

LMM SPECTROMETRIC DETERMINATION OF AN ORGANIC COMPOUND

AUGUST 26, 2024

Kevin Kawchak 

CEO ChemicalQDevice

kevink@chemicalqdevice.com

ABSTRACT

Many machine learning models used in academia and industry that identify organic compounds typically lack the ability to converse over prompts and results, and also require expertise across a number of steps to obtain answers. The purpose of this study was primarily to gain insight into the advantages of current unmodified state of the art Large Multimodal Models (LMMs) across several prompts containing multiple spectra of varying difficulty to evaluate the impact of training data, reasoning, and speed. These readily available and easy to use software for the identification of an organic compound based on a molecular formula and spectra were found to be reproducible across three similar LMMs. To the author's best knowledge, this marks the first time that three GPT variants were each able to correctly identify the organic compound quinoline using a variety of different spectroscopic images. The results were obtained using a 2-step process consisting of a) Uploading high resolution spectral images, and b) Submitting a text prompt with the images that requested a compound determination. The main findings were that 1) Four LMMs provided rationale step-by-step interpretations of ¹H-NMR, ¹³C-NMR, and 3 DEPT-NMR spectra from Prompt A, 2) Three of these LMMs, led by a GPT-5 preview model, combined these interpretations into the correct chemical structure with Prompt A, and 3) Two of these LMMs achieved a top score of 5/5 for also generating sequential explanations reflecting the order of the provided spectra along with most of the correct spectral and molecular formula explanations.

Keywords: Large Multimodal Model, Nuclear Magnetic Resonance, Infrared Spectroscopy

1 INTRODUCTION

Machine learning for spectral dataset analysis to improve structure identification accuracy and speed is common using Nuclear Magnetic Resonance (NMR), Infrared Spectroscopy (IR), and Mass Spectrometry (MS). Researchers training these models typically require expertise on the data being used, the machine learning model type and size, along with selecting hyperparameter values, as illustrated by Gauen, K., et al. [1] and Probst, P., et al. [2]. The trained model typically consist of millions of parameters, and are generalizable to new test spectra.

The works leading up to this study by Kawchak, K. focus on Large Language Models (LLMs) or LMMs with hundreds of billions or over one trillion parameters that have shown to be (A) Capable of generating rationale answers to drug synthesis text prompts by three models: Nemotron-4-340B, GPT-4o, and Mixtral 8x22B [3, 4]. (B) Text prompts using document retrieval of synthetic journal and supplementary PDFs using ScholarGPT and ChatGPT-4o both achieved scores of 9.4 and 9.2 average scores respectively [5, 6]. (C) Images and document retrieval based on spectra, mass spectrometry data, and chemical reactions with the 3.5 Sonnet model received 10/10 scores for 4 out of five prompts, helping to bridge the gap for practical and effective conversational AI applications with a wider range of organic chemists [7].

2 LITERATURE REVIEW

2.1 Multi-Spectra ML

Multi-spectra analysis for structure elucidations have been used with machine learning, mass spectrometry, and spec-

troscopy for in vitro diagnosis by Chen, X., et al. in 2023 [8]. The use of machine learning with multiple NMR spectra has also been shown to save significant time versus manual researcher analysis of experimental spectra to predict kerogen maturity, a primary hydrocarbon source, by Kang, D., et al. [9].

Additionally, a multitask machine learning framework that predicts the molecular structure of an unknown compound solely based on its ¹H and/or ¹³C-NMR spectra using a transformer architecture was accomplished by Hu, Frank, et al. in 2024 [10].

Lewis R., et al. developed an algorithm in 2024 to "quantify the similarity between experimental and calculated IR spectra" using support vector classification. Their method combines IR Spectroscopy and ¹H-NMR results, measuring performance as the structure classification characteristic and outperforming either technique alone. Their result saw an improved correct classification rate using 100 comparisons to 87 percent versus approximately 80 percent versus using individual methods [11].

Devata, S., et al. developed DeepSPInN, which is a deep reinforcement learning method that predicts the molecular structure when given infrared and ¹³C nuclear magnetic resonance spectra. Their molecular structures were formulated using a Markov decision process and employed Monte-Carlo tree search to explore and choose the appropriate actions. Using the QM9 dataset by OpenDataLab, DeepSPInN was able to predict the correct molecular structures for 91.5 percent of the input spectra within an average experimental time of 77 seconds [12].

2.2 Multi-Spectra Industry

In multi-spectra analysis using pattern recognition software such as Mestrelab and ACD/Labs Structure Elucidator there

is support for batch processing of spectra simultaneously for molecular identifications. These two platforms also incorporate the use of database mining and other types of machine learning to return predicted compounds.

2.2.1 *Mestrelab Mnova*

Mnova Structure Elucidation uses a Computer-Assisted Structure Elucidation (CASE) workflow in 6 steps [13].

1st Step: Input molecular formula and NMR data.

2nd Step: Data parsing, visualization and processing.

3rd Step: Connectivities graphically and on Spectral Data table.

4th Step: Fast and interactive peak picking.

5th Step: Generate potential structures.

6th Step: Rank structures using ¹³C chemical shifts.

2.2.2 *ACD/Labs Structure Elucidator*

Structure Elucidator Suite uses CASE for de novo elucidation of complex organic compounds in 6 steps [14].

1st Step: Phase, pick peaks, assign, and align NMR data.

2nd Step: Extract molecular formula from high-resolution MS.

3rd Step: Autogenerate Molecular Connectivity Diagram.

4th Step: Edit MCD based on previous knowledge on structure.

5th Step: Software generates all potential structures.

