The initial molecules were selected by referencing the textbook *Organic Structures from Spectra, 4th Edition*, available as an online PDF on ResearchGate. We chose 234 molecules based on spectrum tasks involving IR, MS, ^{1}H-NMR, and ^{13}C-NMR to reflect a difficulty level suitable for graduate students[36].

To address copyright concerns, we excluded molecules with publicly available mass spectrometry (MS) spectra in open-source databases from our study. The remaining spectra were sourced from public resources, notably the PubChem database[37]. For additional spectra that were not available, we used simulation methods[38][39] and provided a Jupyter notebook to generate these data, ensuring high-quality spectra for analysis. Our final dataset comprised 200 molecules.

Given the challenges associated with NMR spectrum images, some spectra were obtained from simulated data in text format for ^{1}H-NMR and ^{13}C-NMR. This approach ensured clarity and accuracy in the evaluation of molecular structures.

To assess the multiple-stage abilities of LLMs, we designed a unique question-and-answer evaluation. This framework tested the LLMs' capabilities in interpreting and integrating data from different types of spectra, simulating real-world challenges. Details of this evaluation framework are provided in the next section.

A.2 Template design

Each template was crafted to target specific skills within molecular understanding. For instance, saturation identification challenges the models' ability to discern the degree of saturation in a molecule, which is crucial for understanding its chemical reactivity and stability. Aromatic ring identification tests the models' ability to recognize benzene-like structures, which are fundamental in organic chemistry due to their common occurrence and unique properties. Saturation degree calculation pushes the models to apply quantitative analysis, requiring not just recognition but also computation based on molecular structures.

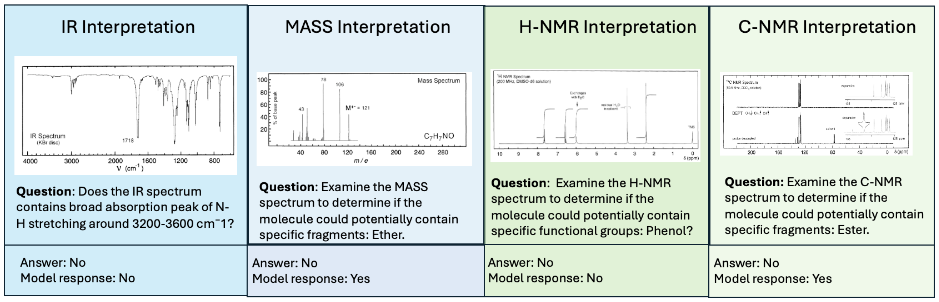
By diving deeper into the rationale behind each template and the kind of chemical knowledge they are designed to test, we can better appreciate how these tasks simulate real-world applications in chemistry. This approach not only tests the models' basic recognition abilities but also their capacity to perform complex reasoning and apply theoretical knowledge practically. The template examples are in A.3.

A.3 Stage1 QA Samples

Table 3: QA samples for the molecule understanding task

Task	Prompt
Saturation Identification	Question: Could the molecule with the formula C8H10O potentially be Saturated? Answer: No Model response: No.
Aromatic Ring Identification	Question: Could the molecule with the formula C8H10O have aromatic rings? Answer: Yes Model response: Yes.
Functional Group Identification	Question:Could the molecule with the formula C6H14O2 potentially contain a Amine group, given the Degree of Unsaturation is 0.0? Answer: No Model response: No, the molecule doesn't contain Amine group
Saturation Degree Calculation	Question: Calculate the Degree of Unsaturation of the molecule with the formula C8H10O? Answer: 4.0 Model response: 2

A.4 Stage2 QA Samples



Question: Does the IR spectrum contains broad absorption peak of N-H stretching around 3200-3600 cm⁻1?

Answer: No
Model response: No

Question: Examine the MASS spectrum to determine if the molecule could potentially contain specific fragments: Ether.

Answer: No
Model response: Yes

Question: Examine the H-NMR spectrum to determine if the molecule could potentially contain specific functional groups: Phenol?

Answer: No
Model response: No

Question: Examine the C-NMR spectrum to determine if the molecule could potentially contain specific fragments: Ester.

Answer: No
Model response: Yes

A.5 Stage3 QA Samples

Table 4: QA samples for the molecule construction task

Task	Prompt
H-NMR Elucidation	Question: Calculate the number of different types of hydrogen atoms present in the molecule, based on the given H-NMR: 4.51-4.61 (4H, 4.56 (s), 4.56 (s)), 7.06-7.32 (10H, 7.13 (dddd, J = 7.9, 7.7, 1.8, 0.6 Hz), 7.13 (dddd, J = 7.9, 7.7, 1.8, 0.6 Hz), 7.25 (dddd, J = 7.9, 1.5, 1.3, 0.6 Hz), 7.25 (dddd, J = 7.9, 1.5, 1.3, 0.6 Hz), 7.26 (tt, J = 7.7, 1.5 Hz), 7.26 (tt, J = 7.7, 1.5 Hz)) Answer: 4 Model response: 3.
C-NMR Elucidation	Question: Analyze the given C-NMR data and determine the number of different types of carbon atoms present in the molecule, based on given C-NMR: 39.3 (1C, s), 63.4 (1C, s), 127.8 (1C, s), 128.4 (2C, s), 128.8 (2C, s), 134.2 (1C, s). Only output the number. Answer: 6 Model response: 8

