A Modular Framework for Automated Hypothesis Validation and Refinement in Scientific Research

Chenhao Chen *, Taiga Masuda, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi

Posted Date: 19 January 2026

doi: 10.20944/preprints202601.1274.v1

Keywords: hypothesis validation; hypothesis refinement; natural language inference; retrievalaugment generation; LLM-based scientific workflow

Article

A Modular Framework for Automated Hypothesis Validation and Refinement in Scientific Research

Chenhao Chen [1,*], Taiga Masuda [1], Tsubasa Hirakawa [2], Takayoshi Yamashita [3] and Hironobu Fujiyoshi [1]

[1] Department of Artificial Intelligence and Robotics, Chubu University, Kasugai-shi, Aichi 487-8501, Japan
[2] Center for Mathematical Science and Artificial Intelligence, Chubu University, Kasugai-shi, Aichi 487-8501, Japan
[3] Department of Computer Science, Chubu University, Kasugai-shi, Aichi 487-8501, Japan
* Correspondence: chinshinkou@mprg.cs.chubu.ac.jp

Abstract

Scientific research typically follows an iterative cycle where hypotheses are proposed, validated against experimental conclusions, and refined accordingly. While recent advances in large language models (LLMs) have enabled significant progress in automating individual stages of this process, existing systems are typically developed as standalone solutions, making it difficult to coordinate multiple research activities within a coherent research workflow. In this study, we present a modular framework for automated hypothesis validation and refinement in scientific research. Rather than introducing new task-specific models, the framework integrates established techniques, including natural language inference (NLI)-based hypothesis validation, attribution-guided hypothesis refinement, and retrieval-augment generation (RAG)-based external evidence retrieval, into a unified and controllable workflow. We evaluate the proposed framework on scientific texts in the chemistry domain to assess its applicability in practical scientific research scenarios. Extensive experiments demonstrate the effectiveness of the proposed framework and suggest that it produces reliable intermediate signals that enhance transparency and traceability throughout hypothesis validation and refinement. Our work offers a modular solution for deploying LLM-based systems into scientific research workflows.

Keywords: hypothesis validation; hypothesis refinement; natural language inference; retrieval-augment generation; LLM-based scientific workflow

1. Introduction

Traditionally, scientific discovery requires human researchers to collect background knowledge, draft initial hypotheses, construct evaluation procedures, assess evidence, and refine their hypotheses accordingly. However, this iterative process of hypothesis formulation, validation, and refinement is inherently limited by human researchers' ingenuity [1]. As the scale of domain-specific knowledge expands continuously, researchers face increasing challenges in efficiently advancing this scientific research workflow. Early efforts to automate scientific discovery focused on providing computer-assisted support for specific stages of the scientific process, such as Automated Mathematician [2,3] and DENDRAL [4]. With the rapid development of artificial intelligence (AI), large language models (LLMs) have demonstrated remarkable capabilities in understanding and reasoning over scientific texts, which introduce a paradigm shift to individual stages of scientific research workflows. For example, LLMs have been applied to scientific literature understanding [5,6], hypothesis generation [1], and experimental planning [7]. Despite these advances, existing approaches for scientific research remain task-oriented. That means most systems are independently designed for specific research activities, without explicitly modeling the interactions and dependencies among multiple stages within a coherent research workflow. Additionally, many LLM-based systems adopt end-to-end architectures that restrict user interaction and provide limited transparency on decision-making and reasoning process.

 doi:10.20944/preprints202601.1274.v1

To address these limitations, in this work, we present a modular framework for automated hypothesis validation and refinement in scientific research. Rather than introducing new task-specific models, the framework systematically integrates hypothesis validation and hypothesis refinement as two interconnected yet decoupled components within a unified scientific research workflow. An overview of the proposed framework is illustrated in Figure 1, where all components are instantiated using existing models and techniques. As shown in the figure, the framework is organized around a common scientific reasoning cycle and consists of three core components: hypothesis validation, hypothesis refinement, and external evidence retrieval.

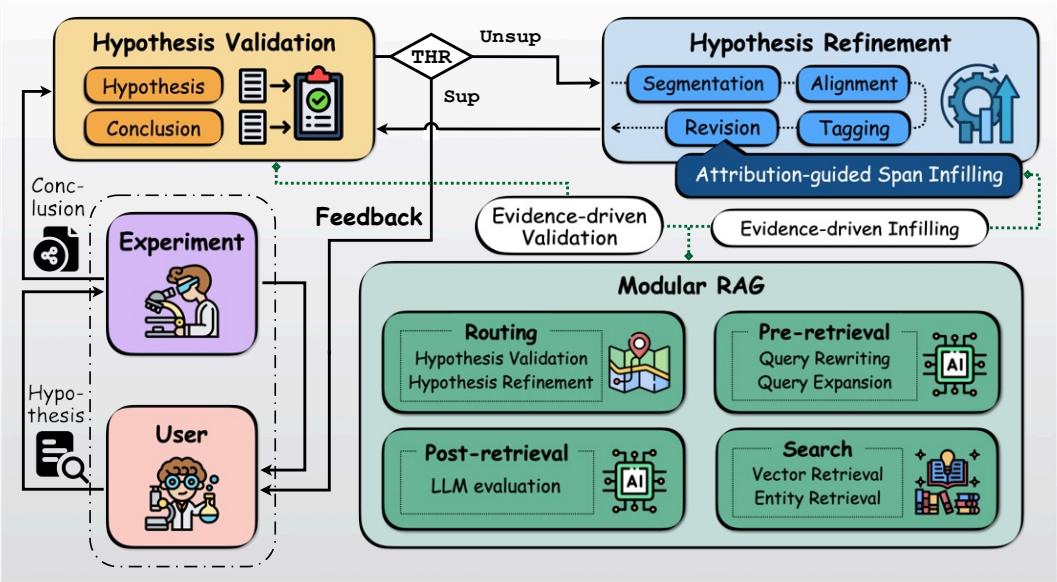

Figure 1. An overview of the proposed modular framework for automated hypothesis validation and refinement. This framework consists of three modules: NLI-based hypothesis validation, attribution-guided hypothesis refinement, and RAG-based external evidence retrieval.

Specifically, given a hypothesis proposed by a researcher and a conclusion drawn from experiments, the hypothesis validation module assesses the logical consistency between the hypothesis and conclusion using a natural language inference (NLI) model to provide validation decisions with confidence scores for downstream reasoning. If the hypothesis is assessed as unsupportive (including contradictory and inconclusive), the hypothesis refinement module is activated to identify and revise hypothesis spans that contribute to the semantic inconsistency. This refinement process is guided by attribution information derived from the NLI model, enabling targeted edits rather than unconstrained full-text rewriting. To support both validation and refinement process with external scientific knowledge, the framework incorporates an evidence retrieval module based on retrieval-augment generation (RAG) [8]. This module serves as a external knowledge base to provide relevant domain-specific evidence that can be leveraged during the process of hypothesis validation and refinement.

The proposed framework offers a modular solution for deploying LLM-based systems into scientific research workflows. It is designed to assist researchers by providing intermediate research signals such as validation decisions, attribution information, and retrieved external knowledge, enabling that researchers maintain control and oversight over the research workflow. We evaluate the proposed framework on scientific texts in the chemistry domain for its rich domain-specific terminology and structured experimental reasoning in practical scientific research scenarios.