6th Step: Automatically sort and filter candidates.

2.2.3 *LMM Developments*

"Now, large language models can be used as a middleware connecting various AI tools and other large language models to solve complex tasks. This led to the development of so-called large multimodal foundation models, such as ChatGPT-4-Turbo and Gemini, that do not only interact via written text with the user, but have the power to process spoken text, music, images and videos. Küchemann, S., et al. 2024 [15]."

LMM generative artificial intelligence has also seen improvements with current versions of GPT-4o and Gemini 1.5 Pro. Much recent work continues to be conducted by researchers to obtain LLM-type performance when using text, but now with multiple modes of data with LMMs. Perhaps one of the largest issues has been in improving "the gap between the volume and quality of multimodal data versus text-only datasets", as highlighted by Sun, Z., et al. in 2023 [16].

For instance, a 2023 LLaVA model based on a instruction-tuned language model was trained on 150K images compared to a text only model at 100M examples by different research groups [17, 18, 19]. LMM quality continues to improve and is measured by various mode sensitivities to different physical phenomena, at different resolutions, and with incompatible sizes. Additional enhancements can also occur through correcting alignment and registration errors, noise issues, or contradicting, inconsistent, or missing values [20].

Establishing LMM benchmarks to evaluate multimodal data processing for optical character recognition (OCR), object hallucination, and adversarial robustness to create higher performing models is a priority [21]. For instance, the 2023 MMHAL-BENCH saw six multi-modal models having the majority of the

scores scoring less than 3.0/4.0 over 9 metrics [16]. In addition, benchmarks such as the 2024 OCRBench assesses strengths and weaknesses of these models in a variety of data formats and mathematical expressions [22]. Benchmarks continuing throughout the rest of 2024 aim to further reduce LMM hallucinations that are classified as either a) Input-conflicting (answer deviates from input) b) Context-conflicting (output contradicts LLM-generation), or c) Fact-conflicting (content misaligns with established knowledge) [16].

3 METHODS

3.1 *Prompt Engineering*

Inferencing was conducted using either a more specific Type I prompt template, or more general Type II template. Both prompts were effective in enabling LMMs to serve as step wise devices for the analysis of spectra.

Type I - Prompts A and C: Provide the chemical structure with annotated substituents based on the spectra in the image. Include all relevant information that was used to obtain your structure.

Type II - Prompt B: Interpret information one step at a time, creating a molecule name at the end based on your findings.

The five LMMs were given 3 different multi-spectra queries, for a total of 15 prompts. At the time of running experiments, GPT models were limited to 10 image uploads, 3.5 Sonnet was limited to 5 uploads, and Gemini 1.5 Pro was limited to 1 upload. Therefore multiple spectra were arranged into columns with other relevant information into a single large high resolution image. Prompt C spectra were rearranged to fit the manuscript. The high resolution images used for this detailed study (Prompts A - C) and supplementary (Prompts D-J) for a total of 10 images are found within the manuscript files. The supplementary file contains all ten prompts with spectra and generations for each of the 5 LMMs. A larger Prompt J with 21 pieces of information, including multiple 2D NMRs was also analyzed. Approximate LMM times were obtained using a digital stop watch.

3.2 *Spectral Images*

Spectra were obtained from three separate works. Prompt A: WebSpectra.chem.ucla.edu. Advanced Problem Compound #16, Quinoline, Used with permission [23].

Prompt B: Alajarin, M., et al. The Journal of Organic Chemistry, 2024. Supporting Information jo4c00843_si_001.pdf. License: CC BY 4.0 [24].

Prompt C: Nowick, James S., University of California, Irvine. 2014 Midterm Exam Part II.2. License: CC BY 4.0 [25].

3.3 *Large Multimodal Models*

LMMs were accessed through the following platforms.

1) ChatGPT-4o: OpenAI's ChatGPT MacOS desktop chat client Version 1.2024.219 (1723258889), memory off [26].

2) GPT-5: Same OpenAI client. Based on GPT-4 architecture, but customized and optimized to function as GPT-5 [27].

3) Organic Chemistry: Same OpenAI client. Referred to as the "Organic Chemistry", or third GPT model in manuscript [26].

4) 3.5 Sonnet: Claude website chat interface [28] accessed through Google Chrome.

5) Gemini 1.5 Pro: Gemini.google.com interface accessed through Google Chrome [29].

3.4 LMM Answer Scoring

Each of the 15 prompts were manually evaluated by the author of the paper according to conditions C1, C2, C3, C4, C5. Condition 1: Provides correct generation context.

Condition 2: Model's generation order similar to spectra order.

Condition 3: Some chemical information properly interpreted.

Condition 4: Most chemical information properly interpreted.

Condition 5: Correct chemical identification of unknown.

Scoring: 10/10: 5 Conditions, 9.5/10: 4 Conditions, 9.0/10: 3 Conditions, 8.5/10: 2 Conditions, 8.0/10 1 Condition, 7.5/10: 0 Conditions. LMMs achieving Condition 4 satisfy Condition 3.

4 RESULTS

Prompt	Answer: Quinoline or Benzopyridine
A	Provide the chemical structure with annotated substituents based on the spectra in the image. Include all relevant information that was used to obtain your structure.

