EvoNarrator: Modeling Scientific Evolution for Feasible Hypothesis Generation

Xiaoying Le[1,2], PengfeiQian[1,2], Yuanzhao Zhai[1,2], Xu Zhang[1,2], Qian Liu[3], Feng Dawei[1,2,*], Bo Ding[1,2,*]

[1]College of Computer Science and Technology, National University of Defense Technology, China
[2]State Key Laboratory of Complex & Critical Software Environment, China
[3]Hunan Institute of Advanced Technology, China

{lexy, pengfeiqian, yuanzhaozhai, dafeng}@nudt.edu.cn

Abstract

Scientific discovery evolution does not emerge in isolation but stems from the structural deepening and recombination of existing functionalities. However, current automated hypothesis generation methods, constrained by the statistical co-occurrence nature of Large Language Models (LLMs), lack perception of temporal causality and the "evolutionary patterns" inherent in scientific development. Consequently, they often yield superficial combinations that are logically infeasible. To address this, we propose **EvoNarrator**, a framework for hypothesis generation based on evolutionary narratives. We first extract structured P-M-L-F (Problem, Method, Limitation, Future Work) quadruples from citation networks. Subsequently, we introduce the **SocketMatch** mechanism, which eliminates logical disconnects between methods and problems by assessing their deep semantic compatibility. Finally, utilizing three macro patterns—*Chain*, *Divergence*, and *Convergence*—we constrain the generation process within historically logical derivation paths. Furthermore, double-blind expert reviews yielded an average score of 4.80/5.00 across novelty, feasibility, theoretical, and Logical. Additionally, hindcasting experiments validated its predictive foresight. Crucially, ablation studies indicate that integrating evolutionary patterns facilitates a paradigm shift from conservative incrementalism to theoretically grounded structural innovation. The code is available at https://github.com/xiyii-star/EvoNarrator.

1 Introduction

Scientific research can be conceptualized as an evolutionary process characterized by the recombination of antecedent components to resolve emerging challenges. This perspective aligns with W. Brian Arthurs analysis in *The Nature of Technology*, which posits that technological evolution does not occur *ex nihilo*; rather, it arises from the structural deepening and novel reconfiguration of existing functional elements (Arthur, 2009). However, the exponential growth of literature creates severe "cognitive overload," preventing researchers from manually discerning evolutionary threads or identifying innovation opportunities. Consequently, Automated Hypothesis Generation has become a critical challenge in AI for Science.

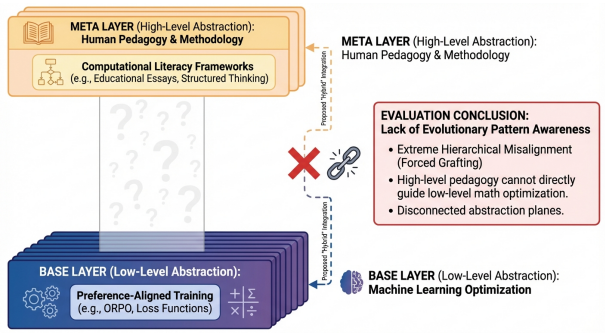

Figure 1: An illustrative case of "Lack of Evolutionary Pattern Awareness."

While Large Language Models (LLMs) excel at creative brainstorming (Si et al., 2024; Kumar et al., 2025), they struggle with rigorous scientific inquiry due to a lack of **Evolutionary Pattern Awareness**. Relying on semantic co-occurrence rather than temporal causality, current models often suggest superficial "hybridizations"such as applying high-level educational frameworks to low-level loss optimization (Figure 1)which are textually novel but technically infeasible. Neglecting structural patterns (e.g., Method Y improving Method X) leads to logical hallucinations rather than actionable insights.

To address this, we propose **EvoNarrator**, a framework based on **Evolutionary Narrative**. We posit that a high-quality hypothesis is not a random sample, but the inevitable "next step" in the scientific network. We first extract structured

*Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11846–11865
July 2-7, 2026

P-M-L-F quadruples (Problem, Method, Limitation, Future Work) from citation networks. We then introduce **SocketMatch**, a mechanism that calculates deep semantic compatibility to translate vague citations into explicit logical operators (e.g., *Overcomes*, *Adapts_to*). Finally, we elevate these micro-connections into macro evolutionary patterns—*Chain*, *Divergence*, and *Convergence*—to target specific unresolved limitations. This transforms hypothesis generation from open-ended writing into constrained logical deduction, ensuring generated hypothesis are both feasible and justifiable.