B Evaluation Experiments

B.1 Experimental Setting

During our testing phase, we selected 100 questions and employed two distinct prompting strategies with the large language model (LLM). Initially, the LLM was tasked with directly answering the questions. In a subsequent approach, the same queries were presented, but the model was prompted to execute a chain-of-thought reasoning process before responding. Each question in our dataset begins with a comprehensive description of the chemical context, along with specified answer formats and detailed guiding rules. To ensure a balanced representation of each task category, for tasks in Stage 1, the distribution ratio for Saturation Identification (SI), Functional Group Identification (FI), Aromatic Ring Identification (AI), and Saturation Degree Calculation (SC) is set at 2:3:3:2. In Stage 2, we have randomly selected 100 questions from each category of the spectrum. For Stage 3, we randomly selected 100 questions focused on H-NMR and C-NMR analyses.

We carried out this evaluation over three rounds, analyzing responses using both accuracy and the F1 score for tasks involving Saturation Identification (SI), Functional Group Identification (FI), and Aromatic Ring Identification (AI). For Saturation Degree Calculation (SDC), which yields numerical results, we assessed accuracy by comparing the count of correct matches to the ground truth data. The detailed results are reported in Table A.3. To ensure that all results are presented in a way that facilitates direct comparison, only those using similar evaluation metrics(AI, FI, AI) are included in the main table. For the SI, AI, and FI tasks, we use the F1 score and Accuracy to evaluate their performance since they are classification tasks. For the SDC task, the answer is a numerical number, so we only use the accuracy score to measure the performance of the LLMs. This approach helps to keep the evaluation coherent and focused on comparable data points.

B.2 Human Evaluation

To evaluate the performance of large language models (LLMs) on specialized tasks against expert humans, we recruited six graduate students from chemistry department to solve the MolPuzzle benchmark. These students, having recently completed a graduate-level course in Molecular Structural Elucidation, represented a highly skilled group of human participants.

For the experiment, we randomly selected six questions from the MolPuzzle dataset for each stage of the study. These questions were presented to the students in different formats according to the stage: In Stages 1 and 2, the questions were simple Yes/No or required short answers. In Stage 3, to align with the conventional methods chemists use to express chemical structures, students were asked to upload images of their hand-drawn structures instead of using SMILES strings. These images were manually compared to the ground truth to calculate scores.

We also imposed self-regulated time constraints to mirror the challenging nature of molecular structural elucidation. Beyond individual stage evaluations, we presented each participant with a complete molecule puzzle, consisting of a formula and four spectral images. The students were tasked with solving these puzzles within a 20-minute time frame. Impressively, all participants successfully submitted their solutions within the allotted period.

Our study included a component where human evaluators were involved to assess the performance of the AI models. To ensure the protection and ethical treatment of all participants, we conducted a thorough risk assessment. Potential risks identified included privacy concerns and the mental strain of repetitive tasks. Mitigation strategies, such as ensuring anonymity and providing breaks, were implemented to protect our evaluators.

The study was submitted for review and received approval from our Institutional Review Board (IRB). The IRB approval number is [insert approval number], which verifies that our protocols met all ethical guidelines for research involving human subjects. Throughout the project, we adhered strictly to these protocols to ensure ongoing compliance with ethical standards.

B.3 Stage1

Molecule understanding requires comprehensive analysis and interpretation of molecular structures, with a focus on chemical properties and spectroscopic data. In our study, we created a dataset of 234 molecules and developed eight distinct question templates across four categories: **Saturation Identification(SI), Functional Group Identification(FI), Aromatic Ring Identification(AI), and**

Saturation Degree Calculation(SC). These templates assess the ability to identify substructures, compute saturation levels, and infer structural presence, incorporating concepts in the chemistry reasoning process. Each question also necessitates a deep understanding of molecular bonding, stereochemistry, and functional group identification. Responses were generated using the RDKit library, ensuring precise and reliable answers grounded in established chemical informatics.

Table 3: The accuracy(↑), F1 score(↑)in 4 different molecule understanding categories, the best LLMs are in bold font.

