Evaluating Chemistry-Guided Filtering Heuristics Using LLM-Extracted Reaction SMILES

Chaewon Lee[1], Shuan Chen[1,2], and Yousung Jung[1,2]*

[*]Equal contribution [1]Department of Chemical and Biological Engineering (BK21 Four) and Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea [2]Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

Correspondence to: Yousung Jung yousung.jung@snu.ac.kr

1. Introduction

Chemical reaction data plays a central role in training machine learning models for synthesis planning, yet most synthesis information exists as unstructured experimental text. Lowe addressed this by extracting a large-scale reaction dataset from USPTO patent documents, which has become a widely used benchmark, further extended by the Open Reaction Database (ORD) [1]. However, Chen and Jung [2] found that many reactions contain products inconsistent with original patent records and lack critical reagent information, rendering them chemically infeasible. To filter such errors, Wigh et al. developed ORDerly [3], a rule-based heuristic framework, yet the root causes of heuristic failures and the types of data noise these rules can or cannot address remain poorly understood.

Recent advances in LLMs offer a scalable way to analyze unstructured chemical descriptions; however, prior work by Zhang et al. [4] and Ai et al. [5] does not reliably reconstruct complete reaction SMILES due to ambiguous compound names (e.g., “title compound”, “compound 1”) in the source text. In this study, we fine-tune LLMs on 75 manually curated synthesis examples to extract full reaction SMILES from patent text, evaluate four chemistry-guided heuristics from ORDerly [3], and proposed a more robust heuristic, *rare-template*, to mitigate the false negative (FN) reactions filtered out by previous heuristics. Our LLM-based extractor achieves 92% extraction accuracy, and the chemistry-guided heuristics exhibit substantially different false-negative rates (FNRs) across categories, with 71% for *many-products* and 8% for *rare-template* reactions.

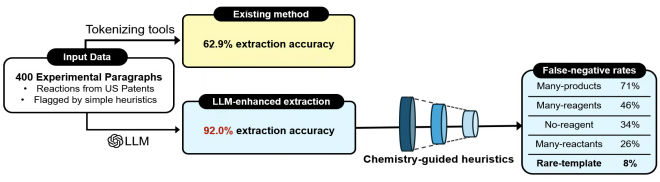


Fig. 1: The overall process and filtering performance of the chemistry-guided heuristics described in this work.

2. Methods

2.1 Flagged Reaction Data in USPTO Dataset

In this study, we focus on re-extracting reaction SMILES from original patent paragraphs and evaluating chemistry-guided heuristics on reactions flagged by ORDerly [3]. We exclude reactions without reactants or products, as none were found in the USPTO dataset, and omit duplicate reactions since they typically originate from nearly identical paragraphs. Solvents and reagents are treated as a unified category, classifying a chemical as a reagent if any of its heavy atoms do not contribute to the major product. Based on these considerations, we evaluate four heuristics: (1) *many-reactants*, reactions with more than two reactants; (2) *many-products*, reactions with more than two products; (3) *many-reagents*, reactions with more than five reagents; and (4) *no-reagent*, reactions with no reagents.

2.2 Extracting Reaction Data using LLMs

We designed an LLM-based workflow to extract reaction data from experimental paragraphs of US patents using gpt-3.5-turbo-0125 and gpt-4-0613, comprising four steps. First, LLMs generate a structured JSON summary of each paragraph, categorizing chemicals into “Reactants, Solvents, Catalysts, Product”, and reaction steps into “Reaction” and “Work-up”. Second, each chemical is converted to SMILES format by sequentially applying OPSIN [6], PubChem [7], CIR [8], and ChemSpider [9] APIs, with all strings canonicalized for consistency. Third, work-up steps are removed, and the remaining synthesis steps are merged into a single reaction SMILES string with each step separated by a “>” token, enabling identification of reagents across multiple synthetic steps. Finally, LocalMapper [10] is applied for atom-mapping.

2.3. Evaluating Heuristics’ Filtering Performance

To evaluate the filtering performance of four chemistry-guided heuristics (*many-reactants, many-products, many-reagents,* and *no-reagent*), we assessed whether the GPT-extracted SMILES satisfy each heuristic. We labeled a reaction as "false negative (FN)" when the heuristic-triggering condition in the original USPTO entry was no longer present in the GPT-extracted SMILES, and as "true negative (TN)" when it persisted. For instance, under the *many-reactants* heuristic, a reaction was labeled as FN if the original entry triggered the rule (e.g., three reactants) but the GPT- extracted SMILES no longer did (e.g., two reactants).

