TOWARDS KNOWLEDGE-AND-DATA-DRIVEN OR-GANIC REACTION PREDICTION: RAG-ENHANCED AND REASONING-POWERED HYBRID SYSTEM WITH LLMS

Anonymous authors

000

001

002

004

006 007 008

009

010 011 012

013

015

016

017

018

019

021

023

025

026

027

028

029

031 032 033

034

037

040

041

042

043

044

045

046

047

048

051

052

Paper under double-blind review

ABSTRACT

In organic reaction prediction, many recent approaches ranging from traditional task-specific models to Large Language Models (LLMs), have demonstrated notable success. However, these methods are inherently data-driven, exhibit constrained interpretability, and have hit fundamental performance bottlenecks. To overcome these limitations, we present **Reaction-Thinker**, a hybrid, knowledge-and-data-driven system that is enhanced by Retrieval-Augmented Generation (RAG) and powered by advanced reasoning, improving both the interpretability of prediction process and the explainability of results. We develop similar-case retrieval database and train a RAG-based LLM through supervised fine-tuning (SFT) to apply both reaction types and similar reaction cases as knowledge. We also construct a reaction reasoning chain-of-thought (CoT) dataset and train a reasoning-based LLM through SFT, then further optimize it using Group Relative Policy Optimization (GRPO). Experimental results show that our method outperforms all compared LLMs and task-specific models, achieving the highest accuracy (Exact Match) and fingerprint similarity (FTS). Ablation study indicates improvements in relative accuracy of 7.5% and 13.9% for RAG and GRPO, respectively. Further analysis of mispredictions reveals limitations in conventional evaluation metrics, which motivates our proposed benchmarking refinement.

1 Introduction

In organic chemistry, predicting reaction outcomes has long been a core challenge. Traditionally, expert chemists relied on years of hands-on experience and well-established principles to design experiments and anticipate products. Today, artificial intelligence offers a powerful augment to this approach, enhancing the efficiency and precision of prediction.

Current approaches to predicting organic reaction outcomes can be mainly categorized into templatebased and template-free methods. Template-based methods integrate machine learning with predefined structural transformation rules, also known as reaction templates, either curated by experts or extracted from atom-mapped datasets Chen & Jung (2022); Sacha et al. (2021). In contrast, template-free methods employ data-driven architectures, such as graph neural networks (GNNs) or Transformer-based sequence models, to infer reaction patterns directly from large reaction corpora without relying on explicit templates. Schwaller et al. (2018); Irwin et al. (2022). Recent advances in large language models (LLMs) have garnered significant attention Achiam et al. (2023); Team et al. (2024); Bai et al. (2023); Liu et al. (2024). By undergoing large-scale pre-training followed by fine-tuning or instruction tuning, these models acquire extensive knowledge, proficiently follow instructions, and exhibit strong reasoning abilities. As a result, LLMs now achieve state-of-the-art (SOTA) performance comparable to or even exceeding that of humans, in general Natural Language Processing (NLP) tasks such as language understanding and question answering, as well as specialized applications including mathematical problem-solving and code generation. Hence, a natural idea is to explore whether LLMs can replicate the cognitive processes of expert chemists, enabling more accurate reaction predictions.

Human chemists predict organic reactions through a multi-step cognitive process. Initially, they analyze molecular structures, identifying functional groups, bond connectivity, stereochemistry, and reactive sites, which are fundamental to mechanistic analysis Smith (2023). Then, they apply core principles to hypothesize bond cleavage and formation. They propose reaction pathways and elucidate mechanistic steps, including identifying reaction centers, considering mechanisms such as SN1, SN2, or pericyclic, and evaluating thermodynamic and kinetic feasibility Levy (2017). Finally, they integrate insights to predict the main product and account for side reactions. Apart from that, known reaction cases are also frequently referenced to inform predictions.

Recent works have applied LLMs to chemistry, particularly targeting organic reactions prediction. Notable examples include the ChemDFM series Zhao et al. (2024b;a), ChemLLM Zhang et al. (2024), ChemCrow M. Bran et al. (2024), and Coscientist Boiko et al. (2023). Some approaches leverage large proprietary models such as GPT-4o OpenAI (2024) directly, exploiting their innate zero-shot reasoning capabilities. Others build on open-source LLMs such as LLaMA Touvron et al. (2023) and fine-tune them on chemical literature and curated datasets, resulting in domain-specific models with enhanced accuracy on chemical question answering (Q&A) and prediction tasks.

However, current fine-tuning methods for LLMs in chemistry primarily rely on data-driven strategies, which are similar to traditional end-to-end deep learning approaches, and often fail to fully leverage the rich chemical knowledge embedded in the pre-trained parameters of LLMs, and fine-tuned models tend to underutilize their robust reasoning and in-context learning capabilities. Consequently, the predictions often lack interpretability and do not outperform established task-specific methods in accuracy. Although LLMs hold immense promise for organic reaction prediction, owing to their pre-trained chemical knowledge as well as robust in-context learning and reasoning capabilities, two critical bottlenecks must still be addressed before this potential can be fully realized.