2. Framework Overview

The proposed framework aims to provide a coherent and controllable reasoning process for hypothesis validation and refinement in scientific research. Rather than focusing on optimizing individual models, the framework emphasizes the organization and coordination of multiple reasoning

 doi:10.20944/preprints202601.1274.v1

components within a unified research workflow. As shown in Figure 1, the proposed framework models the iterative process of scientific discovery where the initial hypothesis is continuously validated and refined based on experimental conclusions. From a macro perspective, the framework adopts a modular architecture composed of three core components: hypothesis validation, hypothesis refinement, and external evidence retrieval. Each component is implemented as an independent module with clearly defined inputs and outputs, making the overall workflow flexible and interpretable. The framework takes as input a hypothesis proposed by a researcher and a conclusion derived from experimental results. Its primary objective is to evaluate the logical relationship between these inputs and revise contradictory spans within the hypothesis in a controlled manner. Next, we would like to give a detailed introduction on the proposed framework.

2.1. Hypothesis Validation as the Decision Anchor

The hypothesis validation module serves as the central decision anchor within the proposed framework. It is responsible for assessing whether a given hypothesis is logically supported by an experimental conclusion, thereby establishing a reliable basis for subsequent reasoning stages. Given a hypothesis-conclusion pair (h, c), the hypothesis validation module assesses their logical relationship and estimates a probability distribution $\boldsymbol{p}_{\text{val}}$ over two validation class: supportive or unsupportive.

$$\boldsymbol{p}_{\text{val}} := \big(p(y \mid h, c)\big)_{y \in \{\text{supportive,unsupportive}\}} \tag{1}$$

Here, supportive corresponds to *entailment* in standard NLI tasks, while unsupportive covers *contradiction* or, depending on the NLI setting, *neutral* cases where the hypothesis is not sufficiently supported by the conclusion. This validation process is performed using an NLI model, which is well-suited for capturing fine-grained semantic consistency between two statements.

The validation results include the predicted entailment relationship and a corresponding confidence score. Rather than functioning as a standalone classifier, this module provides validation signals that guide the overall research workflow. The predicted validation confidence is compared against a predefined threshold, as illustrated in Figure 1. When the confidence score exceeds this threshold, the experimental conclusion is considered sufficiently supportive of the hypothesis, and no further refinement is required. Otherwise, the framework proceeds to the hypothesis refinement stage. Besides the validation results, the hypothesis validation module also produces fine-grained attribution information within the hypothesis that contributes most to the entailment classification, which is also fed back to researchers and subsequently leveraged to guide targeted hypothesis refinement.

2.2. Hypothesis Refinement via Attribution-Guided Local Editing

When a hypothesis is assessed as unsupportive by the validation module, the proposed framework activates the hypothesis refinement module to revise the hypothesis in a controlled and interpretable manner. The process of hypothesis refinement builds upon a pipeline of segmentation, alignment, tagging, and revision that has been systematically investigated in our prior work [9]. In this study, we integrate it as an established hypothesis refinement backbone. The input hypothesis is first decomposed into semantically coherent clauses. Attribution information produced by the hypothesis validation module is then aligned with these clauses to identify key spans within the original hypothesis that contribute most to the semantic inconsistency. Based on this alignment, these spans are tagged as refinement candidates while the remaining contents are preserved unchanged. This attribution-guided mechanism are more semantically transparent compared to rewriting the entire hypothesis using LLMs. Subsequently, the tagged spans are masked and revised through context-aware text infilling.

Notably, the hypothesis refinement stage is not designed to force entailment between the hypothesis and the experimental conclusion. Instead, its primary goal is to perform controllable revisions on the hypothesis, which not only revise local contradictory spans inconsistent with the conclusion, but also explicitly reveal the source of contradiction, enabling researchers to reassess the original

 doi:10.20944/preprints202601.1274.v1

hypothesis and providing transparent evidence for subsequent decision-making (e.g., formulating alternative hypotheses, redesigning experiments, etc.).

2.3. RAG-Based External Evidence Retrieval

Scientific hypotheses and conclusions often include dense domain-specific terminology that involves critical background knowledge required for scientific reasoning. For example, consider a hypothesis claiming that a palladium-based catalyst exhibits enhanced catalytic activity under mild oxidative conditions, while the experimental conclusion reports an increased turnover frequency after introducing molecular oxygen. Analyzing whether the hypothesis can be supported by such a conclusion requires background knowledge including catalytic mechanisms, the role of oxidizing agents, and experimental conditions.

To address this problem, the proposed framework incorporates a RAG-based external evidence retrieval module as an auxiliary component. Notably, this RAG module serves as an external knowledge provider that supplies relevant domain-specific knowledge to support hypothesis validation and refinement in a transparent and controllable manner. By explicitly separating evidence retrieval from decision-making, the framework ensures that the results of hypothesis validation and refinement remain anchored to the original hypothesis and conclusion, while retrieved evidence functions as an optional and interpretable source of contextual support.

As shown in Figure 1, we implement the RAG module following the paradigm of Modular RAG [10], which consists of four functional sub-modules: routing, pre-retrieval, search, and post-retrieval. The routing module serves as a task-level orchestration component. Given an input query composed of a hypothesis h, a conclusion c, and an instruction I, the routing module identifies the task context (e.g., hypothesis validation or hypothesis refinement) and activates the task-specific model. The pre-retrieval module aims to enhance query quality before retrieval. Specifically, it performs query rewriting and expansion on the original query. For query rewriting, we rewrite the query, including (h, c), using LLMs:

$$(h', c') \leftarrow \text{QueryRewrite}(h, c;\ I_{\text{rew}}). \tag{2}$$

For query expansion, inspired by hypothetical document embeddings (HyDE) [11], we generate a hypothetical conclusion $\hat{c}'$ based on the rewritten hypothesis h' as an auxiliary retrieval probe:

$$\hat{c}' \leftarrow \text{QueryExpand}(h';\ I_{\text{exp}}). \tag{3}$$

Here, both generated (h', c') and $\hat{c}'$ are not treated as factual evidence but only serves to explore the retrieval space. The search module bridges refined user queries and external knowledge bases. This module employs a vector-entity joint retrieval strategy that combines semantic similarity and entity relevance, enabling precise and context-aware information retrieval. And the post-retrieval module is responsible for assessing the relevance and potential usefulness of the retrieved chunks $\mathcal{E}$, which is performed using an evaluation LLM:

$$\tilde{\mathcal{E}} \leftarrow \text{Evaluate}(\mathcal{E};\ I_{\text{eva}}). \tag{4}$$

To provide an intuitive overview, we summarize the overall workflow of the proposed framework in Algorithm 1. Given an input hypothesis-conclusion pair, the framework first generate auxiliary representations to facilitate evidence retrieval from external knowledge bases. The retrieved evidence is then incorporated as contextual support for hypothesis validation (and hypothesis refinement when necessary), enabling the validation module to assess semantic consistency while remaining anchored to the original inputs. The validation result serves as the control signal of the framework. When the hypothesis is assessed as supportive, the workflow terminates without hypothesis refinement and returns validation results as well as attribution information and retrieved evidence as interpretable feedback to researchers. Otherwise, the framework activates the refinement module and leverages

 doi:10.20944/preprints202601.1274.v1

Algorithm 1 High-level workflow of the proposed framework (hypothesis validation and refinement driven by RAG retrieval)

Require: hypothesis h, conclusion c,
validation module $\mathcal{M}_{\text{val}}$ with threshold τ,
refinement module $\mathcal{M}_{\text{ref}}$
Ensure: validation result y, validation distribution $\boldsymbol{p}_{\text{val}}$,
attribution information $\mathbf{a}$, revised hypothesis $\tilde{h}$, retrieved chunks $\tilde{\mathcal{E}}$

$(h', c') \leftarrow \text{QueryRewrite}(h, c)$ {Pre-retrieval processing}
$\hat{c}' \leftarrow \text{QueryExpand}(h')$
$Q \leftarrow \{h', c', \hat{c}'\}$
$\mathcal{E} \leftarrow \emptyset$
for each query $q \in \mathcal{Q}$ **do**
  $\mathcal{E} \leftarrow \mathcal{E} \cup \text{VectorEntityRetrieval}(q)$ {External evidence retrieval}
end for
$\tilde{\mathcal{E}} \leftarrow \text{RelevanceFilter}(\mathcal{E})$ {Post-retrieval processing}
$(y, \boldsymbol{p}_{\text{val}}, \mathbf{a}) \leftarrow \mathcal{M}_{\text{val}}(h, c \mid \tilde{\mathcal{E}})$ {Hypothesis validation}
$O \leftarrow \{y, \boldsymbol{p}_{\text{val}}, \mathbf{a}, \tilde{\mathcal{E}}\}$
if $\boldsymbol{p}_{\text{val}}(\text{supportive}) < \tau$ **then**
  $\tilde{h} \leftarrow \mathcal{M}_{\text{ref}}(h, c \mid \mathbf{a}, \tilde{\mathcal{E}})$ {Hypothesis refinement}
  $O \leftarrow O \cup \tilde{h}$
end if
return O

attribution information and retrieved evidence to guide targeted, transparent hypothesis revision. All models used in the proposed framework are introduced in Appendix A.

