Automatic extraction of FAIR data from publications using LLM

Luc Patiny[a] and Guillaume Godin[b]

a) Zakodium Sàrl, Route d'Echandens 6b, 1027 Lonay, Switzerland.
b) DSM-Firmenich, Science and Research, Rue de la bergère 7, Satigny 1242, Switzerland.

Introduction

Since the beginning of modern science[1], researchers have used a specific format to communicate their findings in a standardized language[2] Such formats help to ensure that results can be replicated and published[3]. With the rise of digitalization, artificial intelligence has become increasingly important. When a prominent professor retired, their research group typically remains open for one or two years to allow students to complete their work. After that, the group is disbanded, and most of the materials are lost. This is why scientific literature is so critical[4] - it serves as a foundation of knowledge that can be used to create robust models. Those are the central principles of FAIR (Findable, Accessible, Interoperable, Reusable) data. The more data available, the better the models can be, particularly for neural networks. Large Language Models (LLMs) have proven to be incredibly effective at replicating human tasks, with some models[5] improving computer productivity by up to 40%[6]. This is a significant improvement that not only increases output but also leads to better results. By combining human and LLM assistance, we can produce higher-quality content and solve repetitive tasks that would otherwise take years to complete. Those generative AI assistants can follow instructions to transform and extrapolate existing text.

Our contribution outlines a method for automatically extracting experimental data of molecules from literature. Essentially by our prompt engineering, we demonstrate that this process can be made more cost-effective. Secondly, we use automated fact checking principles to ensure the original data quality as well as the data retrieval by LLM. Ultimately, our aim is to provide guidance for the publication of organic chemical experimental data to assist researchers and enhance FAIR data.

Methodology

In chemical synthesis papers, authors typically include an experimental section that reports on a new procedure for an existing or new product molecule. This section contains structured data such as the name of the molecule, which usually follows IUPAC[7] (International Union Of Pure and Applied Chemistry) nomenclature. There are various tools available, such as rule-based systems (like OPSIN[8]) or neural networks that can convert names to molecules[9]. Similarly, transformers[10] can be used to convert SMILES[11] (Simplified Molecular Input Line Entry Specification) to IUPAC[7] names. We have seen that this conversion is the most difficult part of the automation because IUPAC nomenclature is highly complex and not the best identifier of molecules[12,13].

Other important information that can be found in the experimental section includes the reaction yield, which is the conversion factor of the main reactant to the new product. Predicting this property is difficult[14]. Additionally, the form of the product (e.g. liquid or solid) and its color can be reported, as well as physical properties like melting and boiling points. These properties are often reported with conditions such as the pressure of the measurement and the units used.

Finally, the experimental section contains an analytical portion that typically includes data on infrared, nuclear magnetic resonance, and mass spectra. This analytical data is crucial for validating the synthesis procedure. While the experimental data is structured and can often be parsed by rule-based systems, the way in which it is reported is not consistent and varies depending on the authors and journals. This makes systematic extraction of the data a non-trivial task.

To address this issue, one of the authors worked with a student for several months on a proof of concept, but they did not achieve great success. Recognizing the potential of ChatGPT, the decision was made to give it a chance to automate this task.

We have observed that the experimental section of papers contains various types of structured data that differ slightly from paper to paper. To handle this variability, we have created a prompt that provides a flexible structure for the solution. This type of structure is commonly used in NoSQL databases and applications to pre-define the potential structure of an object to store in MongoDB[15]. ChatGPT requires clear instructions to follow during the conversion process. For example, if a concise response without a deep analysis of the question and answer is desired, one can request "no comment, no explanation."

By applying data template constraints to the prompt, ChatGPT provides a normalized data output that is easily parsable and pre-structured for databases. We conducted an ablation test by removing our data template constraints, and as a result, we obtained unstructured data (see Table A1 in the appendix), where the fields did not match, and the extracted data format was not consistent from one call to another.

However, one constraint of any method is the financial aspect. Considering the cost per 1000 tokens of OpenAI, it is necessary to fine-tune the prompt to maximize the chances that the question and response are short. Our initial idea was to work with a well-known JSON[16] data format, which is commonly used to store data. However, one drawback of this format is the redundancy of the "key" for a "value" series in the JSON object, making it suboptimal in terms of token size. While it is possible to optimize it, the format would be more difficult for ChatGPT to understand.

Given that the training set of ChatGPT includes the full human digital universe, which encompasses any digital data and known structure formats, it is important to use an established format such as JSON. Alternatively, YAML[17] is another useful format for structuring data that allows for flexibility in setting the presence or absence of a field with a "?". It also includes data types and hierarchical substructure similar to the JSON format, which is important for capturing NMR and MS data.

To create an automated process for extracting experimental data, an application and controlling method must be developed. The application takes the experimental data using basic rules, injects the prompt question before it, and sends the question to the assistant. The OpenAI API is then called using this prompt, and it returns the result as a YAML output. This YAML output can be easily parsed by the application to build a database containing all the extracted data per experimental section. This database is then checked by a human through the controlling view. The goal is to fine-tune the prompt to obtain satisfactory coverage of the data included.