Model	CoT	SI		AI		FI		SC
		F1	Acc	F1	Acc	F1	Acc	Acc
GPT-4o	-	**1±0.0**	**1±0.0**	0.943±0.016	0.944±0.015	0.934±0.005	0.966±0.0	0.667±0.003
GPT-4o	✓	**1±0.0**	**1±0.0**	**0.911±0.031**	**0.911±0.031**	0.689±0.025	0.766±0.027	0.816±0.062
GPT-3.5	-	0.451±0.025	0.825±0.075	0.816±0.017	0.816±0.075	0.826±0.075	0.683±0.016	0.5±0.099
GPT-3.5	✓	0.448±0.026	0.816±0.008	0.798±0.025	0.800±0.027	0.526±0.053	0.622±0.031	0.533±0.131
Claude-3-opus	-	0.361±0.009	0.556±0.023	**0.988±0.015**	0.988±0.015	**0.934±0.001**	0.966±0.001	**0.856±0.016**
Claude-3	✓	0.760±0.189	0.903±0.046	0.878±0.025	0.867±0.001	0.547±0.112	0.843±0.081	0.900±0.025
Gemini-pro	-	0.285±0.020	0.399±0.040	0.775±0.093	0.788±0.083	0.646±0.052	0.748±0.051	0.200±0.004
Gemini-pro	✓	0.391±0.045	0.651±0.108	0.685±0.088	0.688±0.087	0.562±0.018	0.629±0.023	0.283±0.062
LLama3	-	0.367±0.018	0.583±0.047	0.490±0.030	0.533±0.027	0.472±0.133	0.588±0.0	0.0±0.0
LLama3	✓	0.473±0.011	0.899±0.040	0.384±0.026	0.533±0.0	0.570±0.035	0.799±0.047	0.017±0.001
Vicuna-13b	-	0.031±0.022	0.033±0.025	0.500±0.087	0.522±0.083	0.308±0.038	0.311±0.041	0.0±0.0
Vicuna-13b	✓	0.380±0.023	0.616±0.062	0.342±0.006	0.522±0.157	0.516±0.080	0.855±0.016	0.0±0.0
Mistral-7b	-	0.221±0.014	0.283±0.025	0.384±0.005	0.500±0.0	0.319±0.014	0.322±0.157	0.0±0.0
Mistral-7b	✓	0.433±0.007	0.766±0.023	0.342±0.006	0.522±0.016	0.601±0.102	0.877±0.031	0.0±0.0

B.4 Stage2

The Spectrum interpretation tasks mainly measure the capability of LLMs in analyzing images related to identifying key substructures indicated by the spectrum plot. In this study, we utilize four distinct types of spectral images: nuclear magnetic resonance (NMR), infrared spectroscopy (IR), mass spectrometry, and others. Each type of data offers insights into various aspects of the molecular structure. We've created specific question templates for each spectrum, targeting peak and substructure identification factors. These templates are designed manually and emphasize the intricate connection between the spikes or troughs in the figures and the structures of the molecules. Responses were generated using the RDKit library to ensure correctness.

The findings from Stage 2 are presented in Table 4. We exclusively focus on the zero-shot learning outcomes, as our observations indicate that implementing chain-of-thought prompting leads to a deterioration in model performance. To address this limitation, we offer qualitative insights in B.6.

Table 4: The accuracy(↑), F1 score(↑) for IR, MASS spectrum, H-NMR, and C-NMR interpretation tasks."-" means the results are not interoperable

Model	Stage-2 Tasks							
	IR Interpretation		MASS Interpretation		H-NMR Interpretation		C-NMR Interpretation	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc
GPT-4o	**0.656±0.052**	**0.713±0.06**	**0.609±0.042**	**0.767±0.042**	**0.618±0.026**	**0.864±0.007**	**0.639±0.107**	**0.892±0.049**
Claude-3-opus	0.440±0.006	0.476±0.055	0.398±0.032	0.466±0.019	0.572±0.190	0.842±0.017	0.554±0.075	0.716±0.042
Gemini-3-pro-vision	0.194±0.002	0.119±0.016	0.116±0.036	0.124±0.038	0.545±0.048	0.851±0.062	0.492±0.016	0.619±0.044
LLava1.5-8b	0.256±0.026	0.414±0.044	0.101±0.021	0.104±0.26	0.118±0.008	0.186±0.011	0.254±0.015	0.472±0.023
Qwen-VL-Chat	0.243±0.027	0.392±0.043	0.125±0.006	0.116±0.021	0.255±0.007	0.611±0.031	-	-
InstructBLIP-7b	0.239±0.020	0.263±0.014	0.101±0.021	0.104±0.26	-	-	0.044±0.006	0.064±0.023
InstructBLIP-13b	0.239±0.020	0.263±0.014	0.101±0.021	0.104±0.26	-	-	0.047±0.014	0.067±0.025

B.5 Stage-3

Constructing a molecule involves a detailed analysis of NMR data, which is critical for understanding its structure. H-NMR data are essential as they provide information about the hydrogen environments within the molecule, including the number and types of hydrogen atoms (such as aliphatic or aromatic), as well as their connectivity. Conversely, C-NMR data offer in-depth insights into the carbon framework, illustrating the distribution and linkage of carbon atoms within the molecule. In our study, to evaluate the ability of large language models (LLMs) to interpret NMR data, we generated 1,171 question-and-answer (QA) pairs. These pairs focus on key NMR interpretation tasks, such as counting hydrogen atom types and identifying substructures, which are critical for accurate analysis.