Our main contributions are:

- **Evolutionary Narrative Framework:** We model hypothesis generation as a dynamic, history-aware evolutionary process rather than static text completion.
- **SocketMatch Mechanism:** A fine-grained alignment algorithm that resolves the "supply-demand mismatch" of innovation elements via explicit logical operators.
- **Pattern-Driven Generation:** By formalizing evolutionary patterns, our method generates hypotheses with clear rationales, significantly outperforming baselines in Dual-Perspective Arena evaluations.

2 Related Work

Evolution-Aware Scientific Analysis. Understanding the trajectory of scientific evolution is a prerequisite for generating plausible hypotheses. Early approaches primarily relied on citation topologysuch as citation counts and PageRank (Bai et al., 2020; Cole and Boutet, 2023)to identify influential works, yet they neglected the semantic nuances of why citations occur. To decode citation motives, research shifted toward citation intent classification. Techniques have evolved from rule-based matching (Valenzuela-Escarcega et al., 2015) to deep learning architectures (CNNs, Transformers) (Budi and Yaniasih, 2022), graph embeddings (Berrebbi et al., 2022), and prompt learning (Lahiri et al., 2023), aiming to capture fine-grained contextual intent. However, these methods typically analyze isolated citation instances, fragmenting the continuous evolutionary process and failing to capture the *Long-range Causal Chains* essential for understanding scientific lineage.

Automated Hypothesis Generation. Early methods mined potential ideas from the static structure of literature, such as identifying missing links between concepts (Rzhetsky et al., 2015) or applying mechanism-transfer analogies (Hope et al., 2017; Amini et al., 2020). Recent agent-based frameworks introduce procedural control over generation. Works like ResearchAgent (Baek et al., 2024) and SciMONSI (Wang et al., 2023) integrate LLMs with knowledge graphs to synthesize novel intersections between previously disjoint concepts. Furthermore, approaches like CoI (Li et al., 2024) construct ordered literature chains to reduce noise, while domain-specific agents like MOOSEChem (Yang et al., 2024) integrate external tools for automated planning. While these methods excel at enhancing *Novelty*, they often fall into the trap of "combinatorial innovation": they tend to stochastically patch concepts together, disregarding the intrinsic logic and historical inertia of scientific development.

3 Method

To mimic the intrinsic logic of scientific discovery and generate well-founded hypotheses, we propose EvoNarrator, a training-free neuro-symbolic framework that reconstructs fragmented literature into cohesive evolutionary narratives. As shown in Figure 2, the framework coordinates macro-topology construction with micro-causal reasoning through four stages: **(1) Evolutionary Graph Construction:** Builds a dynamic citation topology with spatiotemporal awareness; **(2) Deep Semantic Gene Extraction:** Decouples unstructured text into fine-grained scientific entities via multi-agent collaboration; **(3) SocketMatch Causal Inference:** Transforms static citation links into explicit logical operators via structured matching; **(4) Pattern-Driven Generation:** Identifies high-potential evolutionary paths to guide the LLM in synthesizing hypotheses that are both innovative and historically justifiable.

3.1 Evolutionary Graph Construction

We model the target domain literature as a time-varying directed graph $G = (\mathcal{V}, \mathcal{E})$. To comprehensively capture both the established structural core and the emerging frontier without losing relevance, we devise a **Dynamic Hybrid Expansion Strategy**.