To fine-tune the prompt, several metrics are defined to compare the efficiency of different prompts. This is similar to an author examination method[18]. With chemical data, the quality of the data obtained can be easily checked. For example, one can verify if the SMILES can be converted into a true molecule object with OpenChemLib[19,20] or RDKit[21]. As, SMILES is a chemical standard for Wikipedia[22,23]. The first step is to determine if the chemical name can be converted into a molecule, followed by verifying if the molecule corresponds to the number of atoms in the molecular formula (MF) if the MF was provided in the experimental data. These basic metrics help determine the overall efficiency on a subset of papers to determine the best prompt, taking into account the cost of API usage as a second criteria for fine-tuning.

It can be argued that using a human to extract experimental data is also expensive, and that this task cannot be performed by a single human but instead requires a team of humans. However, relying on human standardization can be very painful and likely subject to human errors during the process. This is where automation can be particularly helpful, as it can improve efficiency, reduce costs, and minimize the risk of errors. By using ChatGPT and other automated methods, we can extract experimental data more quickly and accurately, providing a more reliable and cost-effective solution.

Database

In this work, we evaluate our method on the open access Molecules[24] journal, specifically volumes 2 to 28. Only a portion of the articles are related to synthesis or report synthesis experiments, unfortunately there is no flag to select only those specific articles. In order to identify articles with an analytical experiment part, we have introduced a few rules to count the digit proportion included in the string higher than 30% in continuous segments of text, and the minimum and maximum length of the segment respectively from 300 to 2000. This filtration method has detected 39 k individual analytical experiment molecules. We have decided to take volume 21 as a random selection of those experiments to determine the optimal prompt to extract maximum experimental data with high accuracy using ChatGPT.

Discussion

YAML is a markup language that was created in 2001 to establish rules for what markup information can be included in a document and how to combine content to facilitate use by both humans and computer programs. This format, which was officially created in 2006, is related to other markup structure languages such as HTML[25]. Json, on the other hand, is another type of object storage for data that mimics the real object structure in JavaScript[26]. It was created in 2000 and standardized in 2013.

In Table 1, we can see that YAML is more compact than JSON.

Table 1

<table>
<tr><th>Format name</th><th>YAML</th><th>json</th></tr>
<tr><td>data</td><td>Origin:
   author: Dan Brown
   language: English
   publication-date: 2017-10-03
   pages: 461
   description: | When billionaire
researcher Edmond Kirsch is killed,
                  it is up to Robert
Langdon & Ambra Vidal to honor
                  his memory by making
public his findings concerning the
                  origin of human life and
its destiny.</td><td>{
  "Origin": {
    "author": "Dan Brown",
    "language": "English",
     "publication-date": "2017-10-03",
     "pages": 461,
     "description": "When billionaire
researcher Edmond Kirsch is killed,
                     it is up to Robert
Langdon and Ambra Vidal to honor
                     his memory by making
public his findings concerning
                     the origin of human
life and its destiny."
  }
}</td></tr>
<tr><td>Number of words*</td><td>46</td><td>49</td></tr>
</table>

Example of data storage using YAML and json format; source from https://www.geeksforgeeks.org/what-is-the-difference-between-yaml-and-json/ , *: without counting double quotes `"`.

We have observed that ChatGPT performs better with strong instructions and guidance. To provide this, we have created a template of data structure inspired by the YAML format, as shown in Table 2. This template provides a clear structure for the data to be extracted, which helps ChatGPT to better understand and respond to the prompt question. By using this template, we can improve the accuracy and efficiency of the automated extraction process.

Table 2: structuring data request and response

YAML prompt	YAML response	Json response

`{`
`  name: string,`
`  yield?: number,`
`  appearance?: string,`
`mp?: {`
`    low?: number`
`    high?: number`
`    units?: string`
`    otherInfo?: string`
`  }`
	`name:`
	`'2-Acetamido-3-(2-m-tolylthiazol-4-yl)propanoic acid (rac-3b)'`
	`yield: 65`
	`appearance: 'white solid'`
	`mp:`
	`  low: 160`
	`  high: 161`
	`  units: '°C'`
	`},`
		`{`
		`  "name": "2-Acetamido-3-(2-m-tolylthiazol-4-yl)propanoic acid (rac-3b)",`
		`  "yield": 65,`
		`  "appearance": "white solid",`
		`  "mp": {`
		`    "low": 160,`
		`    "high": 161,`
		`    "units": "°C"`
		`  },`

Since we do not always know what is in a given experimental section, we have added the question mark character "?" to provide more flexibility in interpreting our rules by the ChatGPT assistant. This allows for more variability in the data structure while still maintaining a clear and organized format. By using this approach, we can improve the ability of ChatGPT to accurately extract the desired data from the experimental section.

Table 3: Cost comparison between YAML and Json prompts

Tokens example*	YAML	Json	compact Json (without null)
Prompt token size	645	1081	804
Cost & Time factors	1	1.67	1.25

*: see: https://github.com/cheminfo/chatgpt-molecules-parser/blob/main/JSON.md for details

Table 4: Performance for 534 examples.