First, high-quality, structured training data is severely scarce in chemistry. Domains like mathematics benefit from extensive open-source communities (e.g., Lean Community) and web-scale datasets, whereas chemistry lacks publicly available resources for reaction reasoning. As a result, models undergo pre-training and fine-tuning with limited exposure to task-relevant chemical data, hindering their ability to develop advanced capabilities for complex reaction prediction tasks. Furthermore, creating large-scale, annotated chemical datasets is a time-consuming and labor-intensive process that demands substantial domain expertise.

Second, learning strategies for chemistry LLMs remain underdeveloped. Most existing chemical LLM frameworks rely on standard pre-training followed by supervised fine-tuning (SFT), which often fails to unlock the full potential of these models. Recent research highlights two advanced techniques, including Retrieval-Augmented Generation (RAG) Ke et al. (2024); Chen et al. (2025), which can inject domain-specific knowledge as well as mitigating hallucinations, and reinforcement learning (RL) Guo et al. (2025), which can further enhance reasoning and interpretability. However, their adoption in chemical LLMs remains limited. Therefore, developing a hybrid learning framework that integrates SFT, RAG and RL, combining both data-driven and knowledge-driven paradigms, represents a promising direction for achieving interpretable, high-performance organic reaction prediction.

In this paper, we propose **Reaction-Thinker**, a hybrid knowledge-and-data-driven organic reaction prediction system, comprising both a RAG-based predictor and a reasoning-based predictor. The main contributions of our work can be concluded in the following parts.

- We categorize reactions based on the given reaction inputs, define a standardized similarity metric, and construct similar-case retrieval database for each type. Training and test samples are partitioned based on whether similar retrieved cases exist, and each subset is processed through a dedicated, specialized pipeline.
- For samples with similar cases retrieved, we inject reaction type and case-specific knowledge into user prompt and curate a customized SFT dataset for a RAG-based LLM. This enhances the ability of the LLM to retrieve domain-specific contextual information.
- For samples lacking similar reaction cases, we introduce a multi-stage reasoning enhancement pipeline. First, we construct a chain-of-thought (CoT) dataset for organic reaction reasoning. Then, we employ SFT as a cold-start to establish an initial foundation of high-accuracy CoT reasoning. Finally, we refine the deductive reasoning using reinforcement learning through Group Relative Policy Optimization (GRPO).

- The system outperforms all compared LLMs and even exceeds traditional task-specific models, in both accuracy (Exact Match) and fingerprint similarity (FTS). The ablation study indicates improvements in relative accuracy of 7.5% and 13.9% for RAG and GRPO, respectively.
- A detailed error analysis reveals that some incorrect predictions correspond to chemically plausible byproducts or alternative reaction pathways, despite not matching the canonical ground truth. To better account for such chemically plausible outcomes, we propose a novel evaluation paradigm by incorporating retrosynthetic validation. Notably, our analysis indicates that 47.8% of these incorrect predictions are chemically reasonable.

2 METHODS

As illustrated in Figure 1, our proposed organic reaction predict system integrates four core modules: (1) a reaction type classifier, (2) a similar-case retrieval database, (3) a RAG-based reaction predictor, and (4) a reasoning-based reaction predictor. Given a set of reaction inputs including reactants, solvent, and reagents, the system first applies the classifier to determine the most probable reaction type. Based on the predicted type, it queries the similar-case retrieval database for analogous reaction examples. If similar reaction cases are found, they are then incorporated alongside the user prompt into the RAG-based predictor; otherwise, the reaction inputs are routed directly to the reasoning-based reaction predictor for CoT-based analysis. Depending on the pathway, either RAG-enhanced or reasoning-based module generates the final reaction outcome.

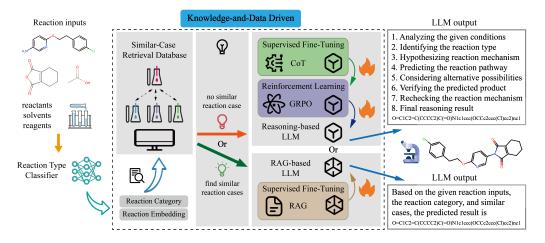


Figure 1: The system architecture, training process, and inference pipeline of **Reaction-Thinker**.

Subsequent sections detail the training process of reaction type classifier, the construction and usage of the similar-case retrieval database, the preparation of a CoT dataset for organic reaction reasoning, and the training strategies for both RAG-based and reasoning-based predictors.