First, utilizing GPT-4o to expand the user query

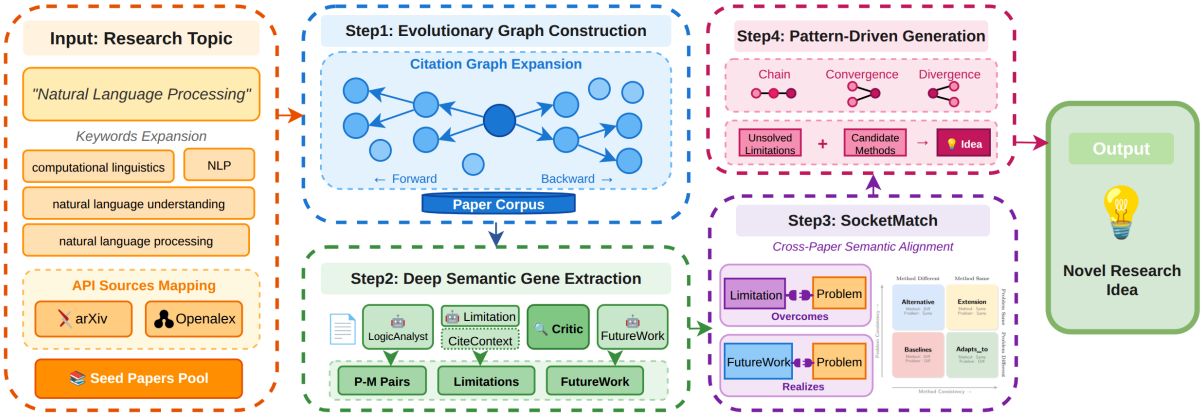

Figure 2: The EvoNarrator framework comprises four phases: (1) **Evolutionary Graph Construction** retrieves and structures the literature; (2) **Deep Semantic Gene Extraction** parses unstructured text into P-M-L-F quadruples; (3) **SocketMatch** infers explicit logical operators between citations; (4) **Pattern-Driven Generation** synthesizes novel hypotheses along identified evolutionary paths.

into a richer semantic description, we initialize seed nodes via a hybrid scoring mechanism. This mechanism balances a paper's global citation impact with its semantic relevance to the expanded query (details in Appendix C.1). This design prevents the graph from being biased toward highly cited but marginally related classics. Subsequently, we perform multi-hop bi-directional citation snowballing anchored on these seeds. Because naive citation expansion often leads to thematic drift, we introduce **Co-citation Gating**: a candidate paper is incorporated only if it is frequently co-cited with papers already in the graph. Intuitively, this ensures that new additions are structurally embedded in the core research conversation rather than being tangential citations. Furthermore, to address the inherent citation lag of purely structural expansion, we implement **Frontier Injection**. We explicitly retrieve recent papers and inject those whose embeddings share high similarity with the graph's overall thematic centroid. Finally, we complete missing edges among all selected nodes to form a dense citation closure containing the full historical context.

3.2 Deep Semantic Gene Extraction

Standard metadata is insufficient for logical reasoning. We parse each paper $d \in \mathcal{V}$ into a structured gene quadruple $\mathcal{G}_d = (P, M, L, F)$ (Problem, Method, Limitation, Future Work). Multi-agent collaboration ensures efficiency and precision: **(1) Causal Alignment:** Rather than independent extraction, we enforce the extractor $\mathcal{A}_{ext}$ to identify "Problem-Solution Pairs," ensuring M is a direct logical response to P. **(2) Dual-Perspective Fusion:** To eliminate author bias, the final limitation L integrates internal self-disclosure (L_{int}) with external criticism from citing papers (L_{ext}), i.e., $L = L_{int} \cup L_{ext}$. Meanwhile, the future work F is directly distilled from the authors' explicit prospective statements without requiring external fusion. **(3) Reflection Loop:** A critic agent $\mathcal{A}_{crit}$ validates completeness and consistency, iteratively refining $\mathcal{G}_d$ until quality thresholds are met.

3.3 SocketMatch

The SocketMatch mechanism reconstructs undirected citation edges into directed evolutionary links with explicit causal logic, leveraging the extracted deep genes (PMLF).

The Plug-Socket Model. We define the predecessor u's elements (M_u, L_u, F_u) as "Sockets" and the successor v's M_v as a "Plug". Analyzing P_u and P_v alongside citation context C_{uv}, an evolutionary relationship is established if and only if the $Plug$ semantically fills the $Socket$. To avoid the ambiguity of traditional similarity matching, we design a discriminant function Φ utilizing the zero-shot reasoning of LLMs to conduct a "scientific review" on $(Socket, Plug)$. The output includes structured JSON containing `is_match`, `reasoning`, and `confidence` (Implementation details in Appendix C.3).