3. Results

We evaluated four heuristics by assessing whether GPT-extracted SMILES still trigger each filtering condition. Since FNs result in irreversible data loss whereas FPs can be corrected through post-processing, minimizing FNR is the primary concern. TN cases fall into three categories: *invalid chemical names, untraceable references*, and *others* (**Fig. 2a**). FNRs decreased in the order of *many-products* (71%), *many-reagents* (46%), *no-reagent* (34%), and *many-reactants* (26%) (**Fig. 2b**).

Notably, many TN reactions exhibit unusual reaction patterns, motivating the introduction of a *rare-template* heuristic that flags reactions associated with extremely infrequent templates (appearing only once among 1.9M reactions) or with no structural change (**Fig. 3a**). This heuristic achieves the lowest FNR (8%), with most TN cases attributed to the *others* (**Fig. 3b**), indicating that flagged reactions reflect genuinely problematic records rather than extraction errors.

While a small number of FN cases remain where missing reactants are correctly recovered by the LLM (**Fig. 3d**), but still exhibits uncommon reaction pattern before and after LLM-re-extraction (**Fig. 3c**). When applied to the full USPTO dataset, it flags 4.16% of reactions (nearly 73,000 reactions) as unrecoverable, demonstrating a more robust and data-preservative filtering strategy than existing ORDerly [3] heuristics.

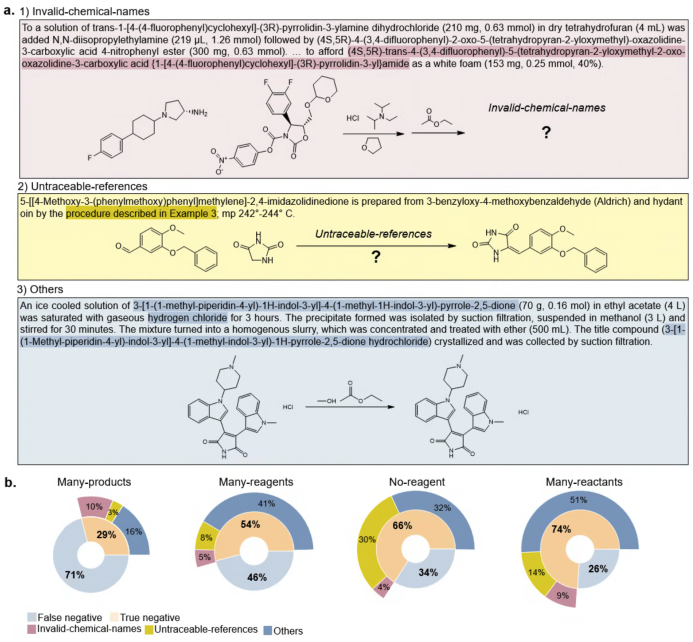


Fig. 2 Analysis of FN reactions and examples of TN reactions for four heuristics: (a) Examples of the TN cases in each category. The *others* example is taken from *many-products*. (b) The inner pie charts show the proportions of FN and TN reactions for each heuristic category: *many-products*, *many-reagents*, *no-reagent*, and *many-reactants*. TN reactions are further broken down into three failure types: *invalid-chemical-names*, *untraceable-references*, and *others*.

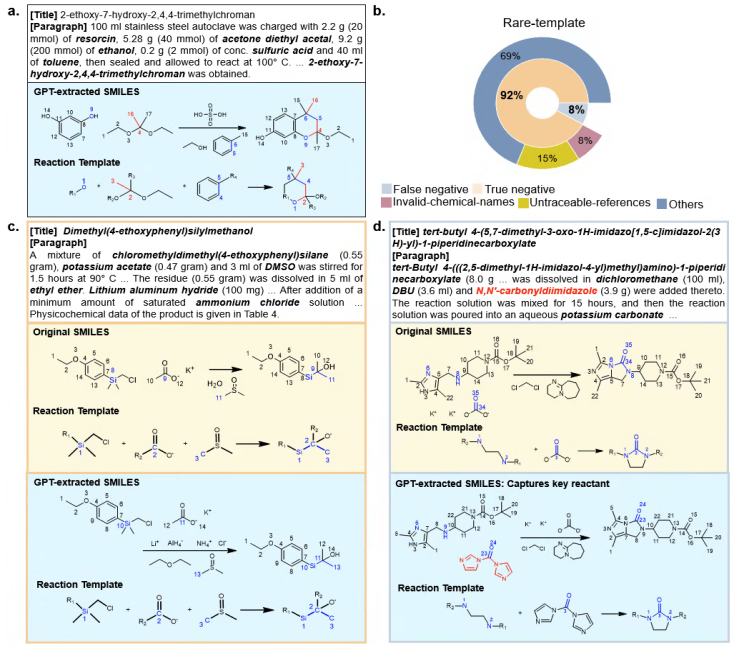


Fig. 3 Examples and analysis for *rare-template*: (a) Representative *rare-template* example showing an atypical reactant set (also triggering *many-reactants*). (b) The inner pie charts show the proportions of FN and TN reactions. TN reactions are further broken down into three failure types: *invalid-chemical-names*, *untraceable-references*, and *others*. (c) Example of TN. (d) Example of FN.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2025-00514706 and RS-2024-00464386).