2.1 REACTION TYPE CLASSIFIER

We implement a two-layer MLP as the classifier. The original reaction inputs, provided as SMILES strings, are processed with RDKit Landrum (2016) to compute multiple structural fingerprints. These fingerprints (*Mol-Fingerprint*) are then concatenated and fed into the MLP to predict reaction type (*Classifier-Out*).

Inspired by previous work Safizadeh et al. (2021), we employ a combination of various molecular fingerprint methods to comprehensively capture molecular information, including RDK (suitable for general molecular similarity searches) Schneider et al. (2015b), LAYERED (useful for substructure screening) RDKit-Book (2025a), PATTERN (focused on identifying specific chemical features) RDKit-Book (2025b), AVALON (effective for both substructure screening and similarity matching in complex molecules) Gedeck et al. (2006), and MORGAN fingerprints (especially suitable for cyclic substructure and comparing structural features) Rogers & Hahn (2010).

We train the reaction type classifier on the Schneider-50K dataset Schneider et al. (2015a), which contains 50K reaction SMILES classified into 50 representative types, providing granularity well-suited for robust classification. After training, we extract the first layer output of the classifier as a compact representation of the reaction inputs, providing a molecular embedding (*Rea-Embedding*) of the reaction. The architecture of the reaction type classifier is described in Equation (1):

$$Rea-Embedding = Layer1 (Mol-Fingerprint)$$

$$Classifier-Out = Layer2 (Rea-Embedding)$$
(1)

2.2 SIMILAR-CASE RETRIEVAL DATABASE FOR RAG

For each reaction in both training and test splits of Open Reaction Dataset (ORD) Kearnes et al. (2021), we feed its reaction inputs (including reactants, solvents, and reagents, all formatted as SMILES) into our trained reaction type classifier. This yields two outputs: (1) a molecular embedding from the first layer, and (2) a predicted reaction type.

Using the ORD training set, we construct a similar-case retrieval library for each reaction type by computing the Euclidean distance (L_2 norm) between the molecular embedding of each query reaction and that of all other embeddings of the same type. Any training sample with an embedding distance smaller than M is considered similar, and its full reaction SMILES (including reactants, solvents, reagents, and products) will be added to the retrieval library for that type.

Based on this, we build a SFT dataset for RAG-based LLM, containing only reactions that successfully retrieve at least one similar case. Each SFT sample includes the reaction inputs, its predicted reaction type, the retrieved similar cases, and the target products.

During inference, we apply the same embedding and classification process to reaction inputs from the test set. If the system retrieves one or more training cases within distance M, it follows the RAG pathway; otherwise, it proceeds to the reasoning pathway without external context.

2.3 COT DATASET FOR REACTION REASONING



Figure 2: An example of chain-of-thought dataset for reasoning, including system prompt, user prompt, and supervised output.

Here a two-stage approach is adopted to generate CoT data for reaction reasoning, involving both the USPTO-MIT Jin et al. (2017) and ORD. Finally, we merge CoT samples obtained from both stages into a unified dataset, serving as the primary CoT resource for this work. Some data are open-sourced (refer to link and details in **Appendix**).

2.3.1 STAGE 1: PRELIMINARY CONSTRUCTION

Following the methodology in HK-O1aw Lab (2024), we extract reaction SMILES from a random subset of the USPTO-MIT training set. These SMILES were processed using Qwen2.5-72B-Instruct Qwen Team (2024) with instructions to reconstruct the reaction mechanism through deductive reasoning, systematically deriving the products from the given reactants, solvents, and reagents through a chain-of-thought reasoning process, even though the model has access to the final answer (refer to **Appendix** for the prompt details). We then apply rigorous post-processing to refine the generated contexts, including format standardization and keyword-based validation, ultimately obtaining 119K high-quality CoT samples after careful filtering. An example of the dataset is shown in Figure 2. Although directly predicting reaction outcomes from conditions is challenging, our approach leverages the observation that when provided with full reaction SMILES, LLMs can systematically deduce the reaction pathway by analyzing the transformation from reactants to products.

2.3.2 STAGE 2: DISTILLATION AND VALIDATION

We first fine-tune DeepSeek-R1-Distill-Qwen-7B DeepSeek-AI (2025c) on the CoT samples generated from USPTO-MIT using SFT. Then, we further train the model on the ORD training set using Group Relative Policy Optimization (GRPO) (see experimental results in **Appendix**). The GRPO will be explained in detail in the following section. During this stage, only those generated reasoning trajectories that lead to correct predicted products are retained. Overall, we collected 575K validated CoT examples, covering approximately 55K unique samples from the original ORD dataset.

2.4 Training Strategy of RAG-based LLM

2.4.1 Supervised Fine-Tuning

We fine-tune RAG-based LLM on the dataset constructed from similar-case retrieval database, using SFT with full parameter updates. The Qwen3-32B Qwen Team (2025) is selected as the backbone model for this process.