Hierarchical Decision. Since a paper may simultaneously extend and cite a predecessor, we define a priority function $f : \mathcal{M} \to \mathcal{T}$ (Algo-

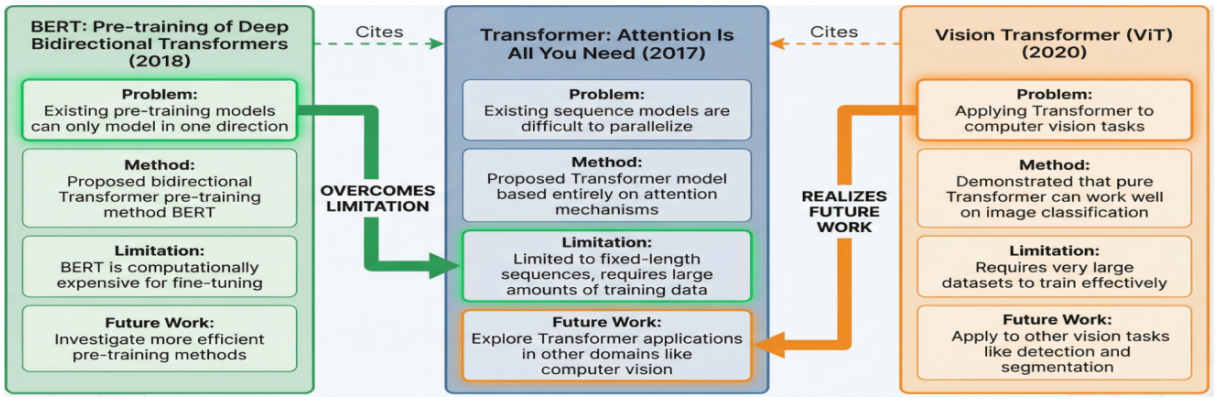

Figure 3: SocketMatch Case Study: Divergent evolutionary paths from a common parent (Transformer).

Algorithm 1 SocketMatch Evolutionary Hierarchy Inference

Require: Predecessor u, Successor v, Genes $\mathcal{G}_u, \mathcal{G}_v$, Context C_{uv}
Ensure: Evolutionary Relation Type $r_{u \to v}$

Initialize:
$Plug_v \leftarrow \mathcal{G}_v.M$, $Socket_u \leftarrow \mathcal{G}_u.M$
$Socket_{Lim} \leftarrow \mathcal{G}_u.L$, $Socket_{Fut} \leftarrow \mathcal{G}_u.F$
$\Phi(\cdot)$: LLM-based semantic entailment function
Step 1: Detect Vertical Breakthrough
if $\Phi_{solve}(Socket_{Lim}, Plug_v, C_{uv}) > \tau$ **then**
  return OVERCOMES {Resolves specific pain point}
end if
Step 2: Detect Blueprint Realization
if $\Phi_{realize}(Socket_{Fut}, Plug_v, C_{uv}) > \tau$ **then**
  return REALIZES {Achieves specific future vision}
end if
Step 3: Detect Horizontal Adaptation
if $\Phi_{reuse}(Socket_u, Plug_v, C_{uv}) > \tau$ **and** $\Delta_{dom}(\mathcal{G}_u.P, \mathcal{G}_v.P) \in \mathcal{K}_{cross}$ **then**
  return ADAPTS_TO {Method reuse across domains}
end if
Step 4: Detect Methodological Evolution
if $\Delta_{dom}(\mathcal{G}_u.P, \mathcal{G}_v.P) \in \mathcal{K}_{similar}$ **then**
  if $\Phi_{inh}(Socket_u, Plug_v, C_{uv}) > \tau$ **then**
    return EXTENDS {Architecture improvement within task}
  else if $\Phi_{div}(Socket_u, Plug_v, C_{uv}) > \tau$ **then**
    return ALTERNATIVE {Competing approach within task}
  end if
end if

return BASELINE {Default: Comparison only}

rithm 1) to resolve conflicts. We follow a **"Necessity First"** principle: *Overcomes* > *Realizes* > *Adapts* > *Extends/Alternative* > *Baseline*. This prioritizes disruptive evolutionary leaps. As illustrated in Figure 3, this logic distinguishes trajectories within the Transformer family: BERT (M_{bert}) resolving Transformer's unidirectionality (L_{trans}) is classified as OVERCOMES, whereas ViT (M_{vit}) actualizing the visual processing vision (F_{trans}) is classified as REALIZES.

3.4 Pattern-Driven Generation

We synthesize feasible hypotheses by identifying evolutionary threads on the logic-enhanced graph.

Pattern Recognition and Pruning. To extract a robust logical skeleton, we prune weak semantic connections (Baseline) via bi-directional BFS. Starting from seeds or high-coverage nodes, we identify three core evolutionary patterns (Figure 4): *Chain* (sequential evolution), *Divergence* (domain expansion), and *Convergence* (paradigm synthesis).