References

[1] S. M. Kearnes, et al. The Open Reaction Database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021. doi:10.1021/jacs.1c09820.

[2] S. Chen and Y. Jung. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4(9):772–780, 2022. doi:10.1038/s42256-022-00526-z.

[3] D. S. Wigh, J. Arrowsmith, A. Pomberger, K. C. Felton, and A. A. Lapkin. ORDerly: Data Sets and Benchmarks for Chemical Reaction Data. *Journal of Chemical Information and Modeling*, 64(9):3790–3798, 2024. doi:10.1021/acs.jcim.4c00292.

[4] W. Zhang, et al. Fine-tuning large language models for chemical text mining. *Chemical Science*, 15(27):10600–10611, 2024. doi:10.1039/D4SC00924J.

[5] Q. Ai, F. Meng, J. Shi, B. Pelkie, and C. W. Coley. Extracting structured data from organic synthesis procedures using a fine-tuned large language model. *Digital Discovery*, 3(9):1822–1831, 2024. doi:10.1039/D4DD00091A.

[6] D. M. Lowe, P. T. Corbett, P. Murray-Rust, and R. C. Glen. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling*, 51(3):739–753, 2011. doi:10.1021/ci100384d.

[7] S. Kim, et al. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi:10.1093/nar/gkac956.

[8] NCI/CADD Group. NCI/CADD Group Chemoinformatics Tools and User Services. https://cactus.nci.nih.gov/, 2025. Accessed: 2025-07-09.

[9] Royal Society of Chemistry. ChemSpider: Search and Share Chemistry - Homepage. https://www.chemspider.com/, 2025. Accessed: 2025-07-09.

[10] S. Chen, S. An, R. Babazadeh, and Y. Jung. Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. *Nature Communications*, 15(1):2250, 2024. doi:10.1038/s41467-024-46364-y.

Appendix A. Statistics of flagged reaction data identified in this work.

We filtered out 536,372 reactions by applying these four criteria, which share around 28% of the total number from the full USPTO dataset of 1.9 million reactions. Notably, the sum of reactions classified under all filtered types exceeds the total number of reactions because one reaction can match multiple filtering criteria.

Filtering criterion	Number of reactions	Ratio in the USPTO dataset
Many reactants (>2)	52,758	2.77%
Many products (>1)	82,203	4.32%
Many reagents (>5)	236,368	12.43%
No reagent	200,594	10.55%
Total unique reactions	536,372	28.20%

Appendix B. Examples of input and output for finetuned-GPT

(a) Input example for LLMs: Example of a title and paragraph provided as input for LLMs. (b) Output example: Illustration of the structured output and response generated by the LLM. (c) Comparative analysis of SMILES: Comparison between original SMILES from the USPTO dataset and SMILES extracted by the fine-tuned GPT model. The GPT-extracted SMILES preserve sequential information by separating merged reaction steps into distinct steps, accurately reflecting the true experimental procedure. In contrast, the original SMILES often combine multiple steps into a single representation.

a. **[Title]** 1-Nitro-3-[(propan-2-ylsulfonyl)methyl]benzene
[Paragraph] A solution of ***1-nitro-3-[(propan-2-ylsulfanyl)methyl]benzene*** (4.0 g) in ***DCM*** (160 mL) was treated at 0° C. with portions of ***m-chloroperbenzoic acid*** (9.3 g, 77%). The mixture was stirred at 0° C. for further 30 minutes and then 18 hours at room temperature. The reaction mixture was diluted with DCM before ***sodium hydrogen sulfite*** and ***sodium bicarbonate*** solution was added and extracted with DCM (2×). The combined organic phases were washed and concentrated. The residue was purified by chromatography (hexane/ethyl acetate 12%-100%) to give the ***title compound*** (4.5 g).

b.