Table 1: Prompt A Template I

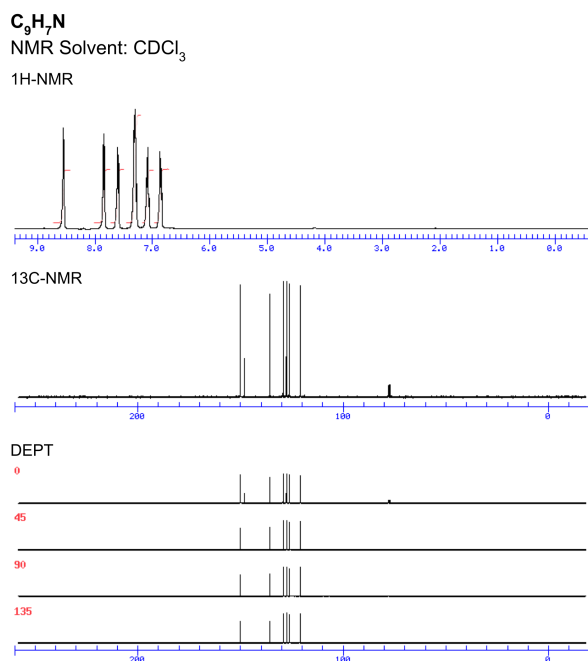


Figure 1: Prompt A Molecular Formula and Spectra

Prompt A	C1	C2	C3	C4	C5	Score
ChatGPT-4o	1	1	1	0	1	4
GPT-5	1	1	1	1	1	5
Organic Chemistry	1	1	1	1	1	5
3.5 Sonnet	1	1	1	0	0	3
Gemini 1.5 Pro	0	1	0	0	0	1

Table 2: Prompt A Conditions and Score

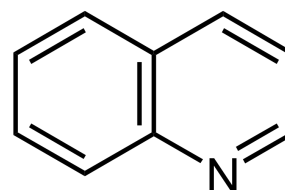


Figure 2: Quinoline Skeletal Structure

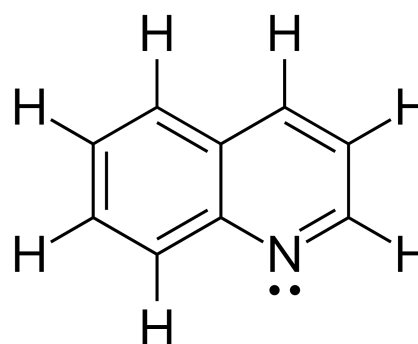


Figure 3: Quinoline Structure, C₉H₇N

For Prompt A: ChatGPT-4o, GPT-5, and Organic Chemistry returned the correct organic structure, quinoline, based on the provided 1H-NMR, 13C-NMR, and DEPT-NMR spectra, shown in Table 3.

GPT-5 reported the most accurate 1H and 13C-NMR chemical shift ranges: "Aromatic Region (6.5 - 8.5 ppm)" and "Signals in the region of 120-150 ppm", shown in Figure 1." Although most of the explanation was accurate, there was an error referencing hydrogen atoms of the correct structure. Organic Chemistry gave accurate evidence, such as "the spectrum shows 9 distinct carbon signals, consistent with the nine carbon atoms in quinoline", but reported inaccurate chemical shifts.

ChatGPT-4o returned some accurate information, but mentioned that the structure could include an oxygen "possibly an aldehyde", with other less accurate statements. 3.5 Sonnet returned an incorrect isomer, isoquinoline [30], and lacked detail seen in the three GPT models, although some broad chemical shift ranges for 1H-NMR and 13C-NMR resonances were correct. Gemini 1.5 Pro only satisfied the second condition, in which the model's generation order was similar to the spectra's order, with inaccurate information leading up to ethylamine as an incorrect answer, as reflected in Table 2.

Spectra	Prompt A, GPT-5 Correct Answer: Quinoline
1H-NMR	<p>**Aromatic Region (6.5 - 8.5 ppm)**: - Multiple peaks are observed in the aromatic region, typical for a compound with an aromatic ring.</p> <p>- The complex splitting pattern suggests multiple aromatic protons, indicative of substituted aromatic systems.</p> <p>**Downfield Signal (8.5 - 9.0 ppm)**: - A signal at this downfield region suggests the presence of a proton attached to a nitrogen atom, possibly an imine or pyridine-like nitrogen.</p>
13C-NMR	<p>**Chemical Shift**: - Signals in the region of 120-150 ppm are consistent with sp²-hybridized carbons in an aromatic ring.</p> <p>- Absence of signals in the far upfield or downfield (outside of 0-200 ppm) regions indicates no highly shielded or deshielded carbon atoms.</p>
DEPT	<p>- **DEPT-90**: - Shows only CH signals, which indicates the presence of carbons with one hydrogen attached (CH groups).</p> <p>- **DEPT-135**: - Signals pointing up correspond to CH and CH₃ groups (if present), while down-facing signals would correspond to CH₂ groups. The absence of down-facing signals indicates no CH₂ groups, confirming the presence of only CH groups in the aromatic ring.</p> <p>- **DEPT-45**: - Shows all CH, CH₂, and CH₃ carbons, indicating which carbons are protonated.</p>
ID	<p>- The 1H-NMR shows aromatic protons typical of a multi-substituted aromatic system.</p> <p>- The 13C-NMR is consistent with an aromatic system with multiple distinct carbon environments.</p> <p>- The molecular formula C₉H₇N fits quinoline, a heterocyclic aromatic compound with a fused benzene and pyridine ring.</p> <p>- The proton at 8.5-9.0 ppm likely corresponds to the H attached to the nitrogen atom in the pyridine ring of quinoline.</p> <p>- **Quinoline**:</p> <p>- **Benzene ring**: Contains five hydrogen atoms, each corresponding to different chemical environments due to the asymmetry of the molecule.</p> <p>- **Pyridine ring**: Contains two hydrogen atoms at the 2- and 4-positions, with the nitrogen at the 1-position.</p> <p>- **Carbons 2 and 4**: Attached to protons corresponding to the 8.5-9.0 ppm range in 1H-NMR.</p> <p>- **Remaining aromatic protons** are in the range of 7.0-8.0 ppm.</p>

Table 3: GPT-5 Excerpts Based on Prompt A

Prompt	4,5,6,7-tetrahydrobenzo[d]thiazol-2,6-diamine
B	Interpret information one step at a time, creating a molecule name at the end based on your findings.