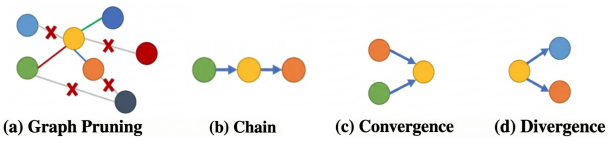

Figure 4: Key Evolutionary Patterns: Chain Iteration, Branching Divergence, and Multi-source Convergence.

Dual-Track Pooling. We automatically construct a problem space (S_{prob}) and a method space (S_{meth}) from the graph topology:

$$S_{\text{prob}} = \left\{ n \in \mathcal{V} \;\middle|\; \begin{array}{l} \text{Out}_{\text{OVER}}(n) = \emptyset \\ \vee\, \text{Out}_{\text{REAL}}(n) = \emptyset \end{array} \right\} \quad (1)$$

This set captures "unconquered" limitations or "unrealized" future works, representing current scientific voids.

$$S_{\text{meth}} = \left\{ m \in \mathcal{V} \;\middle|\; \begin{array}{l} \text{In}_{\text{EXT}}(m) \geq k \\ \vee\, \exists e_{m \to \cdot} \in \text{ADAPTS} \end{array} \right\} \quad (2)$$

This set filters for mature methods with high reuse rates or proven transferability.

Generative Hypothesis Synthesis. The system performs evidence-based reasoning: selecting an open socket $s_{target} \in S_{prob}$ at the end of an evolutionary path and retrieving the best matching plug $m_{cand} \in S_{meth}$, the LLM generates hypothesis H (details in Appendix C.4):

$$H = \text{LLM}(\text{Context}(Path), s_{\text{target}}, m_{\text{cand}}) \quad (3)$$

This mechanism ensures the generated hypothesis is supported by historical logic and methodologically feasible.

4 Experiments

4.1 Experimental Settings

Data. We target "Natural Language Processing" as the domain of interest. To ensure the knowledge base represents a *Dense Logical Closure* rather than a sparse collection, we enforce a strict filtering protocol: performing depth-3 bi-directional snowballing from 5 representative seed papers. This constructs a high-fidelity evolutionary graph containing 226 nodes (papers) and 282 edges (citations) spanning 1980-2025. Notably, despite the compact node count, this subgraph fully captures the evolutionary "skeleton" of the subfield, enabling computationally intensive semantic reasoning that is infeasible on massive raw corpora.

Implementation Details. We employ GPT-4o (Temperature=0.3) as the backbone, generating 17 hypotheses based on the constructed graph. To validate the reliability of *EvoNarrator*, we conduct comparative experiments on the pivotal *Deep Semantic Gene Extraction* and *SocketMatch* modules (detailed evaluation in Appendix A).

Baselines. We benchmark EVONARRATOR against three levels of hypothesis generation systems:

1. **Naive LLM:** Uses standard GPT-4o with a sophisticated "Research Scientist" system prompt (see Appendix C.5), serving as the zero-shot ceiling for general-purpose models.

2. **Standard RAG:** Implements a canonical Retrieval-Augmented Generation pipeline. It indexes the parsed textual content of the papers and retrieves relevant chunks based on semantic similarity, using GPT-4o to generate ideas based on this retrieved contextsee Appendix C.6.

3. **SOTA Agents:** (i) **CoI-Agent** (Li et al., 2024): A framework organizing literature into ordered chains. We select this to represent "Linear Contextualization," assessing whether stacking prior knowledge is sufficient for high-quality generation without non-linear topological reasoning. (ii) **MOOSEChem** (Yang et al., 2024): A SOTA autonomous agent for chemical synthesis via tool integration. We adapt this "Iterative Planning Paradigm" to our task to investigate if general iterative search (zero-shot transfer) can rival topology-aware reasoning.

4. **Ablation (*w/o Path*):** A variant removing evolutionary path awareness while retaining *Unsolved Limitations* and *Candidate Methods*. This validates whether the logical trajectory is essential or if simple component recombination suffices.

Evaluation Metrics. Scientific creativity is inherently subjective. To mitigate bias, we devise a rigorous dual-track protocol:

- **Criteria:** We assess generated hypotheses across four dimensions: *Novelty*, *Feasibility*, *Theoretical Support*, and *Logical Alignment*. To prevent subjective score inflation, we eschew a standard discrete Likert scale in favor of a fine-grained, condition-based rubric (1.0–5.0). Crucially, this rubric enforces strict structural gatekeepers (Hard Caps)e.g., scoring is capped at ≤ 4.4 if explicit systematic deduction chains are missing. The comprehensive quantitative matrix is provided in Appendix D.