2.5 Training Strategy of Reasoning-based LLM

We employ a two-stage training strategy (SFT followed by RL) for our reasoning-based LLM, using DeepSeek-R1-Distill-Qwen-32B DeepSeek-AI (2025b) as the backbone model due to its strong reasoning performance compared to other LLMs of similar size.

2.5.1 Supervised Fine-Tuning

First, we fine-tune the base model using SFT with full parameter updates on the generated CoT dataset for reaction reasoning. Through this process, the model begins to internalize reasoning patterns specific to organic reaction.

2.5.2 Reinforcement Learning

Next, we perform RL with LoRA adapters Hu et al. (2022) on the ORD training set to further enhance reasoning accuracy as well as reliability.

Specifically, we use GRPO as the learning algorithm. Given an input query $q \sim P(Q)$, GRPO samples a group of G responses $\{y_1, y_2, \ldots, y_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$. The core idea is to update the policy π_{θ} by maximizing an objective function that encourages responses with higher-than-average rewards within their group. The GRPO objective function is defined as follows:

$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim P(Q), \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \right]$$

$$\min \left(c_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left(c_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right) - \beta \, \mathbb{D}_{KL} \left[\pi_{\theta} \parallel \pi_{\text{ref}} \right]$$

$$(2)$$

where ϵ is the clipping ratio, β is the coefficient for KL-divergence loss, and $\pi_{\rm ref}$ is the reference policy. $c_{i,t}(\theta)$ is the importance sampling ratio for token $y_{i,t}$ (the t-th token of the i-th response y_i):

$$c_{i,t}(\theta) = \frac{\pi_{\theta}(y_{i,t}|q, y_{i,< t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|q, y_{i,< t})}$$

 $\overline{)}$ (3)

 $\hat{A}_{i,t}$ is advantage estimate for all tokens in response y_i and is calculated by normalizing the rewards $\{r_1, r_2, \dots, r_G\}$ using the group mean and standard deviation:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}$$
(4)

2.5.3 REWARD FUNCTIONS

Here we design a custom reward function for RL, specifically tailored to organic reaction reasoning, composed of four components:

• **Format Reward:** Assess whether the response format strictly follows the user instructions, awarding 0.1 for correct compliance; or otherwise zero.

• **Length Reward:** Encourages concise reasoning, awarding 0.1 if the chain-of-thought length falls within a predefined range (500 to 2000 tokens); or otherwise zero.

• Validity Reward: Assess the validity of the generated product SMILES, awarding 0.1 for chemically valid; or otherwise zero.

• **Accuracy Reward:** Canonicalize the SMILES of generated product and compare it to the ground truth. Award 2.0 if the two match exactly; or otherwise zero.

The final reward is calculated as the sum of these components. This composite reward structure ensures that the model is incentivized to produce well-formed, appropriately concise, and chemically accurate reasoning.

3 EXPERIMENTS

In this section, we first present a comprehensive comparative evaluation between our method and existing baselines. Next, we evaluate the contribution of RAG. We also conduct an ablation study to assess how variations in reward function design and the use of cold-start strategy affect GRPO performance in this task. Finally, through detailed analysis of mispredictions, we identify critical limitations in conventional evaluation metrics. Leveraging these insights, we propose a novel reference metric and present its evaluation result. Additional details are provided in **Appendix**.

3.1 Experimental Setting

3.1.1 Dataset and Metrics

The raw Open Reaction Database (ORD) has been preprocessed by ORDerly Wigh et al. (2024) and split into 832K for training and 86K for testing. For evaluation, we employ various metrics including Validity (whether the product SMILES can be successfully processed by RDKit), Exact Match (after canonicalization) and molecular fingerprint similarity (FTS, including MORGAN, RDK, and AVALON fingerprints). We deliberately avoid relying on text-based similarity metrics (e.g., BLEU and Levenshtein), since they poorly reflect actual molecular differences, even a single alteration in a SMILES string can correspond to a substantial change in the chemical structure.

3.1.2 Baselines

We compare our system with: (1) open-source LLMs including DeepSeek-R1 Guo et al. (2025), DeepSeek-R1-Distill-Llama-70B DeepSeek-AI (2025a), DeepSeek-R1-Distill-Qwen series (32B/14B/7B), and Qwen2.5-72B-Instruct; (2) chemical LLMs including ChemDFM-13B/8B OpenDFM Team (2024) and Text-Chem-T5 Christofidellis et al. (2023); and (3) traditional task-specific models including Chemformer, which is reported to achieve the best Top-1 accuracy on USPTO-MIT Chen & Jung (2022), and Molecular Transformer Schwaller et al. (2019).

For open-source LLMs, we use the same user prompt template as our method, and for chemical LLMs, we adopt the training prompts specified in the relevant papers.

3.2 MAIN RESULT

As shown in Table 1, our method outperforms all compared LLMs and traditional task-specific models, achieving the highest Exact Match and fingerprint similarity (FTS).