Total example*	Extractable	Molecule identification
534	512	384
100%	95.9%	72%

*: see XX for details

For a sample of 534 experimental molecules reported, we were able to extract and send 511 of them to the API using our automated extraction process. We generated 462k tokens for the prompt and produced 419k tokens for the completion response, resulting in a total of 882k tokens. The cost of this process was 0.0017 $ per 1k tokens.

Using this automated process, we were able to extract 384 molecules with relative spectrum data from literature sources at a cost of 1.53 $. Further details can be found in Table 5. This demonstrates the potential for cost-effective and efficient extraction of experimental data using our automated process.

Table 5: ChatGPT cost

Prompt	size (k toten)	$ cost (k token)	Total price
Input	462.4	0.0015	0.693
output	419.3	0.002	0.838
total	881.7		**1.53 $**

For 700k molecules, we estimate that we will need close to 1 million experimental reports based on the 72% conversion ratio. This means that we will require 1 million API calls at an approximate total cost of 3000 $ (aka 1000000 / 500 * 1.5 $ = 3000 $). The cost will be around 3750 $ with compact JSON and 5100 $ with classical JSON.

It's important to note that an increase in the number of tokens written will result in an increase in both cost and processing time. Therefore, we consider cost and time factors to be proportional, and for simplicity, we assume that they are identical. As the number of tokens increases, we will need to wait longer to receive the answer.

Overall, using ChatGPT for automated extraction is a cost-effective solution for extracting experimental data from a large number of molecules. While the cost and time factors are proportional, the cost of using ChatGPT is still significantly lower than the cost of relying on human labor for this task.

We estimate that it would take a trained human approximately 3 minutes to parse experimental data, assuming no parsing errors occur. With 700k extractions required, this would amount to 35k hours of work. The global envelope of 3000 $ provides a fixed cost of 0.1 $ per hour of work per human. This human cost is even lower than the cost of electricity, network, and computer costs. Therefore, it is clear that using ChatGPT for automated extraction is a much more cost-effective solution than relying on human labor for this task. Such a human labor process was done in the past to build ChEMBL[27,28], Pubchem[29], Coconut[30] and other similar scientific databases.

While there are limited open-source NMR and MS databases available for data mining, in 2023 last release MassBank[31] counted 15500 unique molecules for almost 100000 MS references after 17 years of open source collaboration, while for NMRShiftDB2[32] it is 53954 NMR references. Our method can extract a large dataset for a relatively cheap cost. Once the data is validated, it becomes a valuable resource for data mining. However, the lack of available databases for data mining can be a barrier to progress in the field of chemical data analysis. By making a large, open-source database available to researchers around the world, we can promote the field of chemical data analysis and mining legally, and potentially accelerate progress in the field.

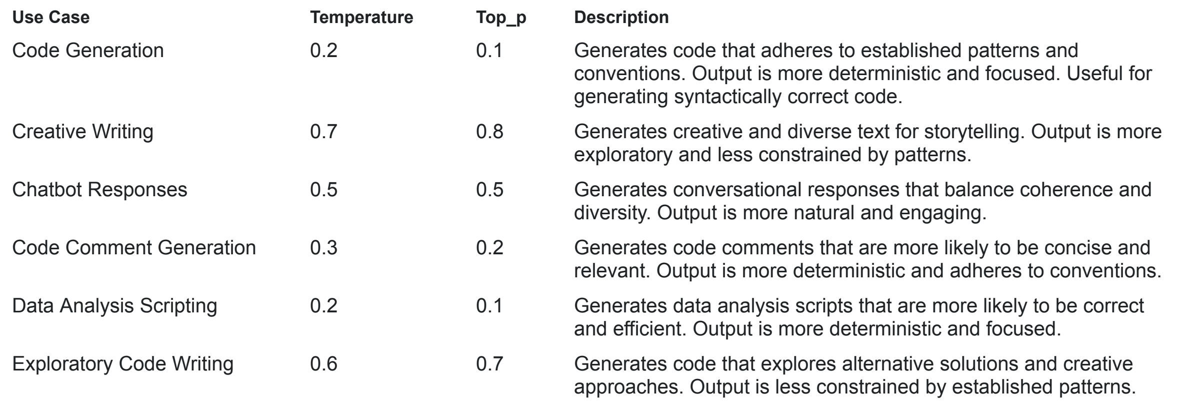
One important application of our method is that our pipeline is open-source. With your own OpenAI API key, you can extract data from your own papers and theses. We hope that you will contribute to our open-source database with your own research work in synthesis.

Furthermore, we believe that our tool can be very useful in the process of submitting an article that includes synthesis experimental data. This will allow chemists to check if the experimental section is fully parsable. While this process is not currently mandatory in the publication process by journals or in preprints, we believe that double-checking the digitalization of experimental data is necessary to preserve our scientific heritage. Our ultimate goal is to enforce data quality in publication and provide common guidance for experimental data publication. In this endeavor, we aim to provide an open-source tool to scientists everywhere, laying the groundwork for digitalization.