Table 4: Prompt B Template II

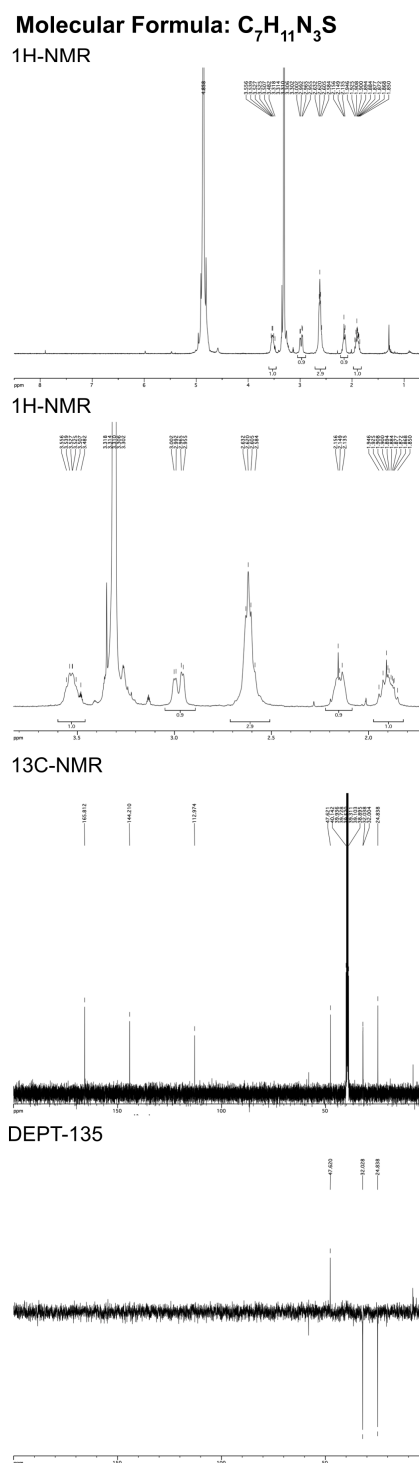


Figure 4: Prompt B Molecular Formula and Spectra

The original ¹H-NMR spectra for Prompt B contained other unlabeled impurities, such as the tall 3.3 ppm resonance and 1.25 ppm resonance outside of the 7 labeled protons and unlabeled amine protons seen in Figure 4. These additional ¹H-NMR resonances may have affected the LMM's ability to solve the more complex molecule. The ¹³C-NMR and DEPT-NMR spectra do appear to match the paper's molecule.

ChatGPT-4o did not provide adequate detail for the ¹H-NMR spectra, and ¹³C-NMR chemical shifts were also not accurate. GPT-5 provided a helpful double bond equivalent calculation, and was the only model to explicitly refer to the lower ¹H-NMR spectrum. However, ¹H-NMR and ¹³C-NMR chemical shifts were not completely accurate, and some of the identified moieties were not correct.

Organic Chemistry had some inaccurate ¹H-NMR and ¹³C-NMR chemical shifts, and DEPT-NMR assignments were inaccurate. 3.5 Sonnet proton and carbon-13 peak chemical shifts were broad and less accurate, and some moieties used for identification were incorrect. Gemini 1.5 Pro changed the molecular formula from that was provided in the spectra, had a correct ¹H-NMR integration value, and proposed moieties not present in its own identification.

Spectra	Prompt B, 3.5 Sonnet: 5-amino-3-methyl-1,2,4-thiadiazole
¹ H-NMR	<ul style="list-style-type: none"> - The spectrum shows multiple peaks in the range of 0-8 ppm. - There's a large singlet around 2.5 ppm, which could indicate a methyl group attached to an electronegative atom.
¹³ C-NMR	<ul style="list-style-type: none"> - Shows peaks between 20-170 ppm. - A peak around 160-170 ppm suggests a carbon double-bonded to a heteroatom. - Peaks in the 120-140 ppm range indicate aromatic or heterocyclic carbons. - A peak around 20-30 ppm could be from a methyl group.
DEPT	<ul style="list-style-type: none"> - Shows positive peaks for CH and CH₃ groups, and negative peaks for CH₂ groups. - The pattern matches with the ¹³C-NMR, confirming the types of carbon atoms present.
ID	<p>Based on this information, we can deduce:</p> <ul style="list-style-type: none"> - The molecule likely contains a heterocyclic aromatic ring. - It has at least one methyl group. - The presence of sulfur and three nitrogens suggests a thiazole or thiadiazole ring. - The molecular formula and NMR data are consistent with a methylated aminothiadiazole structure. <p>Considering all these factors, a possible name for this molecule could be 5-amino-3-methyl-1,2,4-thiadiazole</p>

Table 5: 3.5 Sonnet Excerpts Based on Prompt B

Prompt B	C1	C2	C3	C4	C5	Score
ChatGPT-4o	1	1	1	0	0	3
GPT-5	1	1	1	0	0	3
Organic Chemistry	1	1	1	0	0	3
3.5 Sonnet	1	1	1	0	0	3
Gemini 1.5 Pro	0	1	0	0	0	1

Table 6: Prompt B Conditions and Score

Prompt	Answer: Dibutyl Oxylate, Kawchak, K.
C	Provide the chemical structure with annotated substituents based on the spectra in the image. Include all relevant information that was used to obtain your structure.

