Benchmarking Agentic Systems in Automated Scientific Information Extraction with ChemX

Anastasia Vepreva[1] **Julia Razlivina**[1]
Maria Eremeeva[1] **Nina Gubina**[1] **Anastasia Orlova**[1] **Aleksei Dmitrenko**[1]
Ksenya Kapranova[1] **Susan Jyakhwo**[1] **Nikita Vasilev**[1] **Arsen Sarkisyan**[1]
Ivan Yu. Chernyshov[1] **Vladimir Vinogradov**[1] **Andrei Dmitrenko**[1,2]

[1]Center for AI in Chemistry, ITMO University, St. Petersburg, Russia
[2]D ONE AG, Zurich, Switzerland

dmitrenko@pish.itmo.ru

Abstract

The emergence of agent-based systems represents a significant advancement in artificial intelligence, with growing applications in automated data extraction. However, chemical information extraction remains a formidable challenge due to the inherent heterogeneity of chemical data. Current agent-based approaches, both general-purpose and domain-specific, exhibit limited performance in this domain. To address this gap, we present ChemX, a comprehensive collection of 10 manually curated and domain-expert-validated datasets focusing on nanomaterials and small molecules. These datasets are designed to rigorously evaluate and enhance automated extraction methodologies in chemistry. To demonstrate their utility, we conduct an extensive benchmarking study comparing existing state-of-the-art agentic systems such as ChatGPT Agent and chemical-specific data extraction agents. Additionally, we introduce our own single-agent approach that enables precise control over document preprocessing prior to extraction. We further evaluate the performance of modern baselines, such as GPT-5 and GPT-5 Thinking, to compare their capabilities with agentic approaches. Our empirical findings reveal persistent challenges in chemical information extraction, particularly in processing domain-specific terminology, complex tabular and schematic representations, and context-dependent ambiguities. The ChemX benchmark serves as a critical resource for advancing automated information extraction in chemistry, challenging the generalization capabilities of existing methods, and providing valuable insights into effective evaluation strategies.

1 Introduction

Over the past decade, machine learning has significantly advanced chemical discovery, underscoring the need for well-structured data [1, 2, 3]. Standardized datasets provide essential metrics for comparing algorithms, identifying their limitations, and accelerating progress [4, 5, 6, 7, 8]. However, major gaps persist, particularly in specialized domains, creating an urgent need for robust systems to automatically extract and curate chemical data from diverse sources.

While conventional NLP methods have been used for named entity recognition in the sciences [9, 10, 11], they remain limited in the broader range of tasks required for a chemical data extraction tool. Recent advances in large language models (LLMs) have demonstrated remarkable improvements

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat.

in contextual understanding and reasoning [12]. Autonomous multi-agent systems are becoming a new frontier in the automation of scientific research [13, 14]. Recent advances in automated chemical information extraction have increasingly leveraged agentic AI approaches, which employ autonomous, goal-directed agents capable of reasoning, planning, and executing complex workflows [15, 16, 17]. These agentic systems differ fundamentally from traditional AI methods by integrating domain-specific knowledge with capabilities for contextual understanding and iterative decision-making. Currently, highly specialized systems exist for data extraction in materials science, as well as for the extraction of organic reaction data or deep eutectic solvent knowledge [18, 19, 20, 21, 22, 23, 24]. Applying multi-agent systems to chemical data extraction remains challenging due to domain adaptation, making it an essential research challenge. To support it, we present ChemX, a manually curated multimodal benchmark dataset aimed at extracting chemical features from textual and visual content across diverse chemical domains. By capturing the heterogeneity and interconnectedness of real-world chemical literature, ChemX provides a foundation for evaluating automation extraction systems. This work makes two major contributions:

- We provide the ChemX benchmark, a collection of 10 curated datasets describing various properties of nanomaterials and small molecules. Each dataset is accompanied with detailed documentation, standardized metadata, and cross-verification by domain experts. The datasets are hosted as a collection on the Hugging Face. The accompanying documentation will be provided separately to ensure compliance with anonymization guidelines.
- We present a systematic evaluation of state-of-the-art agentic systems in the task of automated information extraction from domain-specific scientific literature. The code for the extraction experiments is provided in the https://ai-chem.github.io/ChemX.

2 Related works

Recent years have seen a growing ecosystem of chemical science benchmarks, many focusing on machine learning for property prediction, structural analysis, or vision-language tasks [25, 26, 27]. However, these are not designed for evaluating automated information extraction systems. The closest related study, nanoMINER [22], demonstrates structured extraction but is limited to one dataset related to nanozymes. We address this gap with 10 diverse datasets, benchmarking modern LLMs and agentic systems, including nanoMINER for comparison.

3 ChemX

ChemX is a comprehensive multimodal benchmark comprising 10 rigorously validated datasets spanning two major chemical domains: nanomaterials and small molecules (Figure 1). The collection is designed to support robust automated information extraction across heterogeneous data types, including tables, graphs, and unstructured text.

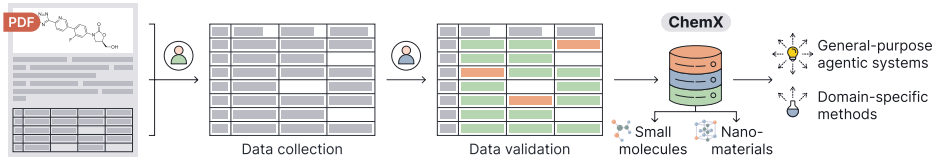

Figure 1: ChemX. This pipeline includes manual collection of multimodal data from scientific articles, further validation by domain experts and benchmarking automated data extraction.