Table 1: Comparison of our method with various baselines on the task of organic reaction prediction. The results for Molecular Transformer are directly taken from existing work Wigh et al. (2024). The top results are marked in **bold**.

Model	Model Type	Validity (%)	Exact Match (%)	FTS (%)		
				MORGAN	RDK	AVALON
Chemformer	Task-Specific Model	98.57	88.13	92.40	94.35	95.12
Molecular Transformer		99.66	85.84	-	-	-
DeepSeek-R1	General LLM	68.54	11.68	55.71	64.09	64.98
Qwen2.5-72B-Instruct		35.27	0.54	37.46	45.95	46.01
DeepSeek-R1-Distill-Llama-70B		66.42	7.20	49.06	59.17	59.67
DeepSeek-R1-Distill-Qwen-32B		57.58	6.52	50.50	61.05	61.05
DeepSeek-R1-Distill-Qwen-14B		43.72	1.69	32.72	40.92	40.95
DeepSeek-R1-Distill-Qwen-7B		47.74	1.21	43.50	54.58	53.77
ChemDFM-13B		98.29	52.41	77.27	82.03	82.15
ChemDFM-8B	Chemical LLM	97.80	48.02	74.69	79.85	80.02
Text-Chem-T5		95.67	47.88	76.45	81.81	81.81
Reaction-Thinker (Ours)	_	98.92	89.86	95.22	96.24	96.37

The final performance of our method is achieved through integration. In the test set, 81.7% of the samples have similar reaction cases available. For these, our RAG-based approach achieves an Exact Match of 94.70%. For the remaining 18.3% samples without similar cases, we apply reasoning-based approach and achieve an Exact Match of 68.24%. Combining these two approaches yields an overall accuracy of 89.86% across the entire test set. The FTS score is computed using the same weighted approach, combining the RAG-based and reasoning-based results according to their respective proportions.

3.3 ABLATION STUDY

3.3.1 CONTRIBUTION OF RAG

The effectiveness of RAG has been widely documented in recent works. By grounding generation with retrieved context, it significantly reduces hallucinations and improves accuracy across many domains. To verify the benefit of RAG for current task, we continue to use Qwen3-32B as the base model and conduct a controlled comparison. Rather than following the conventional RAG setup that retrieves reaction types and similar cases for prompt augmentation, we perform a direct end-to-end supervised fine-tuning, mapping reaction input SMILES directly to product SMILES.

As shown in Table 2, using RAG yields better performance, with a relative accuracy improvement of 7.5%. This matches chemical intuition: just as chemists reference analogous reactions, LLMs benefit from RAG to improve prediction accuracy.

Table 2: Accuracy performance with and without RAG.

w/ RAG w/o RAG (End-to-End)

Exact Match (%) 83.13 77.35

3.3.2 Influence of GRPO

GRPO is an effective reinforcement learning framework for LLMs, where the design of reward functions and selection of the base model critically determine its performance. To further explore its application in current task, we conduct two controlled experiments.

1. Reward Function Ablation Study

Building on the baseline reward function (comprising format, length, validity, and accuracy rewards), we introduce a MORGAN fingerprint similarity reward (FTS reward, ranging from 0.0 to 1.0) in order to structurally align predictions with ground truths.

This modification aims to mitigate reward sparsity by guiding the LLM to generate outputs from structurally similar to fully accurate. We record the reward curves during training in Figure 3 and evaluate the resulting model with results presented in Ta-The experiments reveal a paradoxical phenomenon, while the reward curve shows continuous improvement, the evaluation accuracy actually declines. Through detailed analysis, we identify that the finetuned LLM tended to verbatim copy reactant SMILES in outputs. This indicates a suboptimal optimization strategy, since product and reactants structures share chemical similarities, directly copying reactants could still achieve relatively high reward.

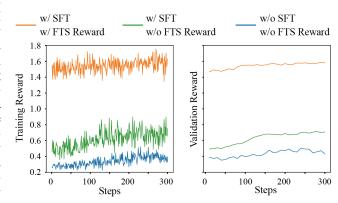


Figure 3: The reward curves under different combinations of (i) whether SFT was applied before RL and (ii) whether the FTS reward is introduced.

Table 3: Accuracy performance with and without fingerprint similarity reward in GRPO.

	w/ FTS reward	w/o FTS reward
Exact Match (%)	56.83	68.24

To address this issue, we downweight the FTS reward and incorporate explicit penalties for reactant copying. However, subsequent experiments demonstrate these measures are insufficient to completely prevent this behavior. This phenomenon exemplifies reward hacking, a well-documented RL failure mode where agents optimize the proxy reward in unintended ways, achieving higher scores while failing the true task objective. The tendency of LLM to cheat by exploiting structural correlations between reactants and products remains a significant challenge, now recognized as a critical focus for our ongoing optimization efforts.