3. Experimental Evaluation

This experimental evaluation is designed to systematically examine the effectiveness and practical implications of the proposed framework. Rather than focusing solely on performance improvements of individual models, the experiments aim to assess whether scientific reasoning benefits from organizing hypothesis validation, hypothesis refinement, and external evidence retrieval into a unified workflow. Specifically, we investigate the following three research questions:

- **(RQ1)** Does the proposed framework improve end-to-end hypothesis validation and refinement compared to standalone baselines?
- **(RQ2)** How does each component contribute to the performance of hypothesis validation and refinement?
- **(RQ3)** How do intermediate signals produced by the framework support reliability and interpretability in practical scientific reasoning?

All experiments are conducted on scientific texts in the chemistry domain for its rich domain-specific terminology and structured experimental reasoning in practical scientific research scenarios.

3.1. Experimental settings

3.1.1. Evaluation Datasets

Since hypothesis validation can be formulated as an NLI problem, we select CRNLI [13], a structured NLI corpus in the chemistry domain, to evaluate the performance of the hypothesis validation module. For hypothesis refinement, we employ the datasets proposed in [9], which are built for hypothesis revision via text infilling in the general and chemistry domain. Details of these datasets are provided in Appendix B. In this section, all experiments are conducted on their test set, including 3,812 instances for hypothesis validation and 1,809 instances for hypothesis refinement.

 doi:10.20944/preprints202601.1274.v1

3.1.2. Implementation Details

For hypothesis validation, the threshold τ of NLI confidence score is set to 0.49 using Youden's index. Through feature attribution, we select refinement candidates by filtering low-contribution words with contribution score below δ (default $\delta = 0.02$ after normalization) and subsequently selecting the top-m (default $m = 3$) words for each hypothesis clause. To perform context-aware text infilling, we form editable spans around these high-contribution words. Each span is defined as a symmetric window of size w (default $w = 5$) centered at a key word. Overlapping spans will be merged. For RAG settings, the external evidence retrieval module returns the top-k (default $k = 5$) chunks for each query. The external knowledge base are shared across all retrieval-based methods during this experiment.

3.1.3. Evaluation Metrics

We employ evaluation metrics including Accuracy, F1-score, and AUC to measure the performance of hypothesis validation. And for hypothesis refinement, we evaluate the revision quality using a series of metrics: BLEU-4 and ROUGE-L for assessing the token-level overlap between the generated infillings and ground truth, BERTScore for their semantic similarity, PPL for textual fluency of the overall revised hypotheses, and NLI score for quantifying their entailment relationship. Additionally, we introduce a binary evaluation metric named Span Completion Rate (SCR) to assess the model's ability to produce outputs that conform to the expected output format:

$$\text{SCR} = \begin{cases} 1, & \text{if} \quad \#G = \#M \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Given a hypothesis with $\#M$ masked spans, the model is expected to generate infillings $\#G$ that correspond one-to-one with the masked spans. To assess the effectiveness of feature attribution, we report MoRF and LeRF [23] by progressively masking top- and low-ranked tokens and tracking the change of NLI confidence. To evaluate the reliability of intermediate results, we employ ECE and BS to assess the calibration quality of the validation confidence scores.

3.2. Comparison of the Proposed Workflow with Baselines (RQ1)

We evaluate the proposed workflow by comparing several representative configurations that progressively incorporate external evidence retrieval and modular retrieval controls. This comparison aims to quantify the system-level benefit of organizing hypothesis validation, refinement, and evidence retrieval as a unified and controllable pipeline, rather than to introduce a new task-specific model.

For hypothesis validation, we compare four workflow configurations: standalone NLI and three retrieval-augmented variants (Naive RAG, Advanced RAG, and our Modular RAG setting, which are introduced in Appendix C). From Table 1, we observe that NLI with Naive RAG yields only marginal improvements compared to the NLI-only baseline, suggesting that simply incorporating external knowledge is not sufficient for chemical NLI when evidence quality is uncontrolled. In contrast, Advanced RAG shows a clearer performance improvement, indicating that stronger indexing and retrieval design can partially mitigate this problem. Our Modular RAG-based framework achieves the best performance across all metrics with 97.09% Accuracy, 97.28% F1, and 99.28% AUC, which outperforms the NLI-only baseline by 5.86% in accuracy, 5.71% in F1-score, and 2.3% in AUC. These results suggest that chemical NLI benefits from the retrieval quality control and vector-entity joint retrieval in our Modular RAG-based framework, which helps yield more reliable validation decisions. Moreover, the confusion matrix and ROC curve of this evaluation are illustrated in Figure 2.

 doi:10.20944/preprints202601.1274.v1

Table 1. Comparison of the proposed framework with baselines in terms of Accuracy, F1-score, and AUC on hypothesis validation (NLI) on the evaluation dataset.

Method	Accuracy (%)	F1-score (%)	AUC (%)
NLI w/o RAG	91.23	91.57	96.98
NLI w/ Naive RAG	91.56	91.62	97.08
NLI w/ Advanced RAG	94.10	94.22	98.12
NLI w/ Modular RAG (Ours)	**97.09**	**97.28**	**99.28**

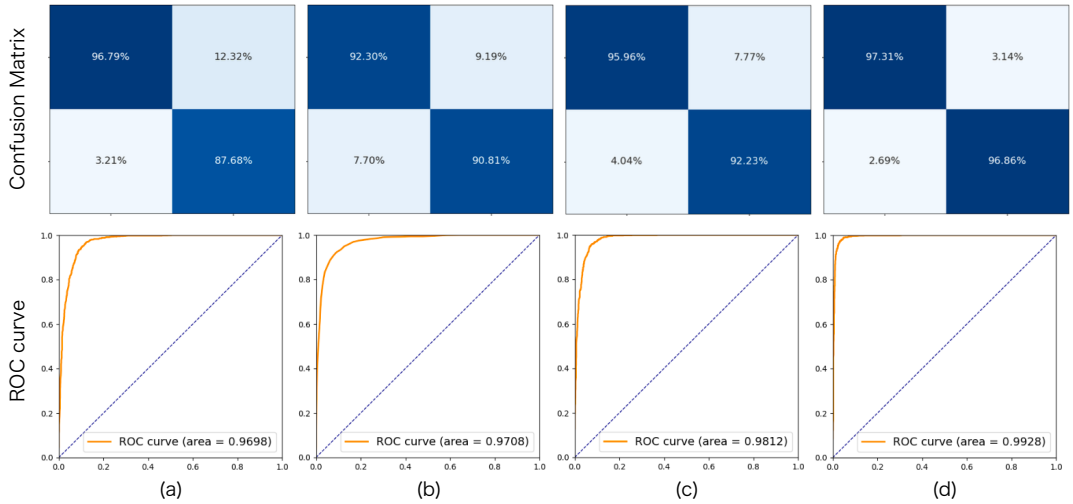

Figure 2. Confusion matrix and ROC curve of (a) NLI w/o RAG, (b) NLI w/ Naive RAG, (c) NLI w/ Advanced RAG, and (d) NLI w/ Modular RAG (Ours).