The datasets' ontology varies between domains:

- Small molecule datasets focus on molecular descriptors like SMILES representations, biological activity metrics (MIC, IC50), and compound metadata
- Nanomaterial datasets encompass a broader range of parameters, including physicochemical properties, synthesis conditions, structural characteristics, and application-specific outcomes

The more detailed description of each dataset, quality control process and dataset analysis are presented in Appendix (Appendix A, Appendix B, Appendix C). Including the datasets of varying sizes and complexity in both domains creates a balanced and practical benchmark for automated information extraction.

All datasets were labeled by complexity level, which is described in detail in the subsection A.1.

4 Experiments

4.1 Information extraction task

This study was designed to evaluate modern agentic information extraction approaches using datasets from ChemX. We selected two datasets of the lowest complexity within the domain, as categorized in Table 4, namely, nanozymes (nanomaterials) and chelate complexes (small molecules). Appendix C demonstrates that closed-access articles constitute the vast majority within each dataset. To ensure the selection was both representative and reproducible, we included two open-access articles for analysis (subsection D.1). An end-to-end information extraction task is, therefore, defined as follows: given the article file (or DOI, in case an attachment is not supported), output the extracted information according to the instructions in the prompt.

4.2 Methods and metrics

A detailed description of the prompts and metrics used to evaluate the quality of extraction is described in the subsection D.2. The latest models GPT-5 and GPT-5 Thinking were selected as baselines. Agent-based approaches were also implemented, encompassing both a general-purpose ChatGPT Agent and domain-specific systems optimized for data extraction in singular and multiple domains such as FutureHouse [28], SLM-Matrix [18], Eunomia Agent [19], ChemOpenIE [23], and nanoMINER systems.

4.3 Single-agent approach

To address OpenAI's opaque PDF/screenshot processing, which risks inconsistent extractions, we develop a single-agent approach for structured text conversion, ensuring reproducibility and semantic integrity. We define our single-agent pipeline as an autonomous, adaptive preprocessing and extraction module that performs structured text conversion, vision captioning, and iterative quality checks without manual intervention. While conceptually simpler than multi-agent orchestration systems, it behaves agentically by controlling tools and reasoning steps toward a defined extraction goal. Using marker-pdf SDK [29], we extract text blocks, tables, and images, preserving document structure. Text and tables are converted to markdown, while images are replaced with local paths. Extracted images are processed by GPT-4o (2024-11-20) to generate descriptions, inserted into markdown at original locations via `<DESCRIPTION_FROM_IMAGE>` tags. The final markdown file is then processed by GPT-4.1, GPT-5, and GPT-OSS-20b for extraction, with results consolidated into dataset-specific CSV files.

5 Results and Discussion

As presented in Table 1, which details the average extraction metrics across all dataset columns, the general methods demonstrated superior performance for both nanomaterial and small molecule datasets. A notable exception is the nanoMINER method, which achieved the highest metrics; however, its applicability is severely limited by its specificity to a single dataset. Contrary to expectations, the GPT-5 Thinking model configured for extended reasoning demonstrates inferior performance on the extraction task compared to standard GPT-5.

Domain-specific multi-agent systems like SLM Matrix and FutureHouse were inadequate for the extraction task. General methods performed better on nanomaterial data (subsection A.1, Table 4), likely because all systems failed to extract SMILES from molecular images. This limitation may systematically underestimate small molecule metrics; detailed per-column metrics are provided in Appendix E.

Table 1: Extraction metrics. The current evaluation covers two low-complexity datasets to establish baseline feasibility. * ChatGPT Agent fails to complete the extraction task for the nanozymes dataset due to alleged policy violations. ** NanoMINER was originally designed to work with the nanozymes dataset only and cannot generalize.

Method	Nanozymes			Complexes		
	Precision	Recall	F1	Precision	Recall	F1
GPT-5	0.33	0.53	0.37	0.45	0.18	0.23
GPT-5 Thinking	0.01	0.04	0.02	0.22	0.18	0.19
Single-agent (GPT-4.1)	0.41	0.73	0.52	0.35	0.21	0.27
Single-agent (GPT-5)	0.47	**0.75**	0.58	0.32	0.39	0.35
Single-agent (GPT-OSS)	0.56	0.67	0.61	0.36	0.31	0.33
ChatGPT Agent*	-	-	-	**0.50**	**0.42**	**0.46**
SLM-Matrix	0.14	0.55	0.22	0.40	0.38	0.39
FutureHouse	0.05	0.31	0.09	0.12	0.06	0.06
NanoMINER**	**0.90**	0.74	**0.80**	-	-	-

Further analysis showed that the single-agent approach outperformed baselines. Pre-processing documents into structured text notably improved GPT-5 recall from 0.53 to 0.75. ChatGPT Agent, while achieving the best small molecule results, issued terms-of-use warnings on the nanozyme dataset.

Table 2: Agentic extraction systems overview.

Method	PDF file	Output format	Generalizability	End-to-end	Multimodality
Single-agent (ours)	yes	yes	yes	yes	yes
ChatGPT Agent	yes	**no**	**no**	yes	yes
SLM-Matrix	yes	yes	yes	yes	yes
NanoMINER	yes	yes	**no**	yes	yes
FutureHouse	**no**	yes	yes	yes	yes
Eunomia	yes	**no**	**no**	**no**	yes
OpenChemIE	yes	yes	**no**	**no**	yes

The pronounced methodological disparities among specialized approaches added complexity. We qualitatively evaluated these methods based on their ability to process PDFs, follow the required output structure, support multimodality, generalize across domains, and perform full data extraction. Methods unable to complete the full extraction task were excluded (Table 2). For example, OpenChemIE was omitted for extracting only molecular IDs and SMILES, while Eunomia was excluded for failing to produce a valid output structure.