Output Structure

{\"**Reactants, Solvents, Catalysts\"**: {\"Code\": \"name\"}, \"**Product\"**: {\"Code\": \"name\"}, \"**Reaction Steps\"**: {\"Number (Reaction/Work-up, Action)\": \"Reactant(s)+Solvent(s)+Catalyst(s)->mixture\"}}

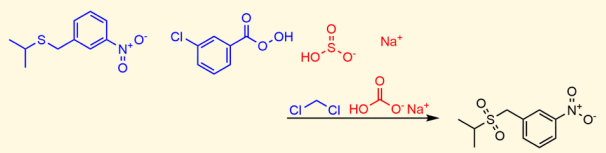
GPT response

{ **"Reactants, Solvents, Catalysts"**: {
"A": "1-nitro-3-[(propan-2-ylsulfanyl)methyl]benzene",
"B": "DCM",
"C": "m-chloroperbenzoic acid",
"D": "sodium hydrogen sulfite",
"E": "sodium bicarbonate",
"F": "hexane",
"G": "ethyl acetate" },
"Product": {
"H": "1-Nitro-3-[(propan-2-ylsulfonyl)methyl]benzene" },
"Reaction Steps":{
"1 (Reaction, Add)": "A+B+C->mixture1",
"2 (Reaction, Heat)": "mixture1->mixture2",
"3 (Work-up, Dilute)": "mixture2+B
->mixture3",
"4 (Reaction, Add)": "mixture3+D+E->mixture4",
"5 (Work-up, Extract)": "mixture4+B->mixture5",
"6 (Work-up, Wash and Concentrate)":
"mixture5->mixture6",
"7 (Work-up, Purify)": "mixture6+F+G->H" } }

c.

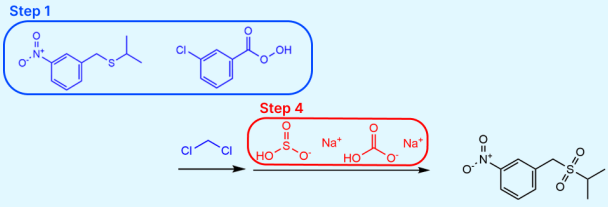
Original data: None-sequential reaction format

CC(C)SCc1cccc([N+](=O)[O-])c1.O=C(OO)c1cccc(Cl)c1.O=S([O-])O.[Na+]>ClCCl.O
=C([O-])O.[Na+]>CC(C)S(=O)(=O)Cc1cccc([N+](=O)[O-])c1



This study: Stepwise reaction format

CC(C)SCC1=CC(=CC=C1)[N+](=O)[O-].C(Cl)Cl.C1=CC(=CC(=C1)Cl)C(=O)OO>C(Cl)Cl>
OS(=O)[O-].[Na+].C(=O)(O)[O-].[Na+]>CC(C)S(=O)(=O)CC1=CC(=CC=C1)[N+](=O)[O-]



Appendix C. Comparison of precision, recall, and accuracy between the original data and the data extracted by LLMs.

We evaluated extraction accuracy by randomly sampling 100 paragraphs from each heuristic category (400 total), using manually annotated ground truth rather than the original data. Extraction results were categorized as TPs, FPs, TNs, and FNs, with precision, recall, and reaction-level accuracy used as evaluation metrics.

To enhance performance, we fine-tuned gpt-3.5-turbo-0125 on 75 manually refined summaries (15 per heuristic), referred to as finetuned-GPT, using non-overlapping paragraphs from the 400 evaluation set. As summarized in Table 2, finetuned-GPT consistently outperforms both the original data and GPT-4, showing the largest gain in reaction-level accuracy of 46.3% over the original data and 5.9% over GPT-4, with incremental improvements in precision and recall across individual components accumulating into higher overall accuracy. We therefore used finetuned-GPT for all subsequent analyses, referring to its outputs as "GPT-extracted SMILES".

Extraction method	**Reactants**		**Reagents**		**Product**		**Reaction**
	Precision	Recall	Precision	Recall	Precision	Recall	Accuracy
Original data	0.992	0.965	0.903	0.910	0.989	0.989	0.629
GPT-4	0.997	0.997	0.969	0.953	0.995	0.995	0.869
Finetuned-GPT	0.997	0.997	0.961	0.974	0.996	1	0.920

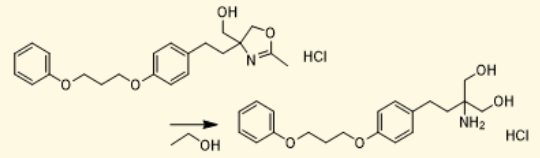
Appendix D. Comparison between original SMILES and GPT-extracted SMILES for different categories

(a) *many-products*, (b) *many-reagents*, and (c) *many-reactants*. Reaction information correctly extracted only by GPT is highlighted in red and blue.