Fine tune ChatGPT parameters

There are two main parameters that can be useful to fine tune API calls in order to have less hallucinations and more accurate responses to specific tasks which are the temperature and the top_p sampling values. Default settings are for Creative writing. We have decided to use Data analysis scripting settings instead which allow us to be more deterministic and less creative in the generative process. The length of the prompt is only very important to avoid truncated results; we set it to 2000.

Table 7 : ChatGPT 3.5 parameters scenario*



Use Case	Temperature	Top_p	Description
Code Generation	0.2	0.1	Generates code that adheres to established patterns and conventions. Output is more deterministic and focused. Useful for generating syntactically correct code.
Creative Writing	0.7	0.8	Generates creative and diverse text for storytelling. Output is more exploratory and less constrained by patterns.
Chatbot Responses	0.5	0.5	Generates conversational responses that balance coherence and diversity. Output is more natural and engaging.
Code Comment Generation	0.3	0.2	Generates code comments that are more likely to be concise and relevant. Output is more deterministic and adheres to conventions.
Data Analysis Scripting	0.2	0.1	Generates data analysis scripts that are more likely to be correct and efficient. Output is more deterministic and focused.
Exploratory Code Writing	0.6	0.7	Generates code that explores alternative solutions and creative approaches. Output is less constrained by established patterns.

table 7: * https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api-a-few-tips-and-tricks-on-controlling-the-creativity-deterministic-output-of-prompt-responses/172683/1

Deeper analysis

After applying all our safety checks (see appendix for more details) to the extracted data, we found that the data retrieval score is only around 27%, while it is close to 70% only based on name to SMILES conversion. Interestingly, on those initial errors, we found missing NMR solvents in the original text, incorrect numbers of protons in 1H NMR, or elementary analysis without close matching between the expected and found results. We will come back to these errors at the end of this section. One recurrent observed error is duplicate or missing numbers between the original data and the extracted data. Generally, ChatGPT fixes human errors (Table 8), such as converting of a string like "172-4 °C" into "172-174 range in °C", LLM Fixes the original string value of "53,30" to "53.30" in an elementary analysis section. The MS peak intensity duplicates like originally "271/273 (100)" became

271 (100) and 273 (100) after extraction. Similarly, NMR delta shifts are reported as a number, but those shifts have been converted into a string, resulting in a new matching error. Another duplication can be an original "J1 = J2 = 7.5", which gives us a duplication alert as we got twice coupling values of 7.5. One interesting observed matching error is due to the replication of the same MS, HRMS, or "Mass" fields in the extracted data. We have seen one potential error in the LLM conversion, a few times, we have no frequency row data but in the NMR output has a 400 Frequency.

The criteria of number mismatch is too strict (see appendix table). Removing the number mismatch safety check, we have a data retrieval score of 39%. The rest of the safety criteria are more related to the control of the reported analytical results by the authors and not linked to ChatGPT extraction itself. We cannot consider this score as a metric of extraction quality. However, this criteria is very important as a data quality source for any additional data mining applications. We can provide additional information very important in the chase of outliers in order to answer questions like "Could this NMR outlier data be wrong?".

The errors in 1H-NMR are generally observed on a labile azote proton case, which is not always visible in the 1H-NMR analysis. By adjusting this condition, the extracted data reaches 39% based on safety checks.

We are not able to give an exact value of the error introduced by extraction, but random manual analysis showed that the data generated were in the original text in a very large majority of the cases, which meets our requirements as this was one of the main prompt constraints.

Table 8: Example of LLM corrections by data type

Data type	Raw data	LLM output	Good
MP	"172-4 °C"	lower: 172, higher:174, unit: °C	yes
Elemental analysis	"53,30"	53.30	yes
Mass	271/273 (100)	mass: 271, intensity: 100 mass: 273, intensity: 100	yes
1H NMR	"J1 = J2 = 7.5"	coupling: 7.5 coupling: 7.5	yes
Molecular formula	No	deduced it based on the name	yes
NMR frequence	No	400	no

Conclusion

Based on our analysis, the remaining major issue is molecule name parsing. Addressing this issue would be very impactful in systematically including SMILES or INCHI data in future publication guidance, additionally the tool can control and enhance data quality in literature.

While this article focuses on molecule properties extraction, the same idea can be applied to other digitalization applications for example in ChEMBL[27,28], PubChem[29] or Coconut[30].

Despite the variability and subtle format changes in each author or journal, we have demonstrated that we can define simple and comprehensible rules to help ChatGPT with the extraction of chemical experimental data from scientific papers. With optimal YAML instructions, we are able to reduce the cost and solve this challenge. We analyzed a subset and found that we can extract 72% of data by name using the ChatGPT 3.5 turbo API version. While it would take nearly one full day of work for humans, the extraction process takes almost 10-30 minutes for a cost of 1.53 $. Additionally, we were able to control the data quality (39%) and data coverage (95%). For 100 $, we get 1 year FTE (full time equivalent) human extraction work. Any chemist or scientist can use our open-source code to extract their own research paper and make the dataset available to the whole community.