- **Carbon Review (Human):** Ten PhD researchers in relevant subfields conduct a **double-blind review** of anonymized, shuffled hypotheses. We achieved a Fleiss' Kappa (κ) of 0.68, indicating Substantial Agreement.

- **Silicon Review (LLM):** To verify automated evaluation consistency and reduce self-preference bias, we employ GPT-4o, DeepSeek-v3.2, and GPT-3.5.

4.2 Quantitative Analysis

Table 1 presents the comprehensive results. **The Reality Gap.** A critical divergence exists between silicon and carbon evaluations. Naive LLM

Table 1: Comparison of different methods under various agent models.

Method	Eval: GPT-4o					Eval: Deepseek-v3.2				
	Novelty	Feasibility	Theoretical	Logical	Average	Novelty	Feasibility	Theoretical	Logical	Average
CoI	4.73	4.63	4.35	4.38	4.52	4.77	4.60	4.40	4.40	4.54
MOOSEChem	**4.90**	**4.70**	4.40	4.40	4.60	4.80	**4.70**	4.40	4.40	4.58
NaiveLLM	4.80	4.60	4.43	4.40	4.56	4.76	4.57	4.37	4.37	4.52
StandardRAG	4.76	4.68	4.40	4.39	4.55	4.68	4.64	4.24	4.36	4.48
w/o Path	4.77	**4.70**	4.80	4.76	4.75	4.66	4.64	4.70	4.78	4.69
Ours	4.81	4.66	**4.86**	**4.86**	**4.80**	**4.81**	4.59	**4.84**	**4.86**	**4.78**

Method	Eval: GPT-3.5-turbo					Eval: Human				
	Novelty	Feasibility	Theoretical	Logical	Average	Novelty	Feasibility	Theoretical	Logical	Average
CoI	4.40	4.43	4.40	4.40	4.41	4.25	3.75	3.25	3.75	3.81
MOOSEChem	4.60	4.50	4.40	4.40	4.48	4.50	4.00	4.00	4.00	4.13
NaiveLLM	4.66	4.54	4.40	4.40	4.50	4.27	3.81	3.27	3.55	3.70
StandardRAG	4.38	4.38	4.36	4.36	4.37	4.44	**4.46**	4.40	4.40	4.42
w/o Path	4.64	4.60	4.70	4.66	4.65	4.60	4.00	4.10	4.10	4.20
Ours	**4.75**	**4.67**	**4.88**	**4.88**	**4.80**	**4.61**	4.06	**4.70**	**4.80**	**4.54**

excels in GPT-4o scoring (>4.4) but plummets in human assessment (Theoretical: 3.27, Logical: 3.55). This quantitatively confirms that generic models are prone to "*Hallucinated Innovation*"superficially plausible text lacking deep theoretical roots. **Superiority of EvoNarrator.** Our method dominates all human metrics. Notably, in *Logicality*, we outperform the SOTA MOOSEChem by a significant margin of **0.8** (4.80 vs. 4.00). Crucially, we resolve the "Novelty-Feasibility Trade-off." While MOOSEChem proposes novel but risky ideas, EvoNarrator achieves peak Novelty (4.61) while establishing the highest Feasibility baseline (4.06), realizing **"Grounded Innovation."** Surprisingly, the ablation variant *w/o Path* achieves high Feasibility (4.70). A deeper inspection reveals that without evolutionary context, the model tends to propose conservative, incremental combinations of existing methods (safe bets). In contrast, *EvoNarrator* leverages historical trajectories to suggest paradigm-shifting ideas. While this slightly increases implementation risk (Feasibility 4.66), it significantly boosts theoretical depth (4.86 vs 4.80) and ensures the innovation is logically grounded in historical gaps rather than random recombination.