Since the proposed hypothesis refinement module follows a context-aware text infilling method introduced in [9], similar to hypothesis validation, we also evaluate four text infilling configurations as shown in Tables 2 and 3. Table 2 reports token- and semantic-level similarity between the generated infillings and ground truth spans. Here, BERTScore is computed using SciBERT [12]. Interestingly, we observe that Naive RAG does not improve over the infill-only setting, indicating that directly incorporating external knowledge may introduce irrelevant or noisy cues that contribute negatively to span infilling. Advanced RAG and our Modular RAG-based framework demonstrate a significant improvement. Especially, our framework achieves the best performance (48.43 BLEU, 48.8 ROUGE, and 0.9258 BERTScore), suggesting that our framework is capable of generating the lexical structure of ground-truth infilling while maintaining semantic consistency.

Table 2. Comparison of the proposed framework with baselines in terms of BLEU-4, ROUGE-L, and BERTScore on hypothesis refinement (text infilling) on the evaluation dataset.

Method	BLEU	ROUGE	BERTScore (%)
Infill w/o RAG	44.21	45.72	90.97
Infill w/ Naive RAG	43.04	45.29	90.74
Infill w/ Advanced RAG	47.15	46.91	91.36
Infill w/ Modular RAG (Ours)	**48.43**	**48.8**	**92.58**

 doi:10.20944/preprints202601.1274.v1

Table 3. Comparison of the proposed framework with baselines in terms of PPL, NLI score, and SCR on hypothesis refinement (text infilling) on the evaluation dataset.

Method	PPL ↓	NLI score (%) ↑	SCR (%) ↑
Original	10.1	83.2	N/A
Masked	N/A	53.57	N/A
Infill w/o RAG	10.72	79.23	100
Infill w/ Naive RAG	**10.64**	80.41	99.83
Infill w/ Advanced RAG	11.2	82.7	100
Infill w/ Modular RAG (Ours)	10.81	**82.92**	100

Furthermore, we evaluate the textual fluency (PPL) of completed hypotheses using a different evaluation model Phi-3.5-mini-instruct [28]. As illustrated in Table 3, compared to the original hypothesis that serves as a baseline, Naive RAG yields the lowest PPL of 10.64, indicating that the textual fluency of completed hypotheses cannot benefit from incorporating external knowledge. Besides BERTScore that evaluates the span-level similarity, we employ LANLI [13] to compute NLI score that assesses whether the completed hypotheses are semantically aligned with the conclusion. Since all hypothesis-conclusion pairs in the hypothesis revision dataset are labeled as entailment, the test set achieves an average NLI score of 0.832 as shown in Table 3, which serves as an upper bound reference. And it drops to 0.5357 when high-attribution spans within hypotheses are masked. We observe that all infilling variants recover the score, demonstrating the effectiveness of hypothesis refinement via span infilling. Especially, our Modular RAG-based framework achieves the highest NLI score of 0.8292, which approaches the original score (0.832), indicating that the revised hypotheses better align with the conclusions. Finally, we report SCR to quantify models' ability to produce outputs that conform to the expected output format (e.g., output a well-formed set of infillings that matches the number of masked spans and aligns one-to-one with each corresponding mask). As illustrated in Table 3, SCR is near-saturated across all configurations, indicating that introducing external knowledge has little influence on SCR and all baselines can reliably follow the expected output format.

To assess the effectiveness of the attribution method (SHAP [14]) used in our framework, we conduct a word masking-based faithfulness evaluation experiment on the hypothesis revision dataset. Specifically, we perform individual masking of top-10 high-attribution words ranked by SHAP and measure the NLI score drop Δ:

$$\Delta = S_{orig} - S_{mask}, \tag{6}$$

where S_{orig} denotes the baseline NLI score without masking, S_{mask} denotes the NLI score after masking. Since the hypothesis revision dataset only consists of entailment-labeled samples, the NLI score is expected to drop after keyword masking, and a larger Δ indicates the masked word is more important for the entailment decision. The results are illustrated in Figure 3 (left). We observe that removing the top-ranked word leads to the largest NLI score decrease (Mean $\Delta = 0.37$), while the drops for lower-ranked words quickly shrink toward near-zero values. Additionally, we further measure the MoRF and LeRF by masking all the top- and bottom-5 high-attribution words ranked by SHAP, respectively. As illustrated in Figure 3 (right), MoRF has a substantially larger and more dispersed Δ distribution compared to LeRF whose values tightly centered around 0. The consistent rank-wise decay in Δ and the great separation between MoRF and LeRF indicate that SHAP reliably identifies words that contribute most to the validation decision, supporting its use as the attribution signal for subsequent refinement steps in our framework.

 doi:10.20944/preprints202601.1274.v1

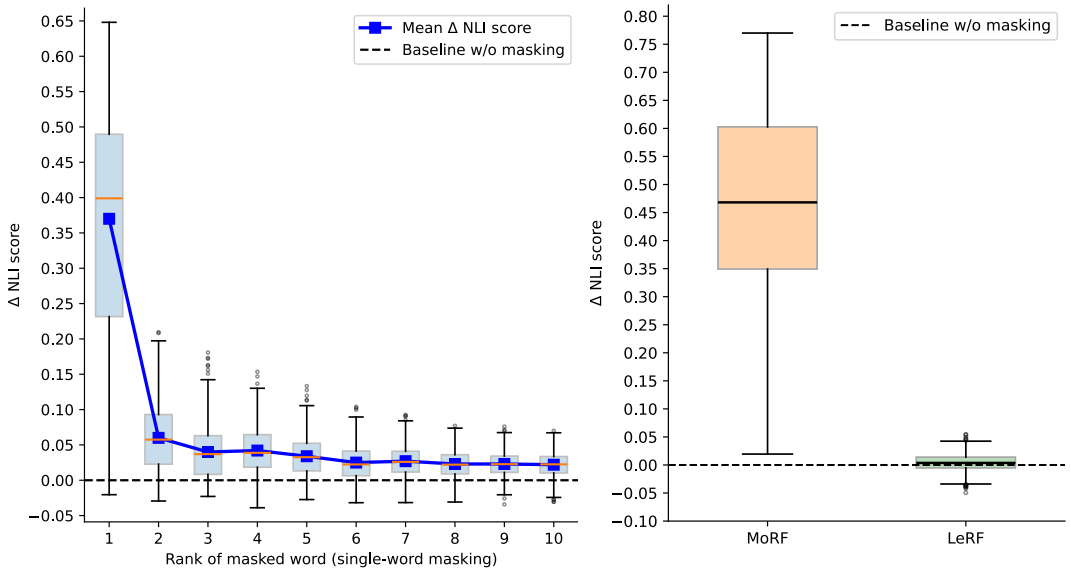

Figure 3. Evaluation of attribution effectiveness. Left: individual masking of top-10 high-attribution words ranked by SHAP. Right: comparison of MoRF and LeRF that mask all top- and bottom-5 high-attribution words respectively. The vertical axis denotes the drop Δ in NLI score by keyword masking ($\Delta = S_{orig} - S_{mask}$).

3.3. Component-Wise Analysis of the Framework (RQ2)

To investigate how each component contributes to the overall performance, we conduct an ablation study over components including pre-retrieval processing, vector-entity joint retrieval, and post-retrieval filtering under the same experimental settings as in RQ1.