We can easily extrapolate that we can fine-tune a much smaller open-source LLM to reproduce our work without additional cost on the API[33]. However, we must also consider that the OpenAI API provides hardware of computation included in the API token cost. Therefore, it will be a compromise between a free open-source LLM using dedicated cloud infrastructure at moderate cost versus OpenAI performance in an all-in-one API.

Finally, we must consider that the fine-tuning is based on one version of ChatGPT (3.5). As newer versions of ChatGPT are released, it may be necessary to adapt our approach to maintain optimal performance. We also expect that the current cost will decrease in the future as data, architecture, and hardware improve following Moore's Law[34]. As technology advances, we anticipate that our method will become even more cost-effective and efficient, making it an increasingly valuable tool for the scientific community.

Future

We may propose in future a fine tune version with an open LLM using this cleaner dataset and extending to a larger database. OpenAI have just announced a cheaper price for ChatGPT 4.0 turbo which is still around ten times more expensive than version 3.5 used in our research. We may expect a little improvement compared to v3.5 that needs to be confirmed.

Contribution

L. Patiny wrote codes, controlling webpage, metrics and ran experiments, G. Godin wrote the paper and suggested using ChatGPT, reviewed data, both designed the experiments, fine tuned prompt and defined metrics.

Reference

(1) Csiszar, A. *The Scientific Journal: Authorship and the Politics of Knowledge in the Nineteenth Century*; University of Chicago press: Chicago, 2018.

(2) Gordin, M. D. *Scientific Babel: How Science Was Done before and after Global English*; The University of Chicago Press: Chicago ; London, 2015.

(3) Scholz, F. Writing and Publishing a Scientific Paper. *ChemTexts* **2022**, *8* (1), 8. https://doi.org/10.1007/s40828-022-00160-7.

(4) Miyakawa, T. No Raw Data, No Science: Another Possible Source of the Reproducibility Crisis. *Mol. Brain* **2020**, *13* (1), 24, s13041-020-0552–2. https://doi.org/10.1186/s13041-020-0552-2.

(5) ChatGPT Version 3.5, 2023. https://openai.com/chatgpt.

(6) Noy, S.; Zhang, W. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* **2023**, *381* (6654), 187–192. https://doi.org/10.1126/science.adh2586.

(7) *The IUPAC Compendium of Chemical Terminology: The Gold Book*, 4th ed.; Gold, V., Ed.; International Union of Pure and Applied Chemistry (IUPAC): Research Triangle Park, NC, 2019. https://doi.org/10.1351/goldbook.

(8) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51* (3), 739–753. https://doi.org/10.1021/ci100384d.

(9) Krasnov, L.; Khokhlov, I.; Fedorov, M. V.; Sosnin, S. Transformer-Based Artificial Neural Networks for the Conversion between Chemical Notations. *Sci. Rep.* **2021**, *11* (1), 14798. https://doi.org/10.1038/s41598-021-94082-y.

(10) Rajan, K.; Zielesny, A.; Steinbeck, C. STOUT: SMILES to IUPAC Names Using Neural Machine Translation. *J. Cheminformatics* **2021**, *13* (1), 34. https://doi.org/10.1186/s13321-021-00512-4.

(11) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(12) Dashti, H.; Westler, W. M.; Markley, J. L.; Eghbalnia, H. R. Unique Identifiers for Small Molecules Enable Rigorous Labeling of Their Atoms. *Sci. Data* **2017**, *4* (1), 170073. https://doi.org/10.1038/sdata.2017.73.

(13) Dashti, H.; Wedell, J. R.; Westler, W. M.; Markley, J. L.; Eghbalnia, H. R. Automated Evaluation of Consistency within the PubChem Compound Database. *Sci. Data* **2019**, *6* (1), 190023. https://doi.org/10.1038/sdata.2019.23.

(14) Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. *When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges*; preprint; Chemistry, 2023. https://doi.org/10.26434/chemrxiv-2023-xdv03.

(15) Guo, R. MongoDB's JavaScript Fuzzer. *Commun. ACM* **2017**, *60* (5), 43–47. https://doi.org/10.1145/3052937.

(16) Crockford, D. The JSON Saga. *YouTube*; 2011.

(17) Clark, C. . E. YAML Draft 0.1. *Yahoo! Tech groups: sml-dev*; 2001.

(18) Van Deursen, R.; Ertl, P.; Tetko, I. V.; Godin, G. GEN: Highly Efficient SMILES Explorer Using Autodidactic Generative Examination Networks. *J. Cheminformatics* **2020**, *12* (1), 22. https://doi.org/10.1186/s13321-020-00425-8.

(19) Wahl, J.; Sander, T. Fully Automated Creation of Virtual Chemical Fragment Spaces Using the Open-Source Library OpenChemLib. *J. Chem. Inf. Model.* **2022**, *62* (9), 2202–2211. https://doi.org/10.1021/acs.jcim.1c01041.