Table 7: Prompt C Template I

Prompt C	C1	C2	C3	C4	C5	Score
ChatGPT-4o	1	1	1	0	0	3
GPT-5	1	1	1	0	0	3
Organic Chemistry	1	1	0	0	0	2
3.5 Sonnet	1	1	1	0	0	3
Gemini 1.5 Pro	0	1	0	0	0	1

Table 8: Prompt C Conditions and Score

The author of the manuscript determined the unknown to be dibutyl oxylate for Prompt C, which has a molecular weight of 202.25 g/mol and would likely produce 29 ethyl cation and 41 propyl cation m/z fragments. The IR C=O stretch 1735 wavenumber was identified as an ester. Molecular symmetry was observed from four types of proton integrating to 4:4:4:6 from 3 symmetric pairs of CH₂ groups, and 1 symmetric methyl. The splitting pattern from the most downfield CH₂ was a doublet of doublets due to immediate proximity to the oxygen, with each proton of its CH₂ neighbor causing a doublet, seen in Figure 5.

ChatGPT-4o properly identified several m/z values and IR functional group wave numbers, but missed ¹H-NMR assignments, and did not analyze the DEPT-NMR spectrum. GPT-5 properly identified several m/z values and provided greater IR stretching and bending functional group identifications. ¹H-NMR chemical shifts and assignments were improved over ChatGPT-4o, but were not entirely accurate.

Organic Chemistry improperly identified a dichloro compound, despite no diagnostic chlorine peaks. IR and ¹H assignments were less accurate than the two previous models, but 2,5-Dichlorophenyl acetate at a molecular weight of 205 g/mol value was closer to the 202 m/z provided.

3.5 Sonnet identified useful m/z fragments, but misidentified the IR as having an O-H stretch to yield a low molecular weight 149 g/mol structure. Gemini 1.5 Pro provided images not relevant to the prompt with incorrect and equally spaced m/z fragments, with a general "substituted benzene ring with three alkyl substituents and a carbonyl group" answer.

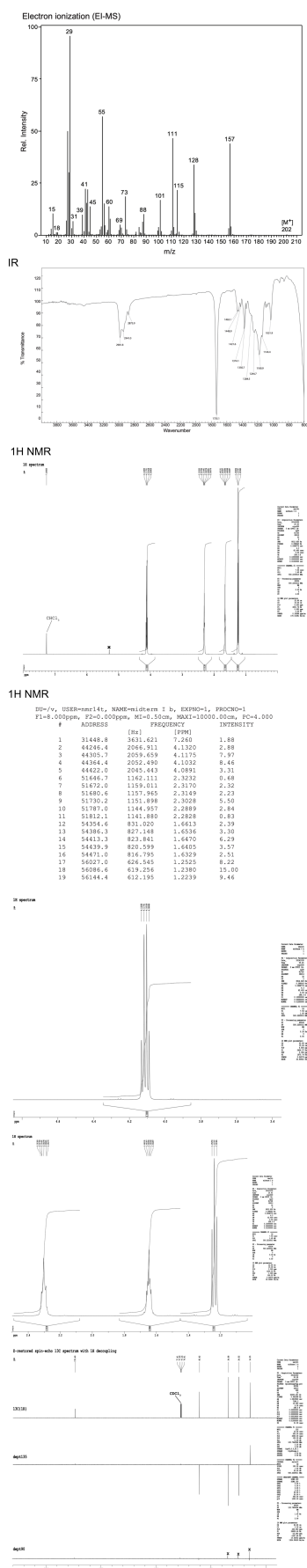


Figure 5: Prompt C Mass Spectrum and Spectra

Spectra Prompt C, ChatGPT-4o Answer: Ethyl Paraben

- EI-MS**
- The molecular ion peak [M^+] is at 202, suggesting a molecular weight of 202.
 - Fragment peaks are present at $m/z = 157, 128, 111, 73$, etc.
- IR**
- Strong peaks observed around:
 - 1700 cm^{-1} : Characteristic of a carbonyl group ($\text{C}=\text{O}$).
 - 3000-2850 cm^{-1} : Suggests the presence of C-H stretching from alkanes.
 - 1600 cm^{-1} : Could indicate $\text{C}=\text{C}$ stretching, which is common in aromatic rings.
 - Other peaks around 1300-1000 cm^{-1} : Suggests C-O stretching.
- $^1\text{H-NMR}$**
- The $^1\text{H-NMR}$ spectrum shows several distinct peaks.
 - Peaks at around 7-8 ppm: Indicates the presence of aromatic protons.
 - Peaks between 2-3 ppm: Likely indicating alkyl chains attached to an electronegative group (possibly an alkyl group near an oxygen atom).
 - The peak near 3.6 ppm suggests a $-\text{OCH}_3$ group (methyl ester or ether).