Despite observing moderate accuracy from the LLMs in Stage 2 of our testing, we enhanced the quality of the QA pairs in Stage 3 by providing the LLMs with verified NMR data, generated by using nmrdb[40]. This approach ensures that the data used is reliable and helps maintain the integrity of our results. The findings from Stage 2 are presented in Table. We exclusively focus on the zero-shot learning outcomes, as our observations indicate that implementing chain-of-thought prompting leads to a deterioration in model performance. To address this limitation, we offer qualitative insights in

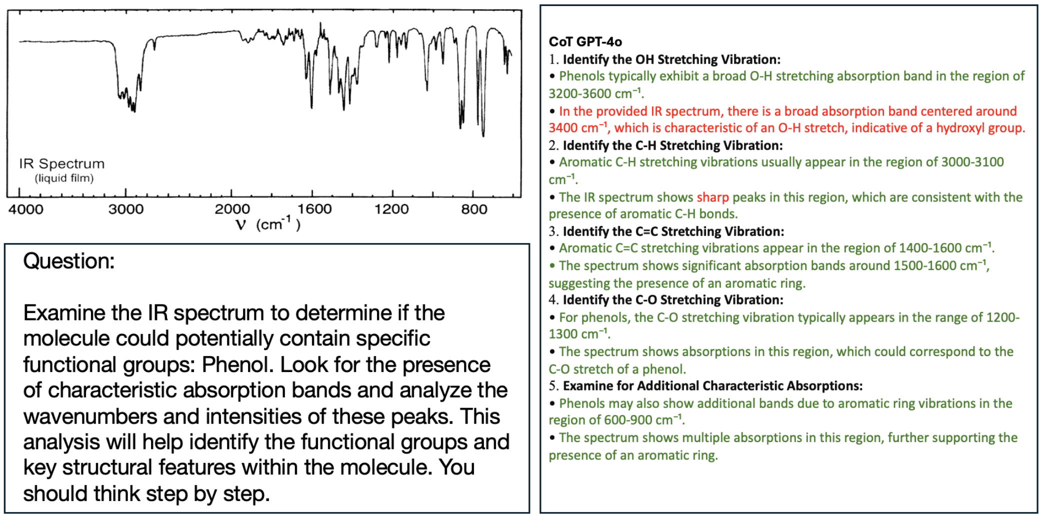
Table 5: The F1 score(↑) for H-NMR, and C-NMR Structure Elucidation

Method	H-NMR Elucidation	C-NMR Elucidation
GPT-4o	**0.433**±**0.013**	**0.411**±**0.034**
Claude-3-opus	0.395±0.008	0.313±0.029
Gemini-pro	0.333±0.012	0.308±0.031
Llama3	0.211±0.012	0.342±0.007
Vicuna-13b	0.181±0.013	0.244±0.001
Mistral-7b	0.131±0.032	0.122±0.027

B.6 Qualitative Results

In this section, we present several examples using GPT-4's chain-of-thought (CoT) reasoning to facilitate a clearer understanding of the results. We have enlisted two Ph.D. candidates from the chemistry department to evaluate these CoT outcomes. The analysis uses color coding to indicate the accuracy of the generated text: green signifies correct responses, red indicates incorrect ones, and yellow denotes responses that are partially correct.

B.6.1 Stage 2 examples



Question:

Examine the IR spectrum to determine if the molecule could potentially contain specific functional groups: Phenol. Look for the presence of characteristic absorption bands and analyze the wavenumbers and intensities of these peaks. This analysis will help identify the functional groups and key structural features within the molecule. You should think step by step.

CoT GPT-4o

1. **Identify the OH Stretching Vibration:**
- Phenols typically exhibit a broad O-H stretching absorption band in the region of 3200-3600 cm^{-1}.
- In the provided IR spectrum, there is a broad absorption band centered around 3400 cm^{-1}, which is characteristic of an O-H stretch, indicative of a hydroxyl group.

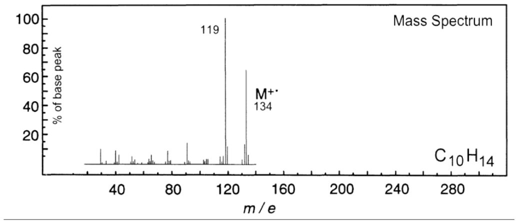
2. **Identify the C-H Stretching Vibration:**
- Aromatic C-H stretching vibrations usually appear in the region of 3000-3100 cm^{-1}.
- The IR spectrum shows sharp peaks in this region, which are consistent with the presence of aromatic C-H bonds.

3. **Identify the C=C Stretching Vibration:**
- Aromatic C=C stretching vibrations appear in the region of 1400-1600 cm^{-1}.
- The spectrum shows significant absorption bands around 1500-1600 cm^{-1}, suggesting the presence of an aromatic ring.