Our findings show that accurate chemical information extraction remains complex despite AI advances. As multi-agent frameworks become more common, research should focus on agent orchestration. ChemX, as the first resource of its kind, provides a foundation for systematically evaluating and improving extraction techniques.

6 Conclusion

ChemX is an expert-curated, multimodal benchmark for chemical information extraction, addressing gaps in existing resources through standardized schemas, domain diversity, and provenance metadata. Its utility was demonstrated by evaluating state-of-the-art agentic systems compared against the leading reasoning LLMs. As the first benchmarking resource of its kind, ChemX provides a critical foundation for advancing automated information extraction in chemistry. By offering rigorously validated datasets, it enables systematic evaluation and refinement of emerging techniques, ultimately driving the progress in chemical information extraction.

7 Acknowledgment

The work was financially supported by the Priority 2030 Federal Academic Leadership Program.

We sincerely thank Olga Kononova for constructive feedback and fruitful discussions that helped us improve the manuscript.

References

[1] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, July 2018.

[2] Seyed Mohamad Moosavi, Kevin Maik Jablonka, and Berend Smit. The role of machine learning in the understanding and design of materials. *Journal of the American Chemical Society*, 142(48):20273–20287, November 2020.

[3] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, July 2018.

[4] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, September 2011.

[5] Colin R. Groom, Ian J. Bruno, Matthew P. Lightfoot, and Suzanna C. Ward. The cambridge structural database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials*, 72(2):171–179, April 2016.

[6] Stephen K. Burley, Helen M. Berman, Gerard J. Kleywegt, John L. Markley, Haruki Nakamura, and Sameer Velankar. *Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive*, page 627–641. Springer New York, 2017.

[7] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

[8] Stefan Ganscha, Oliver T. Unke, Daniel Ahlin, Hartmut Maennel, Sergii Kashubin, and Klaus-Robert Müller. The qcml dataset, quantum chemistry reference data from 33.5m dft and 14.7b semi-empirical calculations. *Scientific Data*, 12(1), March 2025.

[9] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[10] Jiahui Yu, Chengwei Zhang, Yingying Cheng, Yun-Fang Yang, Yuan-Bin She, Fengfan Liu, Weike Su, and An Su. Solvbert for solvation free energy and solubility prediction: a demonstration of an nlp model for predicting the properties of molecular complexes. July 2022.

[11] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 429–436. ACM, September 2019.

[12] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

[13] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang, Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. Magentic-one: A generalist multi-agent system for solving complex tasks, 2024.

[14] Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. Clinicalrag: Enhancing clinical decision support through heterogeneous knowledge retrieval. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, page 64–68. Association for Computational Linguistics, 2024.

[15] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, May 2024.

[16] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023.

[17] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research, 2023.

[18] Xin Li, Zhixuan Huang, Shu Quan, Cheng Peng, and Xiaoming Ma. Slm-matrix: a multi-agent trajectory reasoning and verification framework for enhancing language models in materials data extraction. *npj Computational Materials*, 11(1), July 2025.

[19] Mehrad Ansari and Seyed Mohamad Moosavi. Agent-based learning of materials datasets from the scientific literature. *Digital Discovery*, 3(12):2607–2617, 2024.

[20] Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. Llamp: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation, 2024.

[21] Ankan Mullick, Akash Ghosh, G. Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659, January 2024.

[22] R. Odobesku, K. Romanova, S. Mirzaeva, O. Zagorulko, R. Sim, R. Khakimullin, J. Razlivina, A. Dmitrenko, and V. Vinogradov. Agent-based multimodal information extraction for nanomaterials. *npj Computational Materials*, 11(1), June 2025.

[23] Vincent Fan, Yujie Qian, Alex Wang, Amber Wang, Connor W. Coley, and Regina Barzilay. Openchemie: An information extraction toolkit for chemistry literature, 2024.

[24] Xiting Peng, Yi Shen Tew, Kai Zhao, Chi Wang, Ren'ai Li, Shanying Hu, and Xiaonan Wang. Unlocking deep eutectic solvent knowledge through a large language model-driven framework and an interactive ai agent. *Green Chemical Engineering*, June 2025.

[25] Kuzma Khrabrov, Anton Ber, Artem Tsypin, Konstantin Ushenin, Egor Rumiantsev, Alexander Telepov, Dmitry Protasov, Ilya Shenbin, Anton Alekseev, Mikhail Shirokikh, Sergey Nikolenko, Elena Tutubalina, and Artur Kadurin. ∇^2dft: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials, 2024.

[26] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, and Honglak Lee. Cede: A collection of expert-curated datasets with atom-level entity annotations for optical chemical structure recognition. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27114–27126. Curran Associates, Inc., 2022.

[27] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 125780–125808. Curran Associates, Inc., 2024.

[28] FutureHouse — futurehouse.org. `https://www.futurehouse.org/`. [Accessed 09-05-2025].

[29] GitHub - VikParuchuri/marker: Convert PDF to markdown + JSON quickly with high accuracy — github.com. `https://github.com/VikParuchuri/marker`. [Accessed 09-05-2025].

[30] Michał P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15:1569, 2024.

[31] Kausik Hira, Mohd Zaki, Dhruvil Sheth, Mausam, and N. M. Anoop Krishnan. Reconstructing the materials tetrahedron: challenges in materials information extraction. *Digital Discovery*, 3(5):1021–1037, 2024.

[32] Kohulan Rajan, Henning Otto Brinkhaus, M. Isabel Agea, Achim Zielesny, and Christoph Steinbeck. Decimer.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications*, 14(1), August 2023.