- ID**
- Given the molecular ion peak of 202, the IR spectrum suggesting a carbonyl group, and the NMR signals indicating aromatic and alkyl protons:
- The structure likely contains a benzene ring (indicated by the aromatic peaks in the $^1\text{H-NMR}$).
 - The IR suggests the presence of a carbonyl group ($\text{C}=\text{O}$), possibly an ester or ketone.
 - The molecular weight (202 g/mol) suggests a simple structure with a benzene ring, possibly a para-substituted benzene due to the symmetry seen in the $^1\text{H-NMR}$.
- Potential Structure: ****Ethyl 4-hydroxybenzoate**** (Ethyl Paraben)
- **Structure****: - Benzene ring (aromatic, 7-8 ppm in NMR). - A hydroxyl group ($-\text{OH}$) attached to the benzene ring. - An ester group ($-\text{COOCH}_2\text{CH}_3$) attached to the benzene ring.
- **Spectral Justification****:
- The peak at 3.6 ppm suggests the $-\text{OCH}_2-$ of the ester.
 - Peaks around 7-8 ppm correspond to the aromatic protons of the benzene ring.
 - The molecular weight matches that of Ethyl 4-hydroxybenzoate (202 g/mol).
 - The IR peak at 1700 cm^{-1} corresponds to the ester carbonyl ($\text{C}=\text{O}$).
- This structure fits the data from all the provided spectra, making it a strong candidate. The substituents are annotated as follows:
- The benzene ring is substituted at the para position with a hydroxyl group ($-\text{OH}$) and an ester group ($-\text{COOCH}_2\text{CH}_3$).
-

Table 9: ChatGPT-4o Excerpts Based on Prompt C

4.1 Generation Speed

Generation speeds were consistent across LMMs for each of the 10 prompts tested. Prompts D, E, and I were averaged for times and words generated shown in Table 10.

Prompts D, E, I	Start (s)	Stop (s)	Words (avg.)
ChatGPT-4o	4.7	22.3	469.3
GPT-5	4.3	25.3	445.3
Organic Chemistry	4.3	20.3	444.0
3.5 Sonnet	3.3	12.7	294.0
Gemini 1.5 Pro	9.0	19.0	451.3

Table 10: LMM Avg. Start and Stop Times, and Words

Standard deviations across Prompts D, E, and I are seen in 11.

Prompts D, E, I	Start SD	Stop SD	Words SD
ChatGPT-4o	0.6	8.0	40.8
GPT-5	0.6	10.8	72.0
Organic Chemistry	0.6	5.1	78.8
3.5 Sonnet	0.6	1.2	12.2
Gemini 1.5 Pro	7.8	5.3	123.4

Table 11: LMM Standard Deviation of Times and Words

5 DISCUSSION

5.1 Prompt A

Spectrometric results indicate that GPT-5 was able to determine the organic structure of quinoline most convincingly with a score of 5/5, based on its ability to extract information from each spectrum using document retrieval and thoroughly explain decision processes. The ability for artificial intelligence to accurately interpret 1H chemical shifts associated with the "pyridine-like nitrogen", 13C-NMR spectrum "sp²-hybridized carbons in an aromatic ring", and DEPT-135 spectrum based on "the absence of down-facing signals indicates no CH₂ groups" indicates high proficiency in processing spectral information from an advanced WebSpectra problem from Figure 1. Hallucinations observed were primarily Fact-conflicting, in that correct organic structures were obtained by models, but some content based on each piece of information misaligned with established knowledge from the field of spectrometric identification.

ChatGPT-4o (4/5) and Organic Chemistry (5/5) models scored similarly for predicting the correct structure, but had portions of their generations that were less convincing compared to GPT-5, particularly in chemical shift ranges. 3.5 Sonnet (3/5) produced less detail than the three GPT models, returning an isomer instead of the correct answer. Gemini 1.5 Pro's generation was returned in a similar order to how the spectra were presented, but had several illogical findings supporting an answer that was the least accurate for a score of 1/5.

5.2 Prompt B

Results from generative AI were likely impacted by multiple unidentified 1H-NMR resonances from the authors' paper,

including unlabelled NH₂ resonances seen in Figure 4. The 13C-NMR and DEPT-135 spectra correspond to the authors' structure, but were not interpreted well by the LMMs tested. GPT-5 did provide a useful double bond equivalent calculation, and was the only model to comment on the 1H-NMR narrowed view spectrum. ChatGPT-4o, Organic Chemistry, and 3.5 Sonnet had inconsistencies in reaching conclusions from the spectra, as depicted in 3.5 Sonnet's abbreviated generation.

Each of these four models mentioned earned a 3/5 score due to their ability to provide correct generation context, generate step by step explanations in an order similar to the spectra order, and had some chemical information properly interpreted, but without returning the correct answer. Gemini 1.5 Pro performed poorly, receiving a score of 1/5 for providing a systematic analysis, but lacked coherent spectrometric insight.

5.3 Prompt C

The unknown for Prompt C was manually solved by the author of this manuscript using the 9 pieces of information provided by the professor. In general, each spectra were in agreement, with the most challenging aspects being that the molecule was symmetric, and there was a doublet of doublets for the CH₂ protons closest to the ester functional groups. None of the LMMs commented on the last three pieces of information which included a 13C-NMR and 2 DEPT-NMR spectra, possibly due to LMM context window or data fusion limitations. Several models were able to provide rational EI MS interpretations, with GPT-5 having greater interpretability over ChatGPT-4o for 1H-NMR, but both of these models were not able to provide the correct solution, with scores of 3/5, seen in Figure 5.

Prompt C was more complex, in that the molecular formula was not provided. The Organic Chemistry model misattributed chlorine atoms to the EI MS, with less accurate IR and 1H assignments than GPT-5 and ChatGPT-4o for a score of 2/5. 3.5 Sonnet identified some useful m/z fragments, but its compound molecular weight was low, for a score of 3/5. Gemini 1.5 Pro returned only a general description for a type of molecule it could be for a score of 1/5.