For hypothesis validation, we report the experimental results in Table 4. We observe that removing either vector retrieval or entity retrieval led to the most severe drop in performance. Especially, compared to the full framework, the accuracy of removing vector retrieval decreases from 97.09% to 92.73%. It confirms that dense semantic similarity retrieval serves as a fundamental component for capturing high-level contextual relevance. While entity retrieval provides strong domain specificity, relying solely on structured entities limits the model's capability to retrieve semantically related information. Removing entity-based retrieval also causes a significant accuracy drop from 97.09% to 94.1%. It suggests that the framework benefits from the retrieved knowledge using entity retrieval. While removing pre- and post-retrieval processing cause a relatively lower drop in performance, they still contribute to the effectiveness of our proposed framework.

Table 4. performance of different ablation settings in terms of Accuracy, F1-score, and AUC on hypothesis validation (NLI) on the evaluation dataset, with NLI-only and proposed full framework as reference.

NLI Method	Accuracy (%)	F1-score (%)	AUC (%)
w/o RAG	91.23	91.57	96.98
Ours w/o pre-retrieval	96.27	95.91	99.09
Ours w/o entity retrieval	94.10	94.22	98.12
Ours w/o vector retrieval	92.73	93.04	97.55
Ours w/o post-retrieval	96.73	96.50	99.12
w/ Modular RAG (Ours)	**97.09**	**97.28**	**99.28**

For hypothesis refinement (Table 5), we observe a trend consistent with hypothesis validation: removing core retrieval modules such as vector retrieval leads to a larger performance drop than removing quality-control components. Additionally, Table 5 indicates that retrieval settings affect not only span-level similarity metrics (BLEU and BERTScore) but also the semantic alignment of the completed hypothesis with the conclusion (NLI score).

 doi:10.20944/preprints202601.1274.v1

Table 5. performance of different ablation settings in terms of BLEU-4, BERTScore, and NLI score on hypothesis refinement (text infilling) on the evaluation dataset, with Infill-only and proposed full framework as reference.

Infill Method	BLEU	BERTScore (%)	NLI score (%)
Original	100	100	83.2
Masked	N/A	N/A	53.57
w/o RAG	44.21	90.97	79.23
Ours w/o pre-retrieval	47.63	91.86	82.86
Ours w/o entity reitrieval	47.15	91.38	82.7
Ours w/o vector reitrieval	45.92	91.24	80.75
Ours w/o post-reitrieval	46.18	91.4	82.51
w/ Modular RAG (Ours)	**48.43**	**92.58**	**82.92**

Furthermore, we conduct ablation experiments on the attribution method that is responsible for guiding the targeted hypothesis revision. Here, we compare SHAP with two representative feature attribution methods: integrated gradients and attention weights. Following Figure 3 (left), we respectively mask the top-10 high-attribution words ranked by three attribution methods and summarize the results in Figure 4. We observe a consistent rank-wise Δ decay across all three attribution methods, which indicates their effectiveness as attribution methods. Notably, SHAP exhibits a higher NLI score drop ($\Delta^1_{\text{SHAP}} = 0.37$) when masking the top-1 high-attribution word. It significantly outperforms integrated gradients and attention weights, suggesting that SHAP is more reliable for identifying high-contribution words and providing accurate guidance for targeted hypothesis revision.

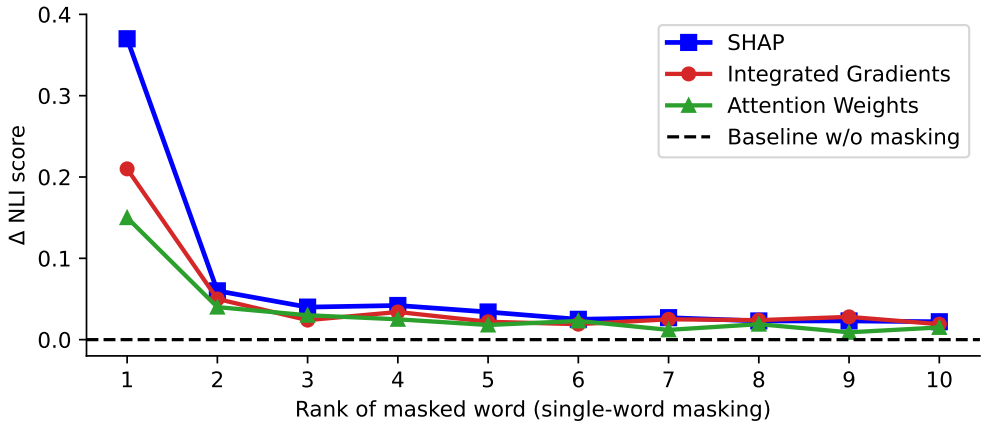

Figure 4. Ablation of attribution methods under individual masking of top-10 high-attribution words ranked by SHAP (blue), integrated gradients (red), and attention weights (green).

3.4. Intermediate Signal Analysis (RQ3)

In this subsection, we investigate the intermediate signals produced by our workflow, including the validation decision and its confidence score, attribution scores over hypothesis words, and retrieved external knowledge chunks.

To evaluate the reliability of confidence estimates for hypothesis validation (NLI), we compute ECE and BS on different NLI configurations to assess whether their predicted probabilities reflect true likelihoods of entailment. The results are illustrated in Table 6. Compared to baselines, the proposed framework achieves the lowest ECE (0.029) and BS (0.051), indicating the most well-calibrated confidence among all compared configurations, which supports the use of confidence as a reliable decision signal.

 doi:10.20944/preprints202601.1274.v1

Table 6. Comparison of the proposed framework with three baselines in terms of ECE and BS on hypothesis validation (NLI).

Method	ECE ↓	BS ↓
NLI w/o RAG	0.068	0.094
NLI w/ Naive RAG	0.065	0.090
NLI w/ Advanced RAG	0.042	0.068
NLI w/ Modular RAG (Ours)	**0.029**	**0.051**

Next, we present an end-to-end case study in Figures 5 and 6 to provide an intuitive understanding on the overall pipeline of the proposed framework. Figure 5 illustrates evidence-driven hypothesis validation, feature attribution, hypothesis revision via span infilling. Figure 6 illustrates the detailed chunk information retrieved from external chemical knowledge base.