A ChemX ontology

Table 3: ChemX benchmark datasets grouped by domain.

Domain	Dataset	Size	Features		Description
			String	Numeric	
Nano-materials	Cytotox	5535	12	9	Cytotoxicity of nanoparticles in normal and cancer cell lines.
	Seltox	3286	9	14	Toxic effects of nanoparticles on bacterial strains.
	Synergy	3326	10	19	Drug–nanoparticle synergy in antibacterial assays.
	Nanozymes	1135	9	11	Catalytic properties of inorganic enzyme mimics.
	Nanomag	2578	8	16	Magnetic nanomaterials and their biomedical uses.
Small molecules	Benzimidazoles	1721	6	1	SMILES molecules with MICs for antibiotic SAR studies.
	Oxazolidinones	2923	6	1	Synthetic antibiotics with biological activity data.
	Complexes	907	4	1	Organometallic chelate complexes with thermodynamic parameters.
	Eye Drops	163	2	1	Drug permeability data across corneal tissue.
	Co-crystals	70	7	0	Drug co-crystals with improved photostability.

For small molecule datasets, the ontology centers around molecular descriptors, including SMILES representations, biological activity metrics (e.g., MIC, IC_{50}), and compound-specific metadata. In contrast, nanomaterials and other material-centric datasets involve a substantially broader set of parameters, encompassing physicochemical properties (e.g., size, zeta potential, surface coating), synthesis conditions, structural characteristics, and application-specific outcomes. This reflects the inherent complexity and multimodality of material-related information in scientific literature.

A.1 Labeling datasets by complexity level for extraction

Table 4: Selection of articles for analysis.

Domain	Dataset	Complexity
Nanomaterials	Cytotoxicity	High
	Seltox	High
	Synergy	High
	Nanozymes	Medium
	Nanomag	High
Small molecules	Benzimidazoles	Medium
	Oxazolidinones	Medium
	Complexes	Low
	Eye drops	Low
	Co-crystalls	Medium

We assess dataset extraction complexity with five interrelated criteria that capture common challenges in automated scientific data extraction. Heterogeneous information formats—continuous text, tables, and figures that often disperse related data and encode values in complex plots or schematics—make parsing difficult [30]. Non-uniform table structures and cases where essential details appear only in the main text require cross-referencing, while semantic ambiguity in parameter labels and variable units demands contextual inference for correct mapping [30]. Records with single numeric values are easier to extract reliably, whereas multi-value records need careful linking of each value to the proper material and unit, increasing error risk. Finally, domain differences matter: inorganic nanomaterials frequently require hierarchical relationship extraction (composition and morphology $\rightarrow$ property), which is harder than extracting properties for small molecules that often use standardized encodings like SMILES [31].

Datasets are classified as low, medium, or high complexity based on these factors, with multi-format parsing, irregular tables, multi-value linking, and hierarchical relationships elevating difficulty (Table 4).

B Quality Control

A critical aspect of ChemX is its rigorous quality control process (Figure 2). To evaluate data integrity, we applied a stratified manual cross-verification procedure depicted on Figure 2.

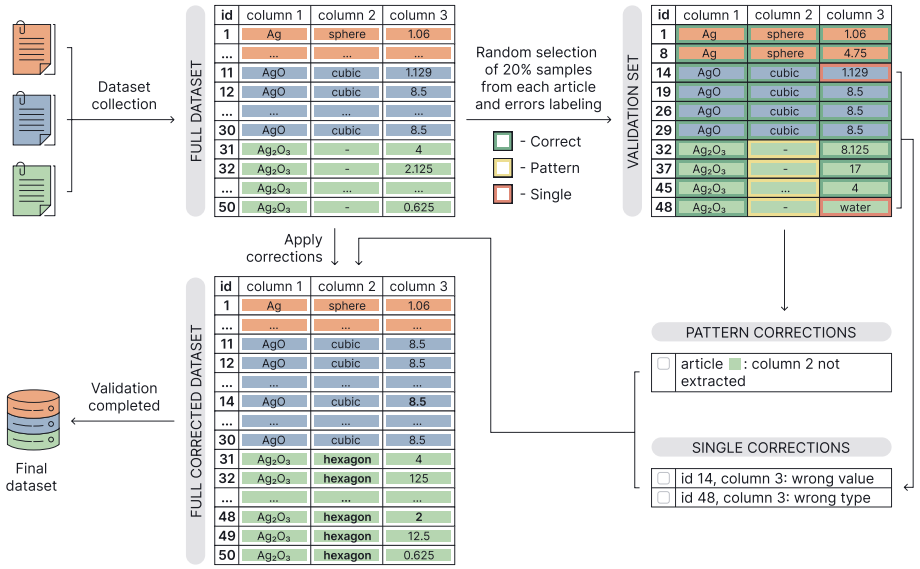

Figure 2: Quality control process for ChemX datasets

From each source article represented in a dataset, approximately 20% of entries were randomly selected and reviewed against the original source material, including PDFs, figures, and supplementary tables. Sampling was rounded up to ensure that at least one entry from each source article was manually reviewed during the verification process.

Errors — including transcription mistakes, structural mismatches, unit inconsistencies, and unsupported inferences — were categorized as either common (recurring patterns) or isolated (single occurrence). Importantly, if an isolated error was identified during review, we systematically checked all the other entries from the same source article, even if they were not part of the original sample. This additional step was intended to determine whether similar issues occurred in other records from the same publication. In many cases, this allowed us to detect recurring patterns that were not evident in the initial sample, enabling the expansion of our correction rules beyond the reviewed subset. As a result, even single-instance errors had the potential to lead to pattern-based corrections across the dataset.