4. **Identify the C-O Stretching Vibration:**
- For phenols, the C-O stretching vibration typically appears in the range of 1200-1300 cm^{-1}.
- The spectrum shows absorptions in this region, which could correspond to the C-O stretch of a phenol.

5. **Examine for Additional Characteristic Absorptions:**
- Phenols may also show additional bands due to aromatic ring vibrations in the region of 600-900 cm^{-1}.
- The spectrum shows multiple absorptions in this region, further supporting the presence of an aromatic ring.

Figure 6: Human annotated IR spectrum interpretation



Question:
Examine the MASS spectrum to determine if the molecule could potentially contain specific fragments: Ether. Look into the number of fragments observed and analyze the differences between the larger fragments. This analysis will help identify the presence of key structural features within the molecule? You should think step by step

CoT GPT-4o: 1.Identify the Molecular Ion Peak (M⁺⁺):

- The molecular ion peak (M⁺⁺) is observed at m/z = 134 . This peak represents the molecular weight of the compound.

2. Identify Significant Fragment Peaks:

- The base peak (the most intense peak) is at m/z = 119 . This fragment is often very informative because it is the most stable fragment.
- Other significant peaks are observed at various m/z values (e.g., around 40, 60, 80, etc.).

3. Determine the Mass Difference Between Key Peaks:

- The difference between the molecular ion peak (134) and the base peak (119) is 15. This suggests a loss of a CH_3 (methyl) group, which is common in mass spectrometry fragmentation patterns.

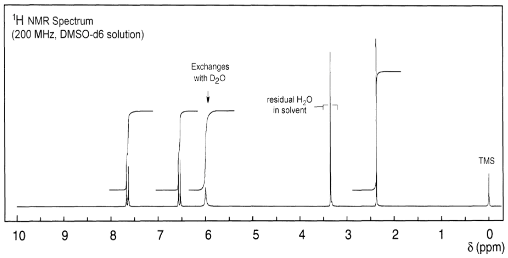
4.Identify Common Ether Fragmentation Patterns:

- Ethers typically show fragmentation patterns where the C-O bond breaks, leading to fragments that indicate the presence of an ether.
- A common ether fragmentation involves the loss of an alkyl group adjacent to the oxygen, often observed as a mass loss of 15 (CH_3), 29 (C_2H_5), or 43 (C_3H_7).

5. Analyze the Presence of Ether Characteristic Peaks:

- Given the mass spectrum data, the loss of 15 units (m/z = 134 to m/z = 119) could suggest the presence of a methyl group, potentially adjacent to an oxygen atom if an ether is present.

Figure 7: Human annotated MASS spectrum interpretation



Question:
Examine the H-NMR spectrum to determine if the molecule could potentially contain specific functional groups: Aldehyde? Look for the presence of characteristic absorption bands and analyze the wavenumbers and intensities of these peaks. This analysis will help identify the functional groups and key structural features within the molecule. You should think step by step

CoT GPT-4o: Analyzing the provided ^1H NMR spectrum gives us various insights into the molecular structure of the compound in question. Here's a step-by-step analysis:

Chemical Shifts and Splitting Patterns

The spectrum shows peaks at different chemical shift regions, which can be indicative of different types of hydrogen environments:

- **Around 9.5 ppm**: This is a typical region for aldehyde protons (CHO). The singlet peak here suggests the presence of an aldehyde group.
- **7.0 to 8.0 ppm**: These peaks typically indicate aromatic protons. The multiplicity of these peaks suggests that there may be multiple aromatic rings or substituted aromatic rings.
- **5.5 to 6.5 ppm**: Peaks in this region are often associated with hydrogen atoms attached to carbon atoms that are double bonded to other carbon atoms (e.g., vinyl protons) or to heteroatoms.
- **Around 1.5 ppm**: This region is characteristic of methyl groups (CH_3) adjacent to heteroatoms or double bonds.

Integration and Relative Intensity

- The spectrum does not provide integration values, but typically, the area under each peak corresponds to the number of equivalent protons contributing to that signal.