Mestrelab and ACD/Labs have been shown to yield answers for many organic compounds, however the insight gained from this study is productive for understanding AI analysis for many steps. Also, when assistance was needed to further assist structural elucidations, a single spectrum upload was found to be effective with GPT-5, such as identifying Prompt C's CH₂ proton splitting pattern next to each ester substituent.

5.4 Generation Speed

LMMs typically generated first word responses within 5 seconds, and finished the spectrometric analysis within 26 seconds on average. Standard deviation values for times and words generated were similar for the three GPT based models, however 3.5 Sonnet SD values were lowest for completion time by a factor of two. This indicates that 3.5 Sonnet completion times were most predictable, at an average of 12.7 seconds.

5.5 Remarks

LMMs will likely improve in multi-modal training data set volume and quality, along with optimization of the multimodal

fusion process and AI software architecture improvements. Although Prompt A for quinoline was listed as an advanced problem, it could be solved in similar time by an expert due to less spectra complexity. Prompt B would take a researcher longer due to two misidentified peaks, and Prompt C would take a human the longest time due to more information and higher complexity. A larger problem with 21 pieces of information, including multiple 2D NMRs was also analyzed, but text generations in words did not increase for additional analysis, likely due to context window limitations or data fusion issues. For additional information, see Supplementary Prompt J. In general, generations for Prompts A-J followed the order of the spectra that was presented to them and were typically in accuracies with higher complexity and number of spectra, as illustrated in Figure 4.

6 LIMITATIONS AND FUTURE WORK

Prompt A: WebSpectra problems and solutions have been available publicly since 2000. Two additional Prompts D, E in Supplementary used WebSpectra problems of lower difficulties, but did not yield correct answers, which likely indicates that LMMs did not memorize answers. Contextual understanding of the problem for three different GPT models was also observed (GPT-5, Organic Chemistry, ChatGPT-4o).

Prompt B: The Journal of Organic Chemistry supplementary spectra from Alajarin, M., et al. were published on August 19, 2024, several days before this manuscript was submitted. This means that the data which were not included in any of the previous LMM trainings. Prompt C: The Nowick, James S., University of California Irvine exam question did not have a solution, and was solved by the author of the manuscript, with some assistance by GPT-5. Given these findings, it is believed that LMMs primarily provided solutions based on generalizability based on prior training data, and were less likely to memorize, obtain solutions posted several days ago, or had other ways to solve unknown problems.

The author of the manuscript performed experiments, analysis, and wrote the manuscript, and primarily used ChatGPT-4o and GPT-5 as research tools. Future studies will focus on further fine-tuning GPT based models and incorporating new LMM updates, as multimodal trainings and data fusion are expected to continue to improve.

7 CONCLUSIONS

Large Multimodal Models continue to rapidly improve the ability for conversational artificial intelligence to process image applications, are readily available, and are easier to use than many current computer vision models. GPT-5 performance using document retrieval was realized, primarily in the accuracy of its reported chemical shift values vs. the four other models for Prompt A. Gemini 1.5 Pro lacked helpful responses for each of the three prompts, and is likely earlier in its development regarding the fields of chemometrics and spectroscopy.

Prompt C contained the greatest number of spectra or data at 9: EI-MS, IR, ¹H-NMR full, 2 ¹H-NMR zoom, Table, ¹³C-NMR, and 2 DEPT-NMR spectra seen in Figure 5. The prior Kawchak, K. study showed high performance using LMMs for

broad utility based on different NMR spectra, mass spectrometry, reaction recommendations, and enzyme interactions [7]. This study provided a greater challenge to LMMs by utilizing many pieces of information simultaneously to yield a correct chemical compound prediction combined with natural language processing. The results revealed current LMM technical limits, and builds upon the 2024 work of Park, N., et al that used similarity searches with LLMs and an OpenClip image model to analyze ¹H-NMR and ¹³C-NMR spectra [31].

Two GPT LMMs achieved 5/5 top scores answering Prompt A, which represents the first time any LMM has analyzed and provided generally accurate stepwise explanations for five spectra, and then combined this information to predict the correct chemical structure, to the author's best understanding. Achieving these types of results more consistently and with more input data will likely improve as a) data fusion of different data types improves, b) optical character recognition is enhanced, and c) context windows are further increased to make existing economical artificial intelligence software both better and faster.

REFERENCES

- [1] Kent et al. Gauen. Comparison of visual datasets for machine learning. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, page 346–355, August 2017. doi: 10.1109/IRI.2017.59. URL <https://ieeexplore.ieee.org/abstract/document/8102956>.
- [2] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32, 2019. ISSN 1533-7928. URL <http://jmlr.org/papers/v20/18-444.html>.
- [3] Kevin Kawchak. Extra large language models benchmarking for medicinal chemistry purpose., 2024. URL <https://github.com/kevinkawchak/Generative-AI-Drug-Synthesis/tree/2dbe2ff3cf0fdddc5c4adcec18d75c549fbbad4a/Code/Drug%20Discovery/XLLM>.
- [4] Kevin Kawchak. Large language models for early phase genai drug discovery large language model performance., 2024. URL <https://github.com/kevinkawchak/Generative-AI-Drug-Synthesis/tree/main/Code/Drug%20Discovery/LLM>.
- [5] Kevin Kawchak. Total synthesis guidance for chemists introduction total synthesis., 2024. URL <https://github.com/kevinkawchak/Generative-AI-Drug-Synthesis/tree/845db1372aeec5236bd5edf551a677a99df78949/Code/Drug%20Discovery/LMM>.
- [6] Kevin Kawchak. Llm-retrieval augmented generation for drug shortages 2024 large language models., 2024. URL <https://github.com/kevinkawchak/Generative-AI-Drug-Synthesis/tree/main/Code/Drug%20Discovery/LLM-RAG>.
- [7] Kevin Kawchak. Lmm chemical research with document retrieval. *chemRxiv*, 2024. URL <https://chemrxiv.org/engage/chemrxiv/article-details/66b88430c9c6a5c07ae26635>.