Error categorization informed the correction strategy. For common errors, we formulated rule-based recommendations that specified the field affected, the observed scope of recurrence, and the appropriate method for correction, such as structural replacement, unit standardization, or removal of inferred content. Corrections were then applied across the whole group. All recommendations were documented in writing and communicated to the dataset curators for implementation across relevant records. Isolated issues were corrected individually.

C Datasets Overview and Analysis

Figure 3B shows the number of openly accessible articles per dataset. The publication year distribution (Figure 3A) reflects literature growth since the early 2000s, with a sharp increase in the past decade. We also assessed missing values across datasets (Figure 3C), with some exhibiting high sparsity due to incomplete reporting. This heterogeneity in data completeness benefits benchmarking by enabling rigorous evaluation of automated extraction systems—testing both accurate retrieval of reported values and correct identification of missing data.

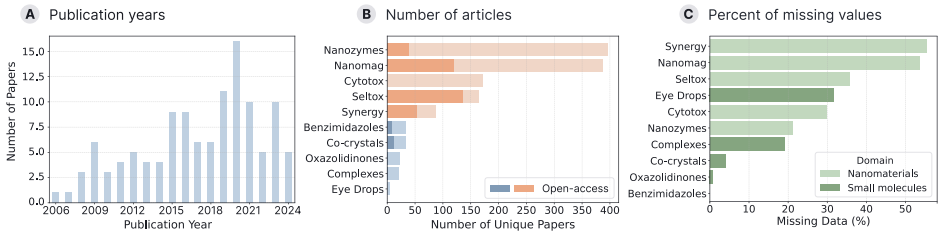

Figure 3: Quality control process for ChemX datasets

D Experiments

D.1 Selected articles

For each domain, we selected the two datasets of lowest complexity (nanozymes and complexes). For each dataset, three articles were picked for the experiments:

1. Nanozymes
   (a) Oxidase-Like Catalytic Performance of Nano-MnO2 and Its Potential Application for Metal Ions Detection in Water **(Open Access)**
   (b) Size Effect in Pd-Ir Core-Shell Nanoparticles as Nanozymes
   (c) Single Nanoparticle to 3D Supercage: Framing for an Artificial Enzyme System
2. Complexes
   (a) Prediction of Gd(III) complex thermodynamic stability
   (b) Coordinating Radiometals of Copper, Gallium, Indium, Yttrium, and Zirconium for PET and SPECT Imaging of Disease
   (c) Technetium and rhenium: coordination chemistry and nuclear medical applications **(Open Access)**

D.2 Prompts and metrics

For evaluating data extraction quality, we calculated the following:

- **True Positives (TP):** The count of values correctly extracted (i.e., the value exists in both the original dataset and the extracted dataset).
- **False Positives (FP):** The count of values incorrectly extracted (i.e., the value does not exist in the original dataset but is present in the extracted dataset).

- **False Negatives (FN):** The count of missing values (i.e., the value exists in the original dataset but is absent from the extracted dataset).

For each PDF in the dataset, we computed precision, recall, and F1 score based on those quantities. The resulting metrics were then aggregated across all PDFs in the dataset and averaged by dividing the total sum by the number of PDFs.

To standardize inputs, we created the following prompt template:

system_prompt = "You are a domain-specific chemical information extraction assistant. You specialize in the chemistry of ... . Your area of expertise includes ... ."

user_prompt = "Your task is to extract **every** mention of ... for ... from a scientific article, and output a **JSON array** of objects **only** (no markdown, no commentary, no extra text):

1. *Feature 1 (string)*: *Description* (e.g., *'example'*).
2. *Feature 2 (numeric)*: *Description* (e.g., *'example'*).
3. ...
4. *Target value (numeric)*: *Description* (e.g., *'example'*).

Extraction rules:

- Extract **each** ... mention as a separate object.
- Do **not** filter, group, summarize, or deduplicate. Include repeated mentions and duplicates if they occur in different contexts.
- If you cannot find a required field for an object, re-check the context; if it's still absent, set that field's value to "NOT_DETECTED"
- *Other rules specific to this dataset*
- The example of JSON below shows only one extracted samples, however your output should contain **all** mentions of ... for ... present in the article.

Output **must** be a single JSON array, like: [{ "feature 1": "example of feature 1", "feature 2": "example of feature 2", ... "target value": "example of target value" }]"

Complexes

system_prompt = "You are a domain-specific chemical information extraction assistant. You specialize in the chemistry of organometallic complexes and their properties."

user_prompt = "Your task is to extract **every** mention of organometallic complexes and chelate ligands from scientific article, and output a **JSON array** of objects **only** (no markdown, no commentary, no extra text).

Fields for each object:

- ‘compound_id‘ (string): ID of a complex within the article, as cited in the text, e.g. ‘"L3"‘, ‘"A31"‘.
- ‘compound_name‘ (string): abbreviated or full name of the complex or ligand as cited in the text, e.g. ‘"DOTA"‘, ‘"tebroxime"‘.
- ‘SMILES‘ (string): full SMILES representation of ligand environment or single ligand. If a complete organometallic complex is shown, extract all ligand structures without mentioning the metal (e.g., "COc1cc(C=CC([O-])CC([O-])CC([O-])C=Cc2ccc(O)c(OC)c2)ccc1O. [C-]#[O+].[C-]#[O+].[C-]#[O+].[OH-]"). For a chelate ligand without a complete organometallic complex, extract only that ligand's structure (e.g., 'O=C(O)CN(CCN(CC(CC(=O)O)CC(=O)O)CCN(CC(=O)O)CC(=O)O').
- ‘SMILES_type‘ (string): one of ‘"ligand"‘ or ‘"environment"‘. "environment" refers to the entire organometallic complex, including one or more ligands and a metal atom.
- ‘target_value‘ (number): the numeric value of logarithms of thermodynamic stability constants lgK or logK (without quotes).