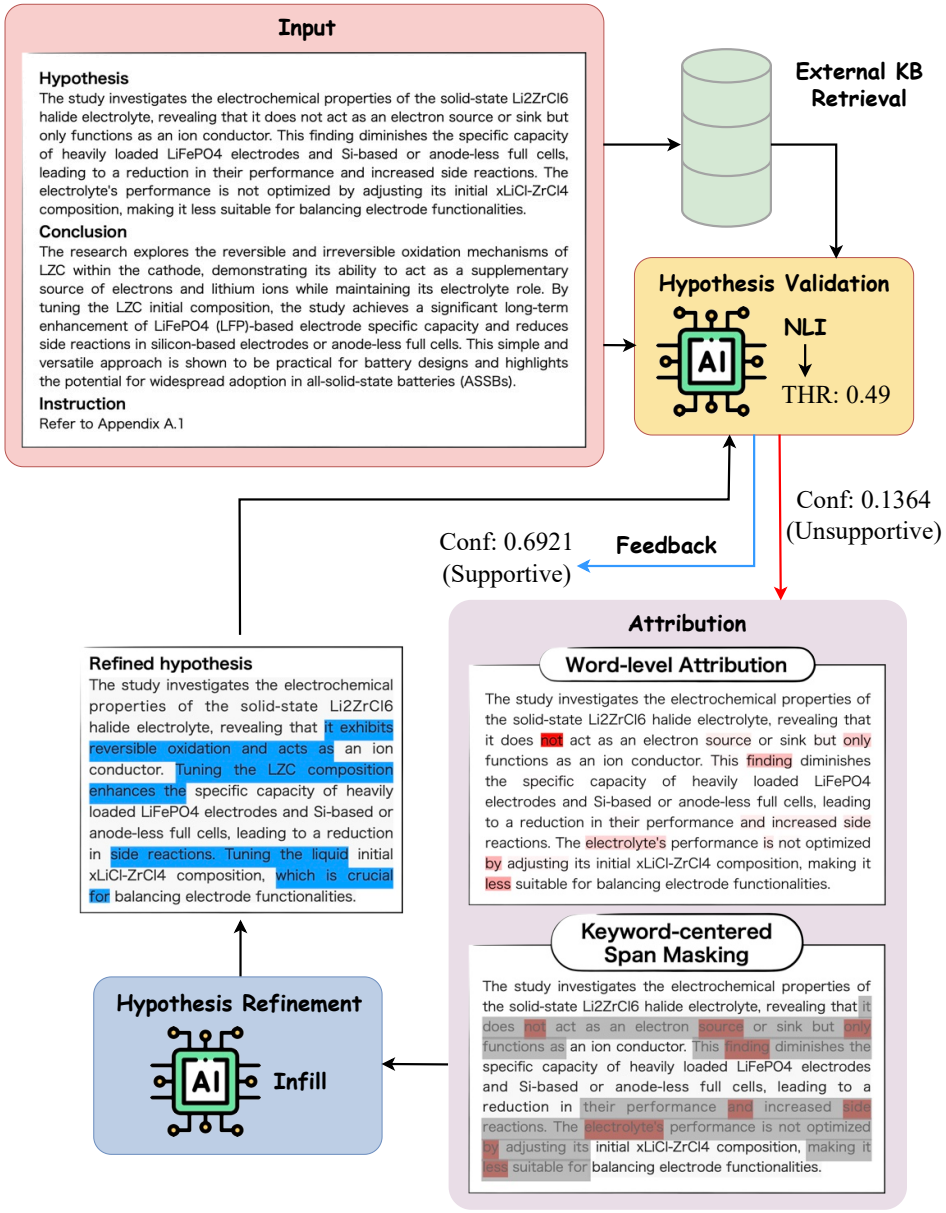

Figure 5. An end-to-end case study of the proposed framework, including input hypothesis and conclusion, validation decision with confidence, attribution map highlighting decision critical words, context-aware span masking, and hypothesis revision results highlighting infilled spans.

 doi:10.20944/preprints202601.1274.v1

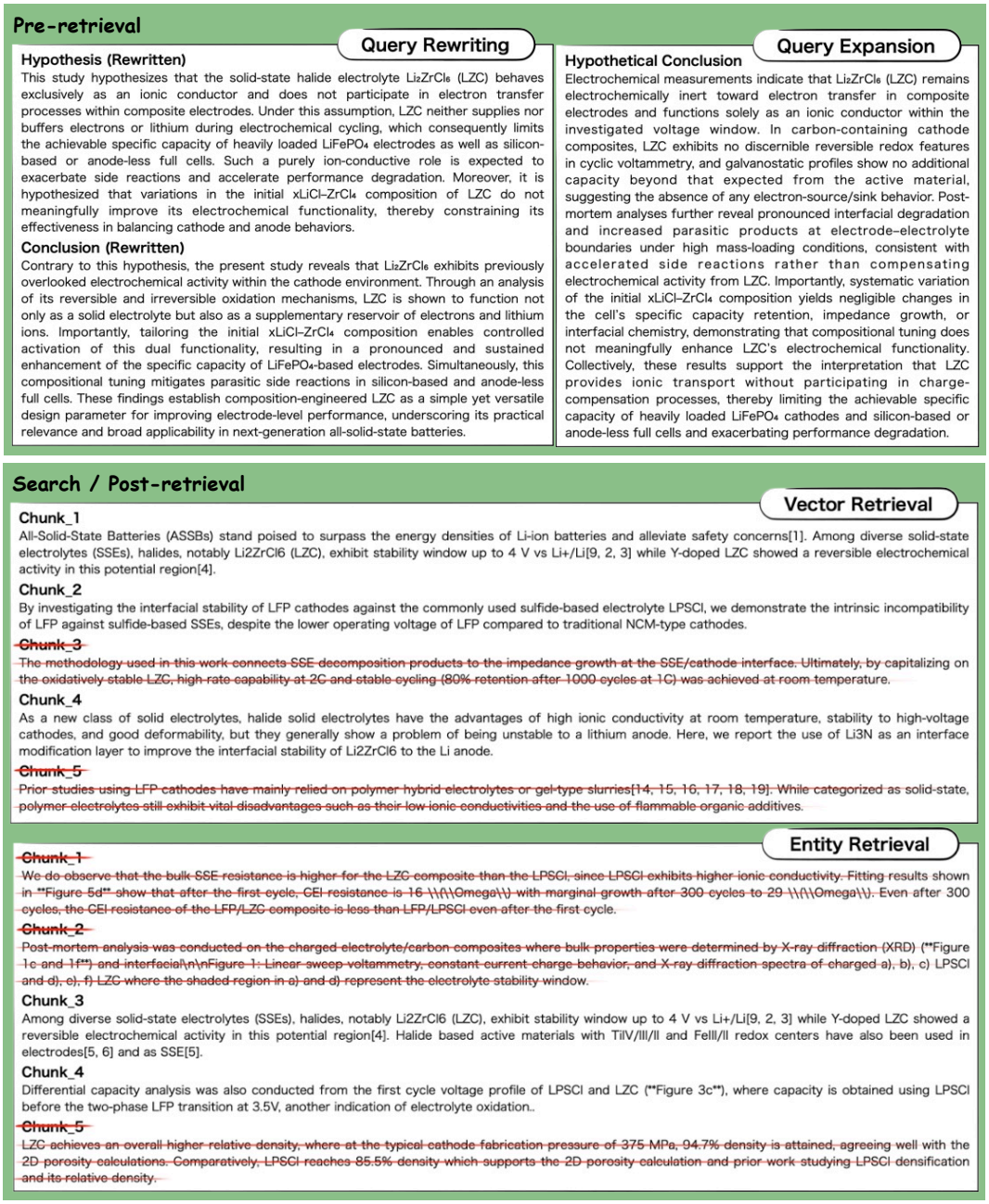

Figure 6. Detailed chunk information retrieved from external chemical knowledge base, including the results of query rewriting and expansion, chunks retrieved by vector-entity joint retrieval, chunks filtered by post-retrieval LLM evaluation.

Given an input hypothesis-conclusion pair, the framework first generate auxiliary representations to facilitate evidence retrieval from the external chemical knowledge base. Specifically, as illustrated in Figure 6, the rewritten hypothesis-conclusion pair and the hypothetical conclusion are generated to enhance query quality before retrieval. Followed by the query rewriting and expansion, the framework explore the external chemical knowledge base for relevant evidence using a vector-entity joint retrieval strategy. Then, the retrieved chunks are evaluated and filtered using an evaluation LLM to ensure the retrieval quality. As illustrated in Figure 5, the retrieved chunks are incorporated as contextual support for hypothesis validation. In this case, the NLI model outputs an NLI confidence of 0.1364, which is below the threshold τ (0.49), indicating the inputs are classified as unsupportive. The attribution module produces word-level contribution scores over the hypothesis. We visualize them by highlighting each word in red based on its score (higher contribution scores correspond to darker red colors). For example, in this case, words such as "not", "only", "less" etc. contribute most to the

 doi:10.20944/preprints202601.1274.v1

NLI decision. Then, we mask spans (top-3 for each hypothesis clause) centered on these words with a context window size of 5 and revise the hypothesis via context-aware text infilling. All infilled contents are highlighted in blue. The refined hypothesis is re-evaluated against the same conclusion, resulting in a new NLI confidence score of 0.6921, which closes the workflow and feed validation results along with retrieved chunks, attribution information, and the refinement hypothesis back to researchers. The case study not only indicates the effectiveness of the proposed framework, but also suggests that it produces reliable intermediate signals that enhance transparency and traceability throughout hypothesis validation and refinement.