4.3 Win-Rate Trend

To quantify the relative advantage, we conducted a pairwise t-test ($p < 0.05$) to verify statistical significance. We performed 80 rounds of pairwise win-rate analysis (Figure 5), employing position swapping during the comparison to rigorously

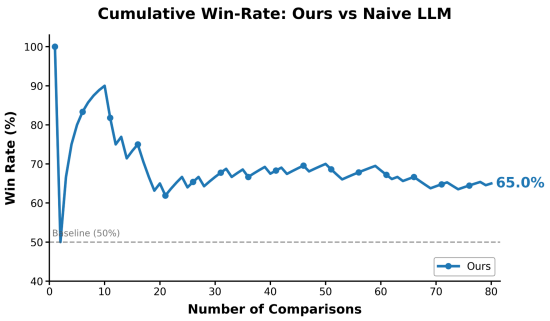

Figure 5: Pairwise Win-Rate Convergence. The x-axis represents the number of evaluation rounds, and the y-axis indicates the cumulative win rate of EvoNarrator against the Naive LLM.

eliminate positional bias. In the supervisor-view blind A/B testing, the win rate of EvoNarrator against the Naive LLM converged and stabilized at **65%+** as the evaluation rounds progressed. This consistent trend confirms that incorporating evolutionary field information enables the model to reliably generate high-quality scientific hypotheses that are preferred by experts, demonstrating statistically significant robustness.

4.4 Hindcasting: The Time-Travel Experiment

The Time-Travel Experiment challenges the agent to forecast future scientific trajectories under strict historical constraints, relying solely on data available up to a specific time point.

To further reduce the possibility that strong results are driven by memorization rather than rea-

Table 2: Analysis of a "Time-Travel" Match: System-generated idea (Input: Pre-2022 data) vs. Real ICLR 2025 Submission (Match Score: 8.0/10).

Generated Idea (Input: Pre-2022 Papers)	**Matched Real Paper (Google Team, 2025)**
Title: Selective Agreement Synthesized **Concept:** Proposes a *synthetic data generation* framework to model opinion dynamics and specifically reduce "sycophancy" (where models incorrectly agree with user opinions to be helpful). The method injects social influence constraints into training data.	**Title:** Simple synthetic data reduces sycophancy in LLMs **Abstract Core:** The paper focuses on *sycophancy* in large language models. It demonstrates that fine-tuning on simple *synthetic data* (QA pairs with specific opinionated features) significantly reduces this behavior and increases robustness.
Evaluation Reason: Both the generated idea and the real paper identify "sycophancy" as a critical failure mode. Crucially, both independently propose "synthetic data intervention" as the core solution. The methodological alignment is high.	

soning, we include an additional NLP-focused time-travel experiment in the Appendix A.3 that uses only papers published before 2024, given that GPT-4o was released in mid 2024, and evaluates predictions against NeurIPS 2024 accepted papers.

Protocol. To ensure a rigorous assessment of foresight untainted by *hindsight bias*, we established the following **Time-Travel Evaluation** protocol.

- **Temporal Cut-off:** The evolutionary graph is strictly constructed using papers published **prior to 2022**. All information from 2023 onwards is masked during inference.
- **Future Ground Truth:** We utilize the **ICLR-Dataset** (sleeping ai, 2024) as the oracle. By filtering for accepted papers between **2023 and 2025** (totaling 9,860 entries), we ensure the target set consists of high-quality, peer-reviewed scientific contributions rather than random text.
- **Matching Mechanism:** We employ a two-stage verification process: (1) *Retrieval*: For each generated hypothesis, we retrieve the top-K semantically similar papers from the ground truth using vector similarity. (2) *Scoring*: An external LLM judge evaluates the alignment between the hypothesis and the retrieved papers on a scale of 0–10, considering problem definition, methodological similarity, and theoretical depth.

Metrics. We report the **Hit Rate** (percentage of hypotheses with a match score $\geq$ 5.0 indicating meaningful relevance) and **High-Fidelity Rate** (score $\geq$ 7.0 indicating strong methodological overlap). Additionally, we analyze the Score Distribution across the generated batch to verify stability.

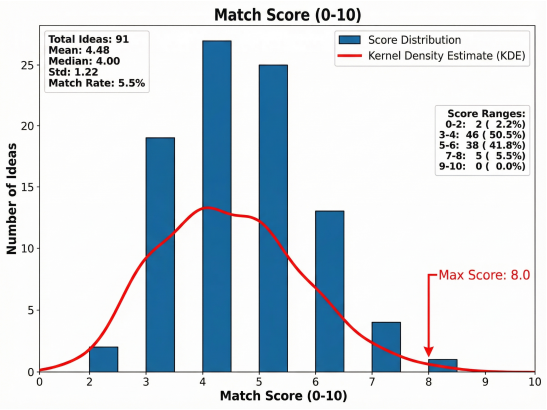

Figure 6: Distribution of Foresight Scores. The histogram illustrates the semantic alignment scores between our generated hypotheses (based on pre-2022 data) and actual ICLR 2023–2025 accepted papers.