Extraction rules:

1. Extract **each** mention of ‘target_value‘ (lgK or logK) as a separate object.
2. Do **not** filter, group, summarize, or deduplicate. Include repeated mentions and duplicates if they occur in different contexts.
3. If a molecule is fully depicted in a figure, write it as a SMILES string. If a molecule is depicted as a scaffold and residues separately in different places of an article, connect them by compound ID or name into one molecule and write it a single SMILES string.
4. If multiple thermodynamic stability constants appear for the same complex or ligand extract each separately.
5. Extract only structures that comply with these rules:
   - The complexes must contain **Ga** as the metal or the ligands must belong to complexes of that metal.
   - The complete molecular structure shall be given without errors in it or identifiers.
   - Compounds must contain more than one carbon (exclude CO, Me).
   - Compounds must not contain polymeric structures, attached biomolecules or carboranes, undefined radicals, undeciphered designations (e.g., amino acids) beyond the simplest abbreviations (i.e., Me, Et, Pr, Bu, Ph, Ac), names of radicals instead of their structure, or incomplete indication of the ligand structure (e.g., L = P, N).
   - Compounds must not be reaction intermediate or precursor.
6. If you cannot find a required field for an object, re-check the context; if it’s still absent, set that field’s value to ‘"NOT_DETECTED"‘.
7. The example of JSON below shows only two extracted samples, however your output should contain **all** mentions of organometallic complexes and / or chelate ligands present in the article.

Output **must** be a single JSON array, like:
[{
"compound_id": "L3",
"compound_name": "DOTA",
"SMILES": "O=C(O)CN(CCN(CC(=O)O)CC(=O)O)CC(=O)O",
"SMILES_type": "ligand",
"target": 21.3 }, {
"compound_id": "A31",
"compound_name": "tebroxime",
"SMILES": "[C-]#[N+]CC(C)(C)OC.[C-]#[N+]CC(C)(C)OC.[C-]#[N+]CC(C)(C)OC.[C-]#[N+]CC(C)(C)OC.[C-]#[N+]CC(C)(C)OC.[C-]#[N+]CC(C)(C)OC",
"SMILES_type": "environment",
"target": 17.9 }]"

Nanozymes

system_prompt = "You are a domain-specific chemical information extraction assistant. You specialize in nanozymes."

user_prompt = "Your task is to extract **every** mention of experiments for **ALL** nanozymes from a scientific article and output a **JSON array** of objects **only** (no markdown, no commentary, no extra text).

Fields for each object:

- ‘formula‘ (string): the chemical formula of the nanozyme, e.g. "Fe3O4", "CuO", etc.
- ‘activity‘ (string): catalytic activity type, typically "peroxidase", "oxidase", "catalase", "laccase", or other.
- ‘syngony‘ (string): the crystal unit of the nanozyme, e.g. "cubic", "hexagonal", "tetragonal", "monoclinic", "orthorhombic", "trigonal", "amorphous", "triclinic".
- ‘length‘ (number): the length of the nanozyme particle in nanometers.

- ‘width‘ (number): the width of the nanozyme particle in nanometers.
- ‘depth‘ (number): the depth of the nanozyme particle in nanometers.
- ‘surface‘ (string): the molecule on the surface of the nanozyme, e.g., "naked", "poly(ethylene oxide)", "poly(N-Vinylpyrrolidone)", "Tetrakis(4-carboxyphenyl)porphine", or other.
- ‘km_value‘ (number): the Michaelis constant value for the nanozyme.
- ‘km_unit‘ (string): the unit for the Michaelis constant, e.g., "mM", etc.
- ‘vmax_value‘ (number): the molar maximum reaction rate value.
- ‘vmax_unit‘ (string): the unit for the maximum reaction rate, e.g., "μmol/min", "mol/min", etc.
- ‘reaction_type‘ (string): the reaction type involving the substrate and co-substrate, e.g., "TMB + H2O2", "H2O2 + TMB", "TMB", "ABTS + H2O2", "H2O2", "OPD + H2O2", "H2O2 + GSH", or other.
- ‘c_min‘ (number): the minimum substrate concentration in catalytic assays in mM.
- ‘c_max‘ (number): the maximum substrate concentration in catalytic assays in mM.
- ‘c_const‘ (number): the constant co-substrate concentration used during assays.
- ‘c_const_unit‘ (string): the unit of measurement for co-substrate concentration.
- ‘ccat_value‘ (number): the concentration of the catalyst used in assays.
- ‘ccat_unit‘ (string): the unit of measurement for catalyst concentration.
- ‘ph‘ (number): the pH level at which experiments were conducted.
- ‘temperature‘ (number): the temperature in Celsius during the study.

Extraction rules:

1. Extract **each** nanozyme mention as a separate object.
2. Do **not** filter, group, summarize, or deduplicate. Include repeated mentions and duplicates if they occur in different contexts.
3. If you cannot find a required field for an object, re-check the context; if it’s still absent, set that field’s value to ‘"NOT_DETECTED"‘.
4. The example of JSON below shows only two extracted samples, however your output should contain **all** nanozymes present in the article.