4. Discussion

The proposed framework aims to organize hypothesis validation and refinement into an inspectable scientific workflow rather than aligning hypotheses with experimental conclusions, which also introduces several practical limitations. First, hypothesis refinement is constrained to attribution-guided local editing. While this design helps enhance the precision and interpretability of hypothesis revision, it may be insufficient for cases that require global restructuring. Second, the refinement quality is bounded by the reliability of upstream results. Since revision span is derived from the attribution on validation results, errors occurred in hypothesis validation can propagate to masking decisions and subsequently affect the refined hypothesis, making the workflow sensitive to NLI calibration and attribution noise. Third, the external knowledge base plays a key role in the proposed framework. Even with pre- and post-retrieval processing, low-quality evidence can limit performance of both validation and refinement, especially when relevant domain-specific knowledge is absent or poorly represented in the corpus. These limitations suggest that the framework is most suitable when researchers investigate intermediate outputs as checkpoints and apply manual review for difficult cases in practical scientific research workflows.

5. Conclusion

In this study, we presented a modular framework for automated hypothesis validation and refinement in scientific research. Rather than introducing new task-specific models, the framework integrates established techniques, including NLI-based hypothesis validation, attribution-guided hypothesis refinement, and Modular RAG-based external evidence retrieval, into a unified and controllable workflow. We conduct extensive experiments on scientific texts in the chemistry domain. By Compared with baseline configurations, we evaluate and verify the effectiveness of the proposed framework on hypothesis validation, hypothesis refinement, and feature attribution. Ablation studies further show that the quality of retrieval and attribution plays a crucial role throughout the overall research workflow. Additionally, the end-to-end case study suggests that the framework produces reliable intermediate signals as checkpoints, enabling researchers to trace the research workflow and apply manual review for difficult cases if necessary. Our work offers a modular solution for deploying LLM-based systems into scientific research workflows. As future work, we plan to explore domain transfer and further optimize the framework on limitations discussed in Section 4.

Author Contributions: Conceptualization, C.C.; methodology, C.C.; software, C.C; validation, C.C. and T.M.; formal analysis, C.C.; investigation, C.C. and T.M.; resources, C.C.; data curation, C.C. and T.M; writing—original draft preparation, C.C.; writing—review and editing, C.C.; visualization, C.C.; supervision, T.H., T.Y. and H.F.; project administration, H.F.; funding acquisition, H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JST Moonshot R&D grant number JPMJMS2236.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experimental code and data related to this paper can be obtained by contacting the corresponding author.

 doi:10.20944/preprints202601.1274.v1

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments that greatly improved the quality of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest regarding the publication of this paper. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

Appendix A. Models Used in the Framework

All models used in the proposed framework are instantiated with chemistry-specific variants, reflecting the target application of this study. The following sections briefly describe each model and its role within the framework.

Appendix A.1. Hypothesis Validation Module

The hypothesis validation module is implemented using a chemistry-domain NLI model, LANLI [13], which is a decoder-only method tailored for long-form hypothesis validation in chemical contexts. LANLI takes as input a hypothesis-conclusion pair and external reference knowledge, and outputs a binary validation decision with a confidence score. Furthermore, it integrates NLI with SHAP [14] to identify key tokens within the hypothesis that contribute most to the NLI result. The prompt for hypothesis validation is illustrated in Figure A1(a).

Appendix A.2. Hypothesis Refinement Module

The hypothesis refinement module employs our previous work [9], which performs clause-level attribution-guided span masking and context-aware text infilling. The model takes as input a masked hypothesis, a conclusion, and external reference knowledge, and outputs infilling contents. In this work, the module is used as a local editing operator to revise contradictory spans identified by the validation module. The prompt for hypothesis refinement is illustrated in Figure A1(b).

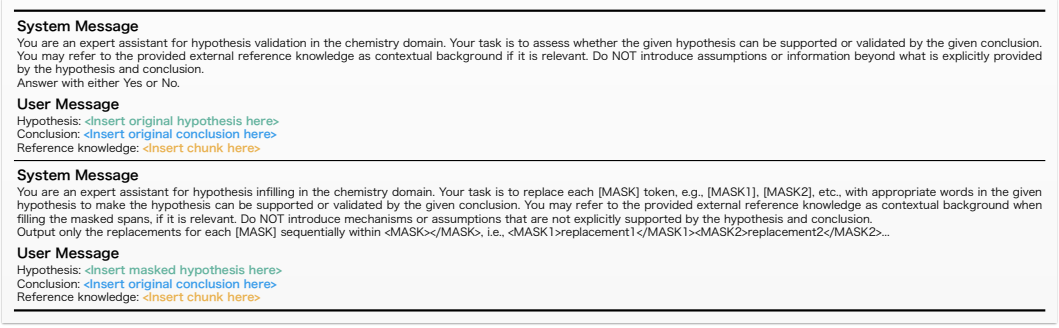

Figure A1. Prompt instructions for (a) hypothesis validation and (b) hypothesis refinement.

Appendix A.3. RAG-based External Evidence Retrieval Module

The external evidence retrieval module follows the Modular RAG paradigm introduced in [10], composed of four sub-modules: routing, pre-retrieval, search, and post-retrieval. For semantic routing strategy (routing module) and chunk embedding (search module), we employ the Sentence-BERT model (all-MiniLM-L6-v2) [19,20] to encode texts. The pre-retrieval (query rewriting and query expansion) and post-retrieval (LLM evaluation) modules are implemented using off-the-shelf LLM (DeepSeek-R1-Distill-Llama-8B) [15,16] with task-specific prompts, as illustrated in Figure A2.

Additionally, the external knowledge base employs a chemistry-specific database proposed in [17]. It is constructed from ChemRxiv, a widely used preprint server in chemical sciences, covering a broad spectrum of chemical sub- fields. The detailed statistics of this chemistry knowledge base are summarized in Table A1. The entity database employs SciSpaCy [21] for named entity recognition (NER), and we default to use FAISS [22] for similarity search between query and chunk embeddings.

 doi:10.20944/preprints202601.1274.v1

System Message
You are an expert assistant for chemical text normalization and retrieval preparation. Your task is to rewrite the given chemical hypotheses and conclusions into a clear, precise, and retrieval-friendly form while strictly preserving their original semantic meaning. Do NOT introduce new information, assumptions, or external knowledge. Do NOT perform reasoning or validation. Focus on clarity, explicitness, and structural completeness.
Return the result in the following format:
Rewritten Hypothesis: <rewritten hypothesis here>
Rewritten Conclusion: <rewritten conclusion here>

User Message
Hypothesis: <Insert original hypothesis here>
Conclusion: <Insert original conclusion here>

System Message
You are an expert assistant for chemical text writing. Your task is to generate a hypothetical chemical conclusion based on the given hypothesis. The generated text does NOT need to be factually correct. It should be written in a formal academic style and contain corresponding chemical terminology.
Return the result in the following format:
Generated Conclusion: <generated conclusion here>

User Message
Hypothesis: <Insert rewritten hypothesis here>

System Message
You are an expert assistant for evidence filtering in retrieval-augmented natural language inference (NLI) in the chemistry domain. Your task is to evaluate whether a given chunk should be kept or discarded based on its relevance and usefulness for assisting reasoning between a hypothesis and a conclusion. You must strictly follow these rules:
- Do NOT perform entailment, contradiction, or neutrality judgment between the hypothesis and conclusion.
- Do NOT validate the truth of the hypothesis or the conclusion.
- Do NOT introduce new information, assumptions, or external knowledge.
- Do NOT rewrite or summarize the hypothesis, conclusion, or chunk.
- Focus solely on whether the chunk provides relevant background knowledge, constraints, entities, conditions, or mechanisms that could assist hypothesis–conclusion inference.