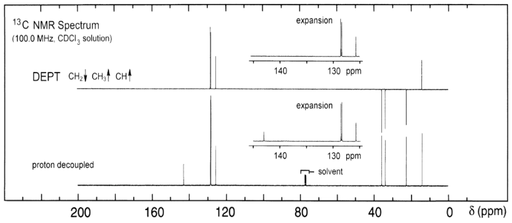
Identification of Functional Groups

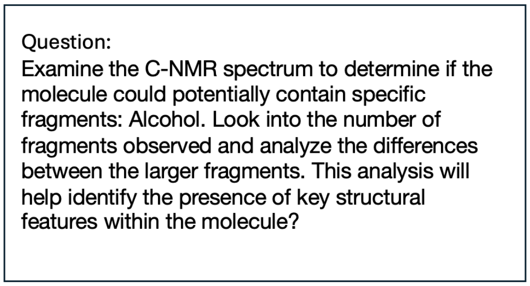
- **Aldehyde Group**: The presence of a sharp singlet peak at around 9.5 ppm strongly suggests an aldehyde functional group.
- **Aromatic Rings**: The cluster of peaks from 7.0 to 8.0 ppm suggests the presence of aromatic protons. The nature of splitting could indicate substitution patterns on the aromatic ring.
- **Vinyl or Alkene Groups**: Peaks in the 5.5 to 6.5 ppm range could indicate the presence of vinyl protons, suggesting a double bond adjacent to the aromatic ring or another functional group.
- **Methyl Groups**: The peak at about 1.5 ppm suggests the presence of methyl groups, possibly adjacent to a heteroatom or a double bond, influencing its chemical shift.

Figure 8: Human annotated H-NMR spectrum interpretation

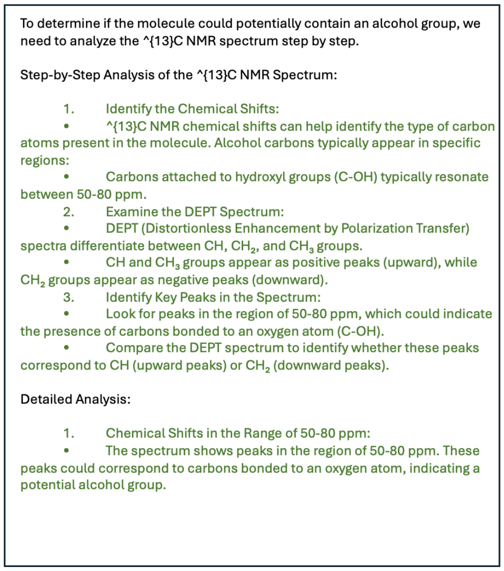
B.6.2 Stage 3 examples

Based on the observation in spectrum interpretation, it is safe to say that the GPT-4o model has a basic grasp of the concept and the task at hand, but their deficiency mainly resides in obtaining fully correct information from the spectrum images and also they lack in-depth understanding. As for the problems in Stage 3, although the GPT-4o outputs are mostly correct, they are too general which shows the poor understanding of GPT-4o models in solving this intricate task. The stage 3 task can be roughly broken down into 3 subtasks: obtaining the correct information from the spectrum image, deducing the correct structural information from the spectral information, and finally translating this structural information into a correct molecular structure. GPT-4o models seem to perform well in the second subtask, and moderately for simple structures in the third subtask but seem to be especially struggling with the first subtask in the case of NMR spectra. This indicates the gap in current LLMs in fully interpreting data therefore more advanced models and approaches should be developed to tackle the problem.





Question:
Examine the C-NMR spectrum to determine if the molecule could potentially contain specific fragments: Alcohol. Look into the number of fragments observed and analyze the differences between the larger fragments. This analysis will help identify the presence of key structural features within the molecule?



To determine if the molecule could potentially contain an alcohol group, we need to analyze the ^{13}C NMR spectrum step by step.

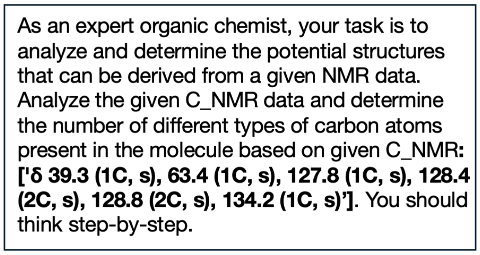
Step-by-Step Analysis of the ^{13}C NMR Spectrum:

1. Identify the Chemical Shifts:
   - ^{13}C NMR chemical shifts can help identify the type of carbon atoms present in the molecule. Alcohol carbons typically appear in specific regions:
   - Carbons attached to hydroxyl groups (C-OH) typically resonate between 50-80 ppm.
2. Examine the DEPT Spectrum:
   - DEPT (Distortionless Enhancement by Polarization Transfer) spectra differentiate between CH, CH_2, and CH_3 groups.
   - CH and CH_3 groups appear as positive peaks (upward), while CH_2 groups appear as negative peaks (downward).
3. Identify Key Peaks in the Spectrum:
   - Look for peaks in the region of 50-80 ppm, which could indicate the presence of carbons bonded to an oxygen atom (C-OH).
   - Compare the DEPT spectrum to identify whether these peaks correspond to CH (upward peaks) or CH_2 (downward peaks).