2. Base Model Capability Analysis

To evaluate the impact of base model capability on GRPO performance, we conduct a controlled experiment comparing two approaches: (1) direct application of GRPO to the initial DeepSeek-R1-Distill-Qwen-32B, and (2) implementing GRPO following SFT on our reaction reasoning dataset.

The experimental results, including the reward curves in Figure 3 and final performance in Table 4, reveal markedly different outcomes between the two settings. This underscores that proper initialization through SFT is critical for unlocking the potential of GRPO in reaction reasoning tasks.

Table 4: Exact Match performance (%) under different SFT and GRPO settings. On the base model with enhanced initial reasoning capabilities (w/ SFT), applying GRPO yields a relative accuracy improvement of 13.9%.

	w/o SFT	w/ SFT
w/o GRPO	6.52	59.93
w/ GRPO	9.67	68.24

3.4 Analysis of Incorrect Predictions

Through systematic error analysis comparing LLM predictions with ground truth, we identified two major failure modes: (1) complex reactions involving multiple functional groups or multi-step processes, and (2) incomplete reaction conditions (e.g., missing temperature or catalysts, which is confirmed by GPT-40). Figure 4 presents an example of incorrect prediction along with detailed analysis using GPT-40 and retrosynthetic validation.

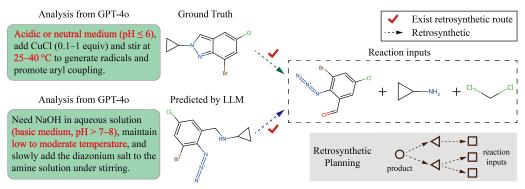


Figure 4: Analysis of incorrect prediction using GPT-40 and retrosynthetic validation. The key reaction conditions influencing final product are marked in bold red. Both products are valid candidate answers in the absence of specific condition constraints.

Fundamentally, many organic reactions inherently generate byproducts via parallel or competing pathways, yet existing datasets typically record only one to three major products. This exposes a critical limitation in current evaluation metrics for reaction prediction tasks, where exclusive comparison to a single ground truth fails to reflect chemical reality and may hinder LLMs from developing a genuine understanding of organic reaction mechanisms.

To address this, we propose a novel evaluation paradigm by incorporating retrosynthetic validation. For each product predicted by reasoning-based LLM, we verify whether a plausible retrosynthetic route exists based on the given reaction inputs. If chemically reasonable, a prediction is deemed correct even if it does not match the ground truth. Applying the retrosynthetic analysis tool Retro* Chen et al. (2020) to our previously mispredicted examples, 47.8% of them are validated as chemically reasonable, increasing our final Exact Match score from 89.86% to 92.64%.

4 Conclusion

In this study, we introduce **Reaction-Thinker**, a hybrid, knowledge-and-data-driven system that significantly advances organic reaction prediction by combining RAG-based LLM with reasoning-based LLM. Experiments on ORD demonstrate that our system achieves state-of-the-art (SOTA) results in Exact Match and fingerprint similarity, outperforming all compared LLMs and task-specific models. The result shows the potential of leveraging LLMs for addressing fundamental challenges in chemical research. We also identify several promising directions for future enhancement. For example, the reasoning-based LLM still has significant room for improvement. The optimization of reward functions will be our focus in further research to help LLMs have better understanding of organic reaction mechanisms. Moreover, enhanced CoT datasets incorporating chemical synthesis processes will be developed and integrated into current training framework, enabling more rigorous analysis of how reaction condition variations affect outcomes. Last but not least, current system implements RAG and reasoning as separate LLM modules, and future work will integrate these capabilities into a unified architecture.

REPRODUCIBILITY STATEMENT

To ensure reproducibility and facilitate the review process, we have included a subset of the code and a dataset sample in the **Appendix**. The full code and complete dataset will be made publicly available upon acceptance.

ETHICS STATEMENT

This research complies with the ICLR Code of Ethics. Our study uses publicly available benchmark data and does not involve human subjects or collection of sensitive information. The authors declare no potential conflicts of interest or sponsorship that could influence the work reported in this paper.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.
- Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: learning retrosynthetic planning with neural guided a* search. In *International conference on machine learning*, pp. 1608–1616. PMLR, 2020.
- Lun-Chi Chen, Mayuresh Sunil Pardeshi, Yi-Xiang Liao, and Kai-Chih Pai. Application of retrievalaugmented generation for interactive industrial knowledge management via a large language model. *Computer Standards & Interfaces*, 94:103995, 2025.
- Shuan Chen and Yousung Jung. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4(9):772–780, 2022.
- Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning*, pp. 6140–6157. PMLR, 2023.
- DeepSeek-AI. Deepseek-rl-distill-llama-70b. https://huggingface.co/deepseek-ai/DeepSeek-Rl-Distill-Llama-70B, 2025a. Distilled Llama-3.3-70B model optimized for reasoning.
- DeepSeek-AI. Deepseek-r1-distill-qwen-32b. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B, 2025b. Distilled Qwen2.5-32B model optimized for reasoning.
- DeepSeek-AI. Deepseek-r1-distill-qwen-7b. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B, 2025c. Distilled reasoning model based on Qwen2.5-7B.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.
- Peter Gedeck, Bernhard Rohde, and Christian Bartels. Qsar- how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of chemical information and modeling*, 46(5):1924–1936, 2006.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

- Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. arXiv preprint arXiv:2410.12837, 2024.
 - Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.
 - Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
 - Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pretrained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.
 - Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023.
 - Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.
 - YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, and Daniel Shu Wei Ting. Development and testing of retrieval augmented generation in large language models—a case study report. *arXiv preprint arXiv:2402.01733*, 2024.
 - Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.
 - HKAIR Lab. Hk-olaw models: Leveraging ol slow thinking in the development of hong kong legal large language models. https://github.com/HKAIR-Lab/HK-Olaw, 2024.
 - Greg Landrum. Rdkit: Open-source cheminformatics software. https://www.rdkit.org, 2016. Version as of your usage date.
 - Daniel E Levy. Arrow-pushing in organic chemistry: an easy approach to understanding reaction mechanisms. John Wiley & Sons, 2017.
 - Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.
 - Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.
 - Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- 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, 2024.
- OpenAI. Gpt-4o: Multilingual, multimodal foundation model. https://openai.com/index/hello-gpt-4o/, 2024. Released May 2024 (real-time text, vision, and audio inference).
 - OpenDFM Team. Chemdfm-13b/8b. https://github.com/OpenDFM/ChemDFM, 2024. Open-source chemical LLM models.

- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.
- Qwen Team. Qwen2.5-72b-instruct. https://huggingface.co/Qwen/Qwen2.5-72B-Instruct, 2024. Instruction-tuned Qwen2.5 model.
 - Qwen Team. Qwen3-32b. https://huggingface.co/Qwen/Qwen3-32B, 2025. Large language model with RAG-friendly architecture.
 - RDKit-Book. RDKit: Open-source cheminformatics. https://www.rdkit.org/docs/RDKit_Book.html, 2025a. Accessed: 2025-08-01.
 - RDKit-Book. RDKit: Open-source cheminformatics. https://www.rdkit.org/docs/RDKit_Book.html, 2025b. Accessed: 2025-08-01.
 - David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
 - Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.
 - Hamid Safizadeh, Scott W Simpkins, Justin Nelson, Sheena C Li, Jeff S Piotrowski, Mami Yoshimura, Yoko Yashiroda, Hiroyuki Hirano, Hiroyuki Osada, Minoru Yoshida, et al. Improving measures of chemical structural similarity using machine learning on chemical–genetic interactions. *Journal of chemical information and modeling*, 61(9):4156–4172, 2021.
 - Nadine Schneider, Daniel M Lowe, Roger A Sayle, and Gregory A Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of chemical information and modeling*, 55(1):39–53, 2015a.
 - Nadine Schneider, Roger A Sayle, and Gregory A Landrum. Get your atoms in order-an open-source implementation of a novel and robust molecular canonicalization algorithm. *Journal of chemical information and modeling*, 55(10):2111–2120, 2015b.
 - John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
 - Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018.
 - Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.
 - Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv* preprint arXiv:2305.15294, 2023.
 - Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
 - David K Smith. Priority and selectivity rules to help students predict organic reaction mechanisms. *Journal of Chemical Education*, 100(3):1164–1178, 2023.
 - Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv* preprint arXiv:2212.10509, 2022.
- Daniel S Wigh, Joe Arrowsmith, Alexander Pomberger, Kobi C Felton, and Alexei A Lapkin. Orderly: data sets and benchmarks for chemical reaction data. *Journal of Chemical Information and Modeling*, 64(9):3790–3798, 2024.
- Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.
- Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pp. 445–460. Springer, 2024.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.
- Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Yansi Li, Zhongyang Dai, et al. Chemdfm-x: towards large multimodal model for chemistry. *Science China Information Sciences*, 67(12):1–2, 2024a.
- Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. Chemdfm: Dialogue foundation model for chemistry. *arXiv e-prints*, pp. arXiv–2401, 2024b.