(20) Michaël Zasso; cheminfo-bot; Luc Patiny; rufenerc; Greenkeeper; Andres Castillo; Daniel Kostro; Robert Fidler. Cheminfo/Openchemlib-Js: V8.7.1, 2023. https://doi.org/10.5281/ZENODO.10016820.

(21) Landrum, G.; Tosco, P.; Kelley, B.; Ric; Cosgrove, D.; Sriniker; Gedeck; Vianello, R.; NadineSchneider; Kawashima, E.; N, D.; Jones, G.; Dalke, A.; Cole, B.; Swain, M.; Turk, S.; AlexanderSavelyev; Vaucher, A.; Wójcikowski, M.; Ichiru Take; Probst, D.; Ujihara, K.; Scalfani, V. F.; Godin, G.; Lehtivarjo, J.; Pahl, A.; Walker, R.; Francois Berenger; Jasondbiggs; Strets123. Rdkit/Rdkit: 2023_09_1 (Q3 2023) Release Beta, 2023. https://doi.org/10.5281/ZENODO.591637.

(22) Wikipedia, contributors. Wikipedia, The Free Encyclopedia. 2017.

(23) Ertl, P.; Patiny, L.; Sander, T.; Rufener, C.; Zasso, M. Wikipedia Chemical Structure Explorer: Substructure and Similarity Searching of Molecules from Wikipedia. *J. Cheminformatics* **2015**, *7* (1), 10. https://doi.org/10.1186/s13321-015-0061-y.

(24) MDPI. Molecules. https://www.mdpi.com/journal/molecules.

(25) Berners-Lee, T. HTML. *W3.org*; 1989.

(26) Powell, T. A.; Schneider, F. *JavaScript: The Complete Reference*; McGraw-Hill: New York, 2001.

(27) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(28) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(29) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51* (D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956.

(30) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminformatics* **2021**, *13* (1), 2. https://doi.org/10.1186/s13321-020-00478-9.

(31) Schulze, T.; Meier, R.; Alygizakis, N.; Schymanski, E.; Bach, E.; D.H. LI; Lauperbe; Raalizadeh; BenildeB; Tanaka, S.; Witting, M.; Bowen, T.; Uly55e5; Treutler, H.; Trchutka; Kohlhoff; Meowcat. MassBank/MassBank-Data: Release Version 2023.09, 2023. https://doi.org/10.5281/ZENODO.3378723.

(32) Kuhn, S. NMRShiftDB2. https://nmrshiftdb.nmr.uni-koeln.de/.

(33) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. *Is GPT-3 All You Need for Low-Data Discovery in Chemistry?*; preprint; Chemistry, 2023. https://doi.org/10.26434/chemrxiv-2023-fw8n4.

(34) Moore, G., E. Cramming More Components onto Integrated Circuits. *Electronics*. January 1, 1965.

Source code: https://github.com/cheminfo/chatgpt-molecules-parser

Parser & controlling view: https://experium.cheminfo.org/

Appendix:

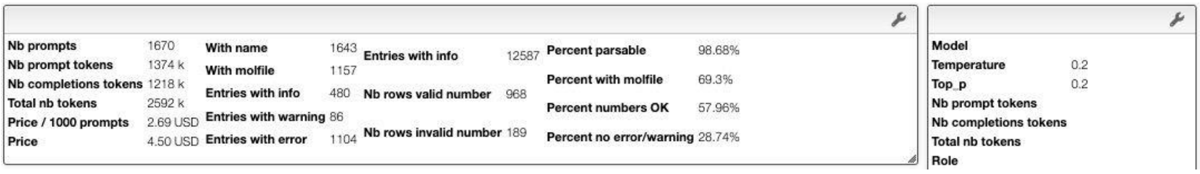
Table A1: semi structured output with minimal prompt output