Output **must** be a single JSON array, like:
[{
"formula": "Fe3O4",
"activity": "peroxidase",
"syngony": "cubic",
"length": 10,
"width": 10,
"depth": 2.5,
"surface": "naked",
"km_value": 0.2,
"km_unit": "mM",
"vmax_value": 2.5,
"vmax_unit": "μmol/min",
"reaction_type": "TMB + H2O2",
"c_min": 0.01,
"c_max": 1.0,
"c_const": 1.0,
"c_const_unit": "mM",
"ccat_value": 0.05,
"ccat_unit": "mg/mL",
"ph": 4.0,
"temperature": 25 }, {

"formula": "CeO2",
"activity": "oxidase",
"syngony": "cubic",
"length": 5,
"width": 5,
"depth": 200,
"surface": "poly(ethylene oxide)",
"km_value": 54.05,
"km_unit": "mM",
"vmax_value": 7.88,
"vmax_unit": "10-8 M s-1",
"reaction_type": "TMB",
"c_min": 0.02,
"c_max": 0.8,
"c_const": 800,
"c_const_unit": "μM",
"ccat_value": 0.02,
"ccat_unit": "mg/mL",
"ph": 5.5,
"temperature": 37 }]"

E Results and Discussion

Table 5: All metrics for complexes dataset (baseline models).

Column	GPT-5			GPT-5 Thinking		
	Precision	Recall	F1	Precision	Recall	F1
compound_id	0.65	0.29	0.35	0.65	0.52	0.58
compound_name	0.41	0.22	0.26	0.44	0.37	0.40
SMILES	0.14	0.03	0.04	0.00	0.00	0.00
SMILES_type	0.67	0.3	0.36	0.00	0.00	0.00
target	0.41	0.1	0.14	0.00	0.00	0.00

Table 6: All metrics for complexes dataset (single-agent approach).

Column	GPT-4.1			GPT-5			GPT-OSS-20b		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
compound_id	0.56	0.35	0.43	0.73	0.88	0.80	0.74	0.63	0.68
compound_name	0.13	0.08	0.1	0.05	0.06	0.05	0.07	0.06	0.07
SMILES	0.06	0.04	0.05	0.00	0.00	0.00	0.00	0.00	0.00
SMILES_type	1.00	0.63	0.77	0.83	1.00	0.91	1.00	0.84	0.91
target	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7: All metrics for complexes dataset (multi-agent approaches).

Column	ChatGPT Agent			SLM Matrix			FutureHouse		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
compound_id	0.64	0.52	0.57	0.93	0.89	0.91	0.06	0.06	0.06
compound_name	0.50	0.41	0.45	0.11	0.11	0.11	0.00	0.00	0.00
SMILES	0.06	0.04	0.05	0.01	0.01	0.01	0.00	0.00	0.00
SMILES_type	0.56	0.47	0.51	0.94	0.90	0.92	0.52	0.24	0.25
target	0.73	0.67	0.70	0.00	0.00	0.00	0.04	0.00	0.00

Table 8: All metrics for nanozymes dataset (baseline models).

Column	GPT-5			GPT-5 Thinking		
	Precision	Recall	F1	Precision	Recall	F1
formula	0.62	1.00	0.71	0.02	0.08	0.03
activity	0.62	1.00	0.71	0.02	0.08	0.03
syngony	0.62	1.00	0.71	0.02	0.08	0.03
length	0.36	0.42	0.38	0.02	0.04	0.03
width	0.25	0.25	0.25	0.02	0.02	0.01
depth	0.47	0.58	0.52	0.01	0.02	0.01
surface	0.00	0.00	0.00	0.00	0.03	0.00
km_value	0.07	0.33	0.11	0.01	0.05	0.02
km_unit	0.07	0.33	0.11	0.01	0.05	0.02
vmax_value	0.40	0.67	0.44	0.01	0.05	0.02
vmax_unit	0.40	0.67	0.44	0.01	0.05	0.02
reaction_type	0.44	0.50	0.47	0.02	0.04	0.02
c_min	0.07	0.33	0.11	0.00	0.03	0.00
c_max	0.07	0.33	0.11	0.00	0.03	0.00
c_const	0.33	0.33	0.33	0.00	0.00	0.00
c_const_unit	0.40	0.67	0.44	0.01	0.05	0.02
ccat_value	0.46	0.83	0.54	0.00	0.04	0.01
ccat_unit	0.33	0.33	0.33	0.01	0.03	0.02
ph	0.62	1.00	0.71	0.02	0.08	0.03
temperature	0.00	0.00	0.00	0.00	0.00	0.00

Table 9: All metrics for nanozymes dataset (Single-agent approach).

Column	GPT-4.1			GPT-5			GPT-OSS-20b		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
formula	0.56	1.00	0.71	0.62	1.00	0.71	0.83	1.00	0.91
activity	0.56	1.00	0.71	0.62	1.00	0.71	0.83	1.00	0.91
syngony	0.56	1.00	0.71	0.62	1.00	0.71	0.17	0.20	0.18
length	0.44	0.80	0.57	0.36	0.42	0.38	0.67	0.80	0.73
width	0.11	0.20	0.14	0.25	0.25	0.25	0.67	0.80	0.73
depth	0.11	0.20	0.14	0.47	0.58	0.52	0.67	0.80	0.73
surface	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
km_value	0.56	1.00	0.71	0.07	0.33	0.11	0.83	1.00	0.91
km_unit	0.44	0.80	0.57	0.07	0.33	0.11	0.67	0.80	0.73
vmax_value	0.56	1.00	0.71	0.40	0.67	0.44	0.83	1.00	0.91
vmax_unit	0.44	0.80	0.57	0.40	0.67	0.44	0.67	0.80	0.73
reaction_type	0.56	1.00	0.71	0.44	0.50	0.47	0.67	0.80	0.73
c_min	0.44	0.80	0.57	0.07	0.33	0.11	0.17	0.20	0.18
c_max	0.44	0.80	0.57	0.07	0.33	0.11	0.17	0.20	0.18
c_const	0.44	0.80	0.57	0.33	0.33	0.33	0.67	0.80	0.73
c_const_unit	0.56	1.00	0.71	0.40	0.67	0.44	0.67	0.80	0.73
ccat_value	0.33	0.60	0.43	0.46	0.83	0.54	0.50	0.60	0.55
ccat_unit	0.44	0.80	0.57	0.33	0.33	0.33	0.67	0.80	0.73
ph	0.56	1.00	0.71	0.62	1.00	0.71	0.83	1.00	0.91
temperature	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 10: All metrics for nanozymes dataset (Multi-agent approaches).