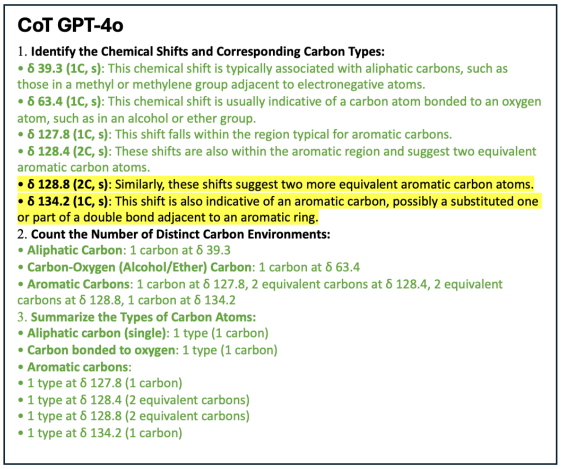
Detailed Analysis:

1. Chemical Shifts in the Range of 50-80 ppm:
   - The spectrum shows peaks in the region of 50-80 ppm. These peaks could correspond to carbons bonded to an oxygen atom, indicating a potential alcohol group.

Figure 9: Human annotated C-NMR spectrum interpretation



As an expert organic chemist, your task is to analyze and determine the potential structures that can be derived from a given NMR data. Analyze the given C_NMR data and determine the number of different types of carbon atoms present in the molecule based on given C_NMR: **['δ 39.3 (1C, s), 63.4 (1C, s), 127.8 (1C, s), 128.4 (2C, s), 128.8 (2C, s), 134.2 (1C, s)']**. You should think step-by-step.



CoT GPT-4o

1. **Identify the Chemical Shifts and Corresponding Carbon Types:**
- **δ 39.3 (1C, s)**: This chemical shift is typically associated with aliphatic carbons, such as those in a methyl or methylene group adjacent to electronegative atoms.
- **δ 63.4 (1C, s)**: This chemical shift is usually indicative of a carbon atom bonded to an oxygen atom, such as in an alcohol or ether group.
- **δ 127.8 (1C, s)**: This shift falls within the region typical for aromatic carbons.
- **δ 128.4 (2C, s)**: These shifts are also within the aromatic region and suggest two equivalent aromatic carbon atoms.
- **δ 128.8 (2C, s)**: Similarly, these shifts suggest two more equivalent aromatic carbon atoms.
- **δ 134.2 (1C, s)**: This shift is also indicative of an aromatic carbon, possibly a substituted one or part of a double bond adjacent to an aromatic ring.

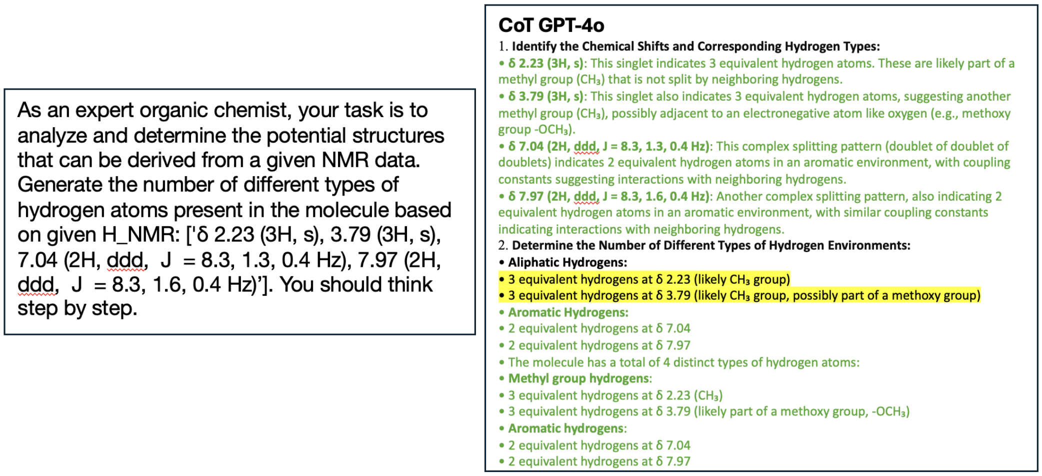
2. **Count the Number of Distinct Carbon Environments:**
- **Aliphatic Carbon**: 1 carbon at δ 39.3
- **Carbon-Oxygen (Alcohol/Ether) Carbon**: 1 carbon at δ 63.4
- **Aromatic Carbons**: 1 carbon at δ 127.8, 2 equivalent carbons at δ 128.4, 2 equivalent carbons at δ 128.8, 1 carbon at δ 134.2

3. **Summarize the Types of Carbon Atoms:**
- **Aliphatic carbon (single)**: 1 type (1 carbon)
- **Carbon bonded to oxygen**: 1 type (1 carbon)
- **Aromatic carbons**:
- 1 type at δ 127.8 (1 carbon)
- 1 type at δ 128.4 (2 equivalent carbons)
- 1 type at δ 128.8 (2 equivalent carbons)
- 1 type at δ 134.2 (1 carbon)

Figure 10: Human annotated H-NMR Elucidation

B.6.3 Complex Molecules

In addition to presenting molecules extracted from textbooks, we also demonstrate how the large language model (LLM) handles complex molecular structures. As illustrated in Figure 12, complex molecules typically have a larger pool of fragments. This expansion results in a greater number of valid elucidation paths, complicating the selection process for an appropriate starting point. Successfully navigating this enlarged pool necessitates an in-depth understanding of each fragment's properties and the associated, more intricate NMR data. In this context, LLMs may struggle because they often lack the nuanced chemical intuition and detailed analytical capabilities that human experts possess. Such limitations can lead to inaccuracies in interpreting complex interactions within NMR spectra, making LLMs less reliable for this task.