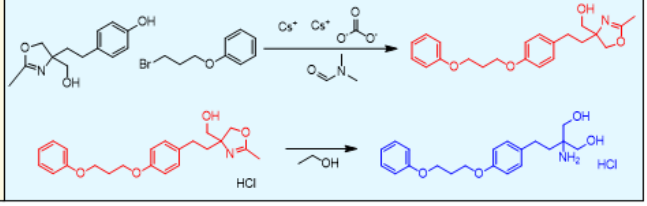
a.

[Title] 2-amino-2-{2-[4-(3-phenoxy-propoxy)-phenyl]-ethyl}-propane-1,3-diol hydro-chloride
[Paragraph] To a solution of ***4-[2-(4-hydroxymethyl-2-methyl-4,5-dihydro-oxazol-4-yl)-ethyl]-phenol*** (300 mg, 1.27 mmol) in ***DMF*** (5 ml) was added ***Cs2CO3*** (1.2 g, 3.83 mmol) and ***(3-bromo-propoxy)-benzene*** ... affords ***(2-methyl-4-{2-[4-(3-phenoxy-propoxy)-phenyl]-ethyl}-4,5-dihydro-oxazol-4-yl)-methanol*** as colorless oil. To a solution of ... in ***ethanol*** (2 ml) was added conc. ***HCl*** (2 ml). The reaction mixture was dried in vacuo to afford ***2-amino-2-{2-[4-(3-phenoxy-propoxy)-phenyl]-ethyl}-propane-1,3-diol hydrochloride*** as a colorless powder.

Original SMILES [*Many-products*]: Single-step (merged)



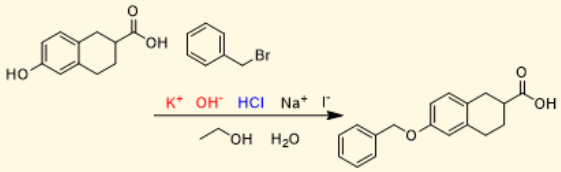
GPT-extracted SMILES: Captures Intermediate, Multi-step (separated)



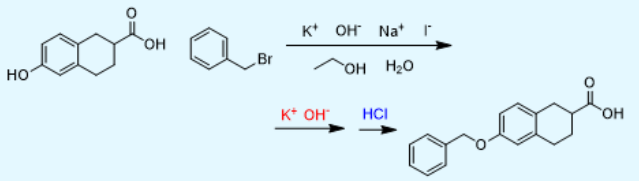
b.

[Title] 1,2,3,4-tetrahydro-6-benzyloxynaphthalene-2-carboxylic acid
[Paragraph] ... ***1,2,3,4-tetrahydro-6-hydroxynaphthalene 2 carboxylic acid***... ***benzyl bromide***, 6.6 g (100 mmol) of 85% ***potassium hydroxide***, 0.525 g (3.5 mmol) of ***sodium iodide***, 200 ml of ***ethanol*** and 25 ml of distilled ***water*** was heated ... To this reaction mixture was added 50 ml of 10% ***potassium hydroxide*** ... the reaction mixture in cold water was acidified with 36% ***hydrochloric acid*** ... obtain ***1,2,3,4-tetrahydro-6-benzyloxynaphthalene-2-carboxylic acid.***

Original SMILES [*Many-reagents*]: Single-step (merged)



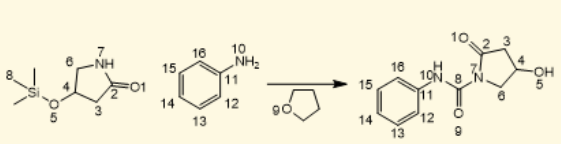
GPT-extracted SMILES: Multi-step (separated)



c.

[Title] 1-phenylcarbamoyl-4-hydroxy-2-pyrrolidinone
[Paragraph] A 8.14 g quantity of ***4-trimethylsilyloxy-2-pyrrolidone*** was dissolved in 100 ml of ***tetrahydrofuran***, 7.62 g of ***N,N'-carbonyldiimidazole*** was added to the solution ... With addition of 4.38 g of **aniline**, the mixture was further refluxed for 7 hours, then cooled and stirred at room temperature for 0.5 hour with addition of 150 ml of 1N ***hydrochloric acid***. ... The crystals were recrystallized from acetone-hexane, giving 7.44 g of ***1-phenylcarbamoyl-4-hydroxy-2-pyrrolidinone*** (yield 72%) having *a melting point of 104° to 104.5° C.*

Original SMILES [*Many-reactants*]: Wrong reactant



GPT-extracted SMILES: Captures key reactant