- [8] Xiaonan Chen, Weikang Shu, Liang Zhao, and Jingjing Wan. Advanced mass spectrometric and spectroscopic methods coupled with machine learning for in vitro diagnosis. *View*, 4(1):20220038, 2023. URL <https://onlinelibrary.wiley.com/doi/full/10.1002/VIW.20220038>.
- [9] Dongliang Kang and Ya-Pu Zhao. Predicting the molecular models, types, and maturity of kerogen in shale using machine learning and multi-nmr spectra. *Energy and Fuels*, 36(11):5749–5761, June 2022. ISSN 0887-0624, 1520-5029. doi: 10.1021/acs.energyfuels.2c00738. URL <https://pubs.acs.org/doi/10.1021/acs.energyfuels.2c00738>.
- [10] 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. *arXiv preprint arXiv:2408.08284*, 2024. URL <https://arxiv.org/abs/2408.08284>.
- [11] Thomas Specht, Justus Arweiler, Johannes Stuber, Kerstin Munnemann, Hans Hasse, and Fabian Jirasek. Automated nuclear magnetic resonance fingerprinting of mixtures. *Magnetic Resonance in Chemistry*, 62(4):286–297, 2024. URL <https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/full/10.1002/mrc.5381>.
- [12] 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. URL <https://pubs.rsc.org/en/content/articlehtml/2024/dd/d4dd00008k>.
- [13] Mestrelab Mnova. Mnova-structure-elucidation, 2024. URL <https://mestrelab.com/software/mnova-software/structure-elucidation/>.
- [14] ACD Labs, 2024. URL <https://www.acdlabs.com/products/spectrus-platform/structure-elucidator-suite>.
- [15] Stefan Kuchemann and et al. Are large multimodal foundation models all we need? on opportunities and challenges of these models in education. *ResearchGate*, 2023. URL https://www.researchgate.net/publication/377144957_Are_Large_Multimodal_Foundation_Models_all_we_need_On_Opportunities_and_Challenges_of_these_Models_in_Education.
- [16] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *arXiv*, na(arXiv:2309.14525), September 2023. doi: 10.48550/arXiv.2309.14525. URL <http://arxiv.org/abs/2309.14525>. arXiv:2309.14525 [cs].
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*, na(arXiv:2304.08485), December 2023. doi: 10.48550/arXiv.2304.08485. URL <http://arxiv.org/abs/2304.08485>. arXiv:2304.08485 [cs].
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, page 8748–8763. PMLR, July 2021. URL <https://proceedings.mlr.press/v139/radford21a.html>.
- [19] LMSYS, 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna>.
- [20] Dana Lahat, Tülay Adalı, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, September 2015. ISSN 1558-2256. doi: 10.1109/JPROC.2015.2460697. URL <https://ieeexplore.ieee.org/abstract/document/7214350>.
- [21] Chunyuan Li. Large multimodal models: Notes on cvpr 2023 tutorial. *arxiv*, na(arXiv:2306.14895), June 2023. doi: 10.48550/arXiv.2306.14895. URL <http://arxiv.org/abs/2306.14895>. arXiv:2306.14895 [cs].
- [22] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arxiv*, na(arXiv:2305.07895), August 2024. doi: 10.48550/arXiv.2305.07895. URL <http://arxiv.org/abs/2305.07895>. arXiv:2305.07895 [cs].
- [23] WebSpectra-Chem-UCLA, 2000. URL <https://webspectra.chem.ucla.edu/>.
- [24] Mateo Alajarin, Jose Cabrera, Delia Bautista, Pilar Sanchez-Andrada, and Aurelia Pastor. Cycloadditions of 4-alkenyl-2-aminothiazoles with nitroalkenes in the formal synthesis of pramipexole: An experimental and computational study. *The Journal of Organic Chemistry*, 0(0):null, 0. doi: 10.1021/acs.joc.4c00843. URL <https://doi.org/10.1021/acs.joc.4c00843>. PMID: 39158141.
- [25] James S. Nowick, 2014. URL <https://www.chem.uci.edu/~jsnowick/organicspectroscopy/index.html>.
- [26] ChatGPT. Introducing chatgpt for ios openai’s latest advancements at your fingertips., aug 2024. URL <https://apps.apple.com/us/app/chatgpt/id6448311069>.
- [27] ChatGPT, 2024. URL <https://chatgpt.com>.
- [28] Claude. Talk with claude, an ai assistant from anthropic. introducing project., 2024. URL <https://claude.ai/>.
- [29] Google, 2024. URL <https://gemini.google.com>.
- [30] NIST Office of Data and Informatics. Search results, 2023. URL <https://webbook.nist.gov/cgi/cbook.cgi?Formula=C9H7N&NoIon=on&Units=SI>.
- [31] Nathaniel H. Park, Tiffany J. Callahan, James L. Hedrick, Tim Erdmann, and Sara Capponi. Leveraging chemistry foundation models to facilitate structure focused retrieval augmented generation in multi-agent workflows for catalyst and materials design. *arxiv*, na(arXiv:2408.11793), August 2024. doi: 10.48550/arXiv.2408.11793. URL <http://arxiv.org/abs/2408.11793>. arXiv:2408.11793 [cs].