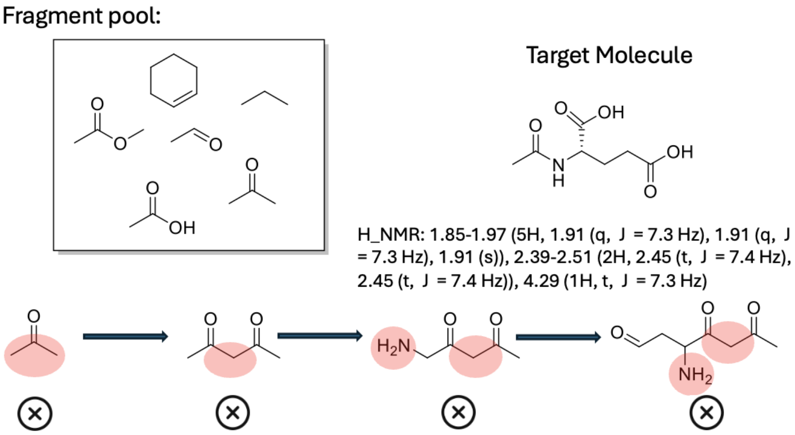
As an expert organic chemist, your task is to analyze and determine the potential structures that can be derived from a given NMR data. Generate the number of different types of hydrogen atoms present in the molecule based on given H_NMR: ['δ 2.23 (3H, s), 3.79 (3H, s), 7.04 (2H, ddd, J = 8.3, 1.3, 0.4 Hz), 7.97 (2H, ddd, J = 8.3, 1.6, 0.4 Hz)']. You should think step by step.

CoT GPT-4o

1. **Identify the Chemical Shifts and Corresponding Hydrogen Types:**
- **δ 2.23 (3H, s)**: This singlet indicates 3 equivalent hydrogen atoms. These are likely part of a methyl group (CH_3) that is not split by neighboring hydrogens.
- **δ 3.79 (3H, s)**: This singlet also indicates 3 equivalent hydrogen atoms, suggesting another methyl group (CH_3), possibly adjacent to an electronegative atom like oxygen (e.g., methoxy group $-OCH_3$).
- **δ 7.04 (2H, ddd, J = 8.3, 1.3, 0.4 Hz)**: This complex splitting pattern (doublet of doublet of doublets) indicates 2 equivalent hydrogen atoms in an aromatic environment, with coupling constants suggesting interactions with neighboring hydrogens.
- **δ 7.97 (2H, ddd, J = 8.3, 1.6, 0.4 Hz)**: Another complex splitting pattern, also indicating 2 equivalent hydrogen atoms in an aromatic environment, with similar coupling constants indicating interactions with neighboring hydrogens.

2. **Determine the Number of Different Types of Hydrogen Environments:**
- **Aliphatic Hydrogens:**
- 3 equivalent hydrogens at δ 2.23 (likely CH_3 group)
- 3 equivalent hydrogens at δ 3.79 (likely CH_3 group, possibly part of a methoxy group)
- **Aromatic Hydrogens:**
- 2 equivalent hydrogens at δ 7.04
- 2 equivalent hydrogens at δ 7.97
- The molecule has a total of 4 distinct types of hydrogen atoms:
- **Methyl group hydrogens**:
- 3 equivalent hydrogens at δ 2.23 (CH_3)
- 3 equivalent hydrogens at δ 3.79 (likely part of a methoxy group, $-OCH_3$)
- **Aromatic hydrogens**:
- 2 equivalent hydrogens at δ 7.04
- 2 equivalent hydrogens at δ 7.97

Figure 11: Human annotated C-NMR Elucidation



H_NMR: 1.85-1.97 (5H, 1.91 (q, J = 7.3 Hz), 1.91 (q, J = 7.3 Hz), 1.91 (s)), 2.39-2.51 (2H, 2.45 (t, J = 7.4 Hz), 2.45 (t, J = 7.4 Hz)), 4.29 (1H, t, J = 7.3 Hz)

Figure 12: Complex molecule Structure Elucidation

C Compute Resources

For the execution of various models in our experiments, distinct compute resources were utilized based on the model's accessibility and computational requirements. Specifically, for models like Claude 3, GPT, and Gemini, we employed API calls to facilitate their operation, leveraging the existing infrastructure provided by their respective platforms. This approach allowed us to access these models without the need for local computational resources, thereby streamlining the process. Conversely, for all other open-sourced models employed in our study, we conducted the experiments locally using an NVIDIA A100 GPU. This high-performance computing unit was chosen due to its advanced capabilities in handling extensive computations and large model requirements efficiently.