Column	SLM-Matrix			FutureHouse			NanoMINER		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
formula	0.25	1.00	0.40	0.12	0.67	0.21	-	-	-
activity	0.25	1.00	0.40	0.12	0.67	0.21	-	-	-
syngony	0.05	0.20	0.08	0.08	0.50	0.14	-	-	-
length	0.20	0.80	0.32	0.04	0.17	0.07	-	-	-
width	0.20	0.80	0.32	0.00	0.00	0.00	-	-	-
depth	0.20	0.80	0.32	0.04	0.17	0.07	-	-	-
surface	0.20	0.80	0.32	0.08	0.33	0.13	-	-	-
km_value	0.05	0.20	0.08	0.04	0.33	0.07	0.97	0.91	0.94
km_unit	0.05	0.20	0.08	0.08	0.50	0.14	-	-	-
vmax_value	0.05	0.20	0.08	0.04	0.33	0.07	0.96	0.83	0.89
vmax_unit	0.05	0.20	0.08	0.04	0.33	0.07	-	-	-
reaction_type	0.20	0.80	0.32	0.04	0.17	0.07	-	-	-
c_min	0.05	0.20	0.08	0.04	0.33	0.07	0.97	0.54	0.69
c_max	0.05	0.20	0.08	0.08	0.50	0.14	0.97	0.53	0.69
c_const	0.20	0.80	0.32	0.00	0.00	0.00	0.78	0.51	0.62
c_const_unit	0.20	0.80	0.32	0.04	0.33	0.07	-	-	-
ccat_value	0.05	0.20	0.08	0.04	0.33	0.07	0.88	0.81	0.84
ccat_unit	0.00	0.00	0.00	0.00	0.00	0.00	-	-	-
ph	0.25	1.00	0.40	0.12	0.67	0.21	0.98	0.82	0.89
temperature	0.20	0.80	0.32	0.00	0.00	0.00	0.70	0.96	0.81

F Limitations

While this benchmark encompasses ten datasets across two chemical domains, its scope is necessarily constrained and does not extend to other critical areas of chemistry, including organic reaction schemes, spectral data, quantum chemical calculations, and others.

Our experimental results on structure extraction underscore the inherent limitations of both general-purpose large language models (LLMs) and agent-based methodologies for the specific task of chemical structure recognition. Furthermore, even specialized agent-based systems demonstrated suboptimal performance. Although dedicated tools such as DECIMER [32] exist for converting molecular images into SMILES strings, their practical integration into automated extraction pipelines is presently precluded by two unresolved technical challenges: (1) the reliable detection of individual molecular images within the complex layouts of scientific articles, and (2) the accurate segmentation of images exhibiting heterogeneous formats and styles. Future advancements in computer vision, particularly in automated molecular localization and standardized image preprocessing, may eventually facilitate the incorporation of such tools. However, due to these extant limitations, tools like DECIMER were deliberately excluded from the present experimental framework. It is critical to note that the incorrect extraction of chemical structures poses significant risks; hallucinations or errors can propagate through automated workflows, leading to failures in reproducibility, invalid computational results, and ultimately, the generation of erroneous scientific data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the release of ChemX, a curated benchmark of 10 datasets for automated information extraction in chemistry, and the evaluation of both mono- and multi-agent LLM-based systems.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section F discusses multiple limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In this article, we provide full documentation for each dataset, describe the methodology of the extraction experiments, and also include the code for conducting these experiments in Sections 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in the supplemental material?

Answer: [Yes]

Justification: Datasets and code are available via HuggingFace and GitHub with accompanying documentation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 4 and D outline LLM setup, prompt structure, document formats, and evaluation procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Experimental errors were not incorporated into the analysis, as the central claim of this work is not the comparative performance of the methods. Rather, we assert that all evaluated methods perform inadequately for the task. Consequently, the consideration of measurement error is immaterial, as its inclusion would not alter this overarching conclusion.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments involving large language models such as GPT-4o were executed via the OpenAI API. All other computations, including preprocessing, single-agent pipeline execution, and evaluation metrics, were performed locally on a laptop with the following specifications: Intel Core i7-11800H (8 cores, 2.3–4.6 GHz), 16 GB RAM, and a 512 GB SSD. The GPU was not used for local execution.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All content was extracted from publicly accessible scientific literature or subscription-based academic access with proper institutional rights. No sensitive data or human participants were involved.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Section F discusses the risks of incorrect extraction, hallucination in chemical contexts, and implications for reproducibility and automation in cheminformatics.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No high-risk pretrained models or internet-scraped data were released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The dataset is manually extracted from open-access and subscription-based articles accessed under institutional license, and all external tools and models are properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All 10 datasets are fully documented with schemas, annotation examples, and feature descriptions in the supplementary material and HuggingFace page.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human participants or crowdworkers were involved in data collection or validation.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: No human subjects were involved in the study. All data were derived from published scientific literature and manually curated by the authors.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper explicitly discusses GPT-4.1, GPT-5, GPT-OSS-20b use for both baseline model and single-agent pipelines in Section 4.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.