input minimal prompt	This is a scientific experimental part describing a chemical molecule and we need to parse all the available properties. This must be very accurate and no data may be invented. 4-Methoxycinnamyl 6-O-(α-L-arabinopyranosyl)-β-D-glucopyranoside (22). Prepared according to synthetic method described for the preparation of 20 from 19b (100 mg, 0.156 mmol). Yield: 71 mg (81%). m.p. 93–95 °C; [α]25D−40.2 (c 1.0, MeOH). 1H-NMR (600 MHz, MeOH-d4,): δ 7.37 (d, 2H, J = 8.3 Hz), 6.83 (d, 2H, J = 8.3 Hz), 6.60 (d, 1H, J = 15.9 Hz), 6.20 (td, 1H, J = 6.8, 15.9 Hz), 4.48 (dd, 1H, J = 5.8, 12.6 Hz), 4.36 (d, 1H, J = 7.8 Hz, H-1′), 4.33 (d, 1H, J = 6.8 Hz, H-1″), 4.30 (dd, 1H, J = 7.8, 12.6 Hz), 3.88 (dd, 1H, J = 3.0, 12.4 Hz), 3.82-3.78 (m, 1H), 3.78 (s, 3H), 3.74 (dd, 1H, J = 5.8, 11.4 Hz), 3.63 (t, 1H, J = 6.8 Hz), 3.55–3.51 (m, 2H), 3.46–3.40 (m, 1H), 3.37–3.33 (m, 3H), 3.25–3.20 (m, 1H); 13C-NMR (75 MHz, MeOH-d4): δ 160.9 (Ar-C), 133.7 (=CH), 130.9 (Ar-C), 128.8 (2C, 2Ar-C), 124.3 (=CH), 115.0 (2C, 2Ar-C), 105.1 (C-1″), 103.2 (C-1′), 78.0 (C-3′), 76.9 (C-5′), 75.1 (C-2′), 74.2 (C-3″), 72.4 (C-2″), 71.7 (C-4′), 71.1 (-OCH2-), 69.5 (C-6′), 69.4 (C-4″), 66.7 (C-5″), 55.7 (OCH3); HRMS (ESI-TOF): C22H31O13 (M+HCOO)– 503.1765; found 503.1773. Properties can be expanded and empty properties may not be present.	This is a scientific experimental part describing a chemical molecule and we need to parse all the available properties. This must be very accurate and no data may be invented. 2-Acetamido-3-(2-m-tolylthiazol-4-yl)propanoic acid (rac-3b): Yield: 65%; white solid; m.p. 160–161 °C; 1H-NMR (600 MHz, DMSO) δ 7.97 (1H, NH), 7.73 (s, 1H), 7.68 (d, J = 7.7 Hz, 1H), 7.36 (t, J = 7.6 Hz, 1H), 7.30 (s, 1H), 7.26 (d, J = 7.4 Hz, 1H), 4.43 (m, 1H), 3.13 (ddd, J = 23.2, 14.6, 6.6 Hz, 2H), 2.37 (s, 3H), 1.79 (s, 3H). 13C-NMR (151 MHz, DMSO) δ 173.32, 168.80, 165.97, 154.70, 138.51, 133.24, 130.66, 129.08, 126.42, 123.28, 115.31, 52.99, 33.75, 22.61, 20.91; ESI+-MS: 305.0960 (calculated: 305.0954 for C15H16N2O3S [M + H]+); m/z (%): 327 (85.3, [M + Na]+), 307 (4.4, [M + 3H]+), 306 (16.5, [M + 2H]+), 305 (100, [M + H]+), 292 (3.0), 291 (21.2), 288 (4.2), 284 (1.7), 263 (30.4), 251 (21.8), 210 (4.8). Properties can be expanded and empty properties may not be present.

semi structured output	Here are the properties of the chemical molecule 4-Methoxycinnamyl 6-O-(α-L-arabinopyranosyl)-β-D-glucopyranoside (22) based on the provided experimental data: 1. Molecular Formula: C22H31O13 2. Molecular Weight: 503.1765 g/mol (as determined by HRMS) 3. Synthesis: - Prepared according to the synthetic method described for the preparation of compound 20 from compound 19b. - Starting material: 100 mg of compound 19b (0.156 mmol). - Yield: 71 mg (81%) 4. Physical Properties: - Melting Point (m.p.): 93–95 °C - Specific Rotation ([α]25D): -40.2 (c 1.0, MeOH) 5. NMR Spectroscopy (1H-NMR, 600 MHz, MeOH-d4): - Chemical shifts (δ) and multiplicity: - 7.37 (d, 2H, J = 8.3 Hz) - 6.83 (d, 2H, J = 8.3 Hz) - 6.60 (d, 1H, J = 15.9 Hz) - 6.20 (td, 1H, J = 6.8, 15.9 Hz) - 4.48 (dd, 1H, J = 5.8, 12.6 Hz) - 4.36 (d, 1H, J = 7.8 Hz, H-1′) - 4.33 (d, 1H, J = 6.8 Hz, H-1″) - 4.30 (dd, 1H, J = 7.8, 12.6 Hz) - 3.88 (dd, 1H, J = 3.0, 12.4 Hz) - 3.82-3.78 (m, 1H) - 3.78 (s, 3H) - 3.74 (dd, 1H, J = 5.8, 11.4 Hz) - 3.63 (t, 1H, J = 6.8 Hz) - 3.55–3.51 (m, 2H) - 3.46–3.40 (m, 1H) - 3.37–3.33 (m, 3H) - 3.25–3.20 (m, 1H) 6. NMR Spectroscopy (13C-NMR, 75 MHz, MeOH-d4): - Chemical shifts (δ): - 160.9 (Ar-C) - 133.7 (=CH) - 130.9 (Ar-C) - 128.8 (2C, 2Ar-C) - 124.3 (=CH) - 115.0 (2C, 2Ar-C) - 105.1 (C-1″) - 103.2 (C-1′) - 78.0 (C-3′) - 76.9 (C-5′) - 75.1 (C-2′) - 74.2 (C-3″) - 72.4 (C-2″) - 71.7 (C-4′) - 71.1 (-OCH2-) - 69.5 (C-6′) - 69.4 (C-4″) - 66.7 (C-5″) - 55.7 (OCH3) 7. High-Resolution Mass Spectrometry (HRMS, ESI-TOF): - Accurate mass of the molecule (M+HCOO): 503.1765 g/mol - Observed mass: 503.1773 g/mol Please note that these properties are based on the provided experimental data, and no additional information has been invented.	Here are the properties of the chemical molecule 2-Acetamido-3-(2-m-tolylthiazol-4-yl)propanoic acid (rac-3b) based on the provided information: 1. Chemical Name: 2-Acetamido-3-(2-m-tolylthiazol-4-yl)propanoic acid (rac-3b) 2. Yield: 65% 3. Physical State: White solid 4. Melting Point (m.p.): 160–161 °C 5. 1H-NMR (600 MHz, DMSO) data: - δ 7.97 (1H, NH) - δ 7.73 (s, 1H) - δ 7.68 (d, J = 7.7 Hz, 1H) - δ 7.36 (t, J = 7.6 Hz, 1H) - δ 7.30 (s, 1H) - δ 7.26 (d, J = 7.4 Hz, 1H) - δ 4.43 (m, 1H) - δ 3.13 (ddd, J = 23.2, 14.6, 6.6 Hz, 2H) - δ 2.37 (s, 3H) - δ 1.79 (s, 3H) 6. 13C-NMR (151 MHz, DMSO) data: - δ 173.32 - δ 168.80 - δ 165.97 - δ 154.70 - δ 138.51 - δ 133.24 - δ 130.66 - δ 129.08 - δ 126.42 - δ 123.28 - δ 115.31 - δ 52.99 - δ 33.75 - δ 22.61 - δ 20.91 7. Electrospray Ionization Mass Spectrometry (ESI+-MS) data: - Observed [M + H]+ m/z: 305.0960 - Calculated [M + H]+ m/z: 305.0954 for C15H16N2O3S - Additional m/z values and their percentages: - 327 (85.3, [M + Na]+) - 307 (4.4, [M + 3H]+) - 306 (16.5, [M + 2H]+) - 305 (100, [M + H]+) - 292 (3.0) - 291 (21.2) - 288 (4.2) - 284 (1.7) - 263 (30.4) - 251 (21.8) - 210 (4.8) Please note that these properties are based on the provided experimental data, and no additional information has been invented.