A APPENDIX

A.1 RELATED WORK

A.1.1 RETRIEVAL-AUGMENTED GENERATION IN LLM

RAG can be an effective paradigm for infusing LLMs with non-parametric knowledge Gao et al. (2023); Gupta et al. (2024); Li et al. (2025), with demonstrated impact in knowledge-intensive domains such as medicine Li et al. (2024) and law Wiratunga et al. (2024). By retrieving and conditioning on external documents, RAG significantly improves performance on generation tasks. RAG methods can be broadly categorized into single-round and multi-round strategies. Basic RAG approaches typically retrieves knowledge based solely on the initial query Guu et al. (2020); Borgeaud et al. (2022). Some works have also explored multi-round retrieval strategies that iteratively refine or rewrite queries across steps Shao et al. (2023); Jiang et al. (2023), interleave retrieval with reasoning Trivedi et al. (2022), or utilize multi-stage self-asking mechanisms Press et al. (2022). Depending on the task, either single-round or multi-round retrieval strategies can be employed.

A.1.2 REINFORCEMENT LEARNING FOR COT REASONING

Chain-of-Thought (CoT) reasoning Trivedi et al. (2022) represents a significant methodological advancement in enhancing the reasoning capabilities of LLMs. This approach prompts models to explicitly generate intermediate reasoning steps before arriving at a final output. Such structured reasoning processes substantially improve prediction accuracy, as higher-quality intermediate contexts often contribute to more reliable and consistent final results. Reinforcement learning (RL) has also emerged as a powerful technique for improving the reasoning ability of LLMs, particularly in domains such as mathematics, where structured reward signals allow models to learn beyond what SFT alone can achieve. Recent developments have introduced RL frameworks with numerical feedback, often relying on online policy optimization algorithms such as Proximal Policy Optimization Schulman et al. (2017), Group Relative Policy Optimization (GRPO) Shao et al. (2024), REINFORCE Williams (1992), and Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) Yu et al. (2025).

A.2 SUPPLEMENTARY MATERIAL

An anonymous link¹ has been provided to open-source the code and data of our work.

A.2.1 CHAIN-OF-THOUGHT DATASET

Supplementary to Section CoT Dataset for Reaction Reasoning.

Data Generation Script

The data generation script and associated prompt template for Section **Stage 1: Preliminary Construction** in the main text, has been made publicly available in *CoT-Gen.py* and *User-Prompt.txt*, respectively.

Open-Source Samples

A curated subset of 100 randomly selected examples is provided in *Open-Source.jsonl* for demonstration purposes. The full dataset, consisting of 119K (from USPTO-MIT) and 575K (from ORD) reaction reasoning samples, will be released upon the official acceptance of this paper.

A.2.2 Training DeepSeek-R1-Distill-Qwen-7B

Supplementary to Stage 2: Distillation and Validation of CoT Dataset for Reaction Reasoning in the main text. We document the experimental setting and training progress of DeepSeek-R1-Distill-Qwen-7B during the CoT dataset construction process.

https://anonymous.4open.science/r/AI4Chem-854A

Experimental Settings and Results

 The implementation details are specified in the scripts SFT-DeepSeek-7B.sh and GRPO-DeepSeek-7B.sh. The experiments are run on 8 NVIDIA A800 GPUs.

We record the reward curves during GRPO in Figure 5. We also evaluate the final models after SFT and GRPO, with detailed results presented in Table 5. This is a preliminary study to validate the integration of LLMs and RL for organic reaction prediction. This pilot study, which also yields a dataset of CoT reasoning traces, provides compelling evidence for the viability of our method. However, anticipating the limitations of a 7B model, we proceed with a more powerful DeepSeek-R1-Distill-Qwen-32B in the main experiment.

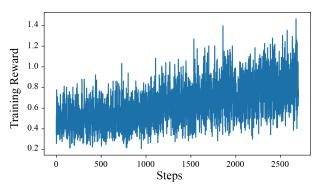


Figure 5: The reward curve during GRPO training for DeepSeek-R1-Distill-Qwen-7B.

Table 5: The Validity and Exact Match performance of DeepSeek-R1-Distill-Qwen-7B on the USPTO-MIT test set after SFT and GRPO, respectively.

	Validity (%)	Exact Match (%)
After SFT	86.4	22.4
After GRPO	91.9	35.9

A.2.3 TRAINING REACTION TYPE CLASSIFIER

Supplementary to Section Reaction Type Classifier.

Experimental Settings and Results

The training and test code, as well as parameter configurations are implemented in *Classifier.py* file.

We randomly shuffled the raw Schneider-50K dataset and split it into a 40K training set and a 10K validation set. The final classifier achieves a Top-1 accuracy of 94.35% on the validation set.

A.2.4 TRAINING RAG-BASED LLM

Experimental Settings

The implementation details are specified in RAG-Qwen-32B.sh. All experiments are run on 8 NVIDIA A800 GPUs.

A.2.5 Training Reasoning-based LLM

Experimental Settings

The SFT implementation details are specified in SFT-DeepSeek-32B-1.sh, SFT-DeepSeek-32B-2.sh. The GRPO implementation details are specified in GRPO-DeepSeek-32B-1.sh and GRPO-DeepSeek-32B-2.sh. All experiments are run on 16 NVIDIA A800 GPUs.