Return the result in the following format:
Decision: <keep or discard>

User Message
Hypothesis: <Insert original hypothesis here>
Conclusion: <Insert original conclusion here>
Chunk: <Insert chunk here>

Figure A2. Prompt instructions for (a) query (hypothesis and conclusion) rewriting, (b) query expansion (hypothetical conclusion generation), and (c) post-retrieval evaluation.

Table A1. Detailed statistics of the external chemistry knowledge base.

	Meta-paper	Referenced paper	Total
Domain	Chemistry	Chemistry, others.	N/A
Source data	ChemRxiv	Crossref	N/A
Data volume	35,373	747,693	783,066
w/ PDF	35,224	4,181	39,372
w/ Abstract only	149	357,634	357,783
Chunk volume	529,569	394,912	924,481
Entity volume	7,607,812	2,359,335	9,967,147

Appendix B. Evaluation Datasets Used in Experiments

For hypothesis validation, the evaluation dataset CRNLI is a structured NLI dataset in the chemistry domain. We compare CRNLI with SNLI (general-domain NLI dataset) [18] and summarize their statistical information in Table A2. Besides chemistry-specific data, CRNLI features an average token length of 122.8, which is substantially longer than that of traditional NLI datasets such as SNLI (11.2 tokens).

Table A2. Statistical information of SNLI and CRNLI.

	SNLI	CRNLI
Domain	General	Chemistry
Source data	Flickr 30k, VisualGenome	ChemRxiv
Dataset volume	570k	53.5k
Label distribution	3 (E/C/N)	2 (E/C)
Data distribution	Balanced	Balanced
Train/Dev/Test	550k/10k/10k	45.5k/4.2k/3.8k
Avg. token length	11.2	122.8

For hypothesis refinement, we employ the hypothesis revision datasets proposed in [9], which are constructed from SNLI and CRNLI by applying attribution-guided masking to entailment-labeled NLI samples. Detailed dataset statistics are illustrated in Table A3.

 doi:10.20944/preprints202601.1274.v1

Table A3. Dataset statistics of hypothesis revision datasets in the general and chemistry domain.

	General	Chemistry
Source data	SNLI	CRNLI
Data volume	174k	25k
Train/Dev/Test	147k/14k/13k	21.3k/1.9k/1.8k
Avg. token length	11.2	122.8
Avg. # of masked spans	1.7	2.6
Avg. masking window size	2.8	7.1

Appendix C. RAG Paradigms Introduced During the Evaluation of Hypothesis Validation

[10] categorizes the RAG research paradigm into three stages: Naive RAG, Advanced RAG, and Modular RAG. Naive RAG represents the earliest RAG methodology, following a traditional process that includes indexing, retrieval, and generation, which is also characterized as a "Retrieve-Read" framework [24]. However, Naive RAG suffers from several limitations such as sensitivity to retrieval noise. As a result, Advanced RAG introduces specific improvements to overcome these limitations. For example, Advanced RAG optimizes the indexing structure by enhancing data granularity, adding metadata, mixed retrieval etc. [11,25] Recently, Modular RAG has been proposed as a flexible and extensible architectural framework that decomposes the RAG pipeline into loosely coupled functional modules. Compared to Naive RAG and Advanced RAG, innovations like restructured RAG modules [26] and rearranged RAG pipelines [27] have been introduced to tackle specific challenges.

References

1. Lu, C., Lu, C., Lange, R.T., Foerster, J.N., Clune, J., & Ha, D. (2024). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. ArXiv, abs/2408.06292.
2. Lenat, D.B. (1977). Automated Theory Formation in Mathematics. In Proceedings of the 5th International Joint Conference on Artificial Intelligence, 2, 833-842.
3. Lenat, D.B., & Brown, J.S. (1984). Why AM and EURISKO Appear to Work. *Artificial intelligence*, 23, 269-294.
4. Buchanan, B.G., & Feigenbaum, E.A. (1978). Dendral and Meta-Dendral: Their Applications Dimension. *Artificial intelligence*, 11, 5-24.
5. Freire, J., Fan, G., Feuer, B., Koutras, C., Liu, Y., Peña, E., Santos, A.S., Silva, C., & Wu, E. (2025). Large Language Models for Data Discovery and Integration: Challenges and Opportunities. *IEEE Data Engineering Bulletin*, 49, 3-31.
6. Luo, Z., Yang, Z., Xu, Z., Yang, W., & Du, X. (2025). LLM4SR: A Survey on Large Language Models for Scientific Research. ArXiv, abs/2501.04306.
7. Zhu, Y., Qiao, S., Ou, Y., Deng, S., Zhang, N., Lyu, S., Shen, Y., Liang, L., Gu, J., & Chen, H. (2025). KnowAgent: Knowledge-Augmented Planning for LLM-Based Agents. In Findings of the Association for Computational Linguistics: NAACL 2025, 3709–3732.
8. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv, abs/2005.11401.
9. Chen, C., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2025). Hypothesis Alignment via Clause-level Attribution-guided Span Masking and Infilling. In Proceedings of the 5th International Conference on Communications, Networking and Machine Learning
10. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv, abs/2312.10997.
11. Gao, L., Ma, X., Lin, J.J., & Callan, J. (2022). Precise Zero-Shot Dense Retrieval without Relevance Labels. Annual Meeting of the Association for Computational Linguistics.
12. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Conference on Empirical Methods in Natural Language Processing.
13. Chen, C., Masuda, T., Ushiku, Y., Tanaka, S., Saito, K., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2025). CRNLI: A Textual Entailment Dataset in the Chemistry Domain. International Conference on Text, Speech and Dialogue.

 doi:10.20944/preprints202601.1274.v1

14. Lundberg, S.M., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Neural Information Processing Systems.
15. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv, abs/2302.13971.
16. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., & Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 645, 633-638.
17. Chen, C., Masuda, T., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2026). Toward Retrieval-Augmented Automatic Hypothesis Validation in the Chemistry Domain. *IEEE Access (Submitted)*.
18. Bowman, S.R., Angeli, G., Potts, C., & Manning, C.D. (2015). A large annotated corpus for learning natural language inference. Conference on Empirical Methods in Natural Language Processing.
19. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv, abs/1908.10084.
20. Yin, C., & Zhang, Z. A Study of Sentence Similarity Based on the All-minilm-l6-v2 Model With "Same Semantics, Different Structure" After Fine Tuning.
21. Neumann, M., King, D., Beltagy, I., & Ammar, B.W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. ArXiv, abs/1902.07669.
22. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazar'e, P., Lomeli, M., Hosseini, L., & J'egou, H. (2024). The Faiss library. ArXiv, abs/2401.08281.
23. Li, X., Du, M., Chen, J., Chai, Y., Lakkaraju, H., & Xiong, H. (2023). M4: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods across Metrics, Modalities and Models. Neural Information Processing Systems.
24. Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query Rewriting for Retrieval-Augmented Large Language Models. ArXiv, abs/2305.14283.
25. Zheng, H.S., Mishra, S., Chen, X., Cheng, H., Chi, E.H., Le, Q.V., & Zhou, D. (2023). Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models. ArXiv, abs/2310.06117.
26. Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., & Jiang, M. (2022). Generate rather than Retrieve: Large Language Models are Strong Context Generators. ArXiv, abs/2209.10063.
27. Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023). Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy. ArXiv, abs/2305.15294.
28. Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H.H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H.S., Benhaim, A., Bilenko, M., & Yang, Y. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. ArXiv, abs/2404.14219.