Prompt Engineering analysis:

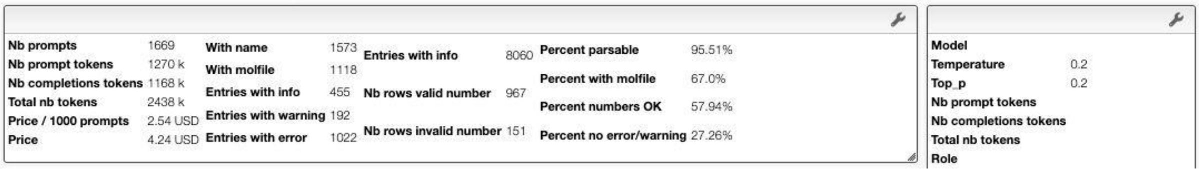
Volume 10 to 13 results with our Prompt



Nb prompts	1670	With name	1643	Entries with info	12587	Percent parsable	98.68%
Nb prompt tokens	1374 k	With molfile	1157			Percent with molfile	69.3%
Nb completions tokens	1218 k	Entries with info	480	Nb rows valid number	968		
Total nb tokens	2592 k					Percent numbers OK	57.96%
Price / 1000 prompts	2.69 USD	Entries with warning	86				
Price	4.50 USD	Entries with error	1104	Nb rows invalid number	189	Percent no error/warning	28.74%

Model	
Temperature	0.2
Top_p	0.2
Nb prompt tokens	
Nb completions tokens	
Total nb tokens	
Role	

Ablation study: removing the “quoted” string criteria



Nb prompts	1669	With name	1573	Entries with info	8060	Percent parsable	95.51%
Nb prompt tokens	1270 k	With molfile	1118			Percent with molfile	67.0%
Nb completions tokens	1168 k	Entries with info	455	Nb rows valid number	967		
Total nb tokens	2438 k					Percent numbers OK	57.94%
Price / 1000 prompts	2.54 USD	Entries with warning	192				
Price	4.24 USD	Entries with error	1022	Nb rows invalid number	151	Percent no error/warning	27.26%

Model	
Temperature	0.2
Top_p	0.2
Nb prompt tokens	
Nb completions tokens	
Total nb tokens	
Role	

We clearly see a degradation of the parsing performances as well as the name to smiles conversion. We cannot remove this constraint.

Reliable fact checking criterias:

Type	Matching	Question
NMR	Proton	Is number of protons are in matching the molecular formula ?
All	Numbers	Is there is created numbers in the answer ?
Mass	HRMS	Is the HRMS accurate versus the given molecule ?
Elemental analysis	MF	is the element analysis accurate versus the given molecule ?
NMR	Solvent	Is solvent is given ?