Results. Among the 80 hypotheses generated by EvoNarrator, **41.25%** achieved a meaningful relevance score ($\geq$ 5.0), significantly surpassing the baseline of random combination. The overall distribution of these semantic alignment scores is illustrated in Figure 6. The histogram shows a central distribution, indicating that while the majority of ideas achieve moderate relevance (4.0–6.0), the system is capable of producing high-fidelity predictions reaching scores up to 8.0. Table 2 highlights one such striking instance (Score: 8.0) where our system proposed "Selective Agreement Synthesized" to mitigate LLM sycophancy. This idea coincides precisely with a Google DeepMind submission to ICLR 2025, *"Simple synthetic data reduces sycophancy in LLMs."* Both works independently identify the same failure mode and propose a "synthetic data intervention" as the core solution. To understand the nature of these high-scoring hypotheses, we performed

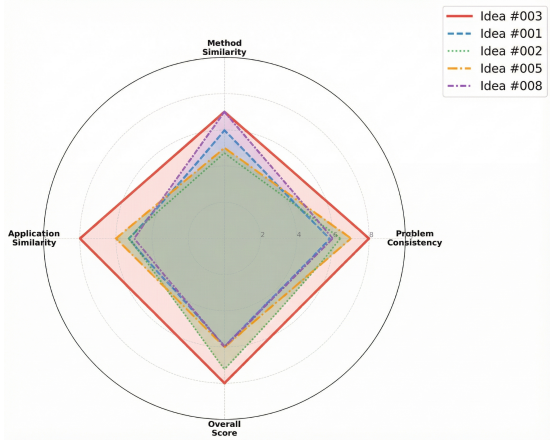

Figure 7: Multi-dimensional Alignment Analysis. A radar chart decomposing the similarity metrics for the top-5 generated ideas across different semantic dimensions.

a multi-dimensional alignment analysis, shown in **Figure 7**. This decomposition reveals that top-performing hypotheses, such as Idea #003 (highlighted in red), demonstrate consistent high alignment across Problem Definition, Methodology, and Application Domain, rather than relying on superficial keyword matching. This suggests that EVONARRATOR does not merely summarize history; it successfully models **"Evolutionary Momentum"**the derivative of scientific trendsenabling the extrapolation of valid future states three years in advance.

5 Conclusion

This work validates the premise that scientific breakthroughs are not stochastic events but ordered evolutions of existing knowledge structures. To address the "Hallucinated Innovation" prevalent in LLMs, we propose **EVONARRATOR**, a framework that shifts the hypothesis generation paradigm from superficial statistical correlation to **constrained causal deduction**. By transforming static citation networks into dynamic reasoning graphs via the SocketMatch mechanism, we ensure that generated innovations are not merely conceptual heaps but precise, structurally aligned responses to historical limitations.

Empirical evidence supports this theoretical shift. Double-blind evaluations confirm that EVONARRATOR surpasses SOTA baselines in logicality and feasibility, while hindcasting experiments demonstrate its foresight in capturing **"Evolutionary Momentum"** to predict high-value research trends.

Ultimately, EVONARRATOR represents a shift from the passive retrieval of history to the active deduction of the future. It highlights a key principle for AI-driven discovery: to authentically infer "What's Next," an agent must possess the cognitive capacity to deeply comprehend "How it Evolved.

Limitations

Despite EvoNarrator's demonstrated proficiency in capturing scientific evolutionary trajectories and generating highly feasible hypotheses, we acknowledge several limitations in the current framework:

- **Inherent Biases in Citation Metrics:** Because our graph construction and expansion mechanisms utilize citation signals (e.g., co-citation gating), the system inevitably inherits known biases associated with citation metrics. These include popularity bias (the "Matthew effect"), where historically highly-cited papers may disproportionately dominate the network structure, potentially overshadowing novel but less-cited niche research. Furthermore, heterogeneous citation practices across different sub-domains can introduce structural skewness into the evolutionary graph.
- **Dependence on Full-Text Availability:** The core precision of *SocketMatch* hinges on the accurate extraction of deep semantic genes (P-M-L-F), which typically reside within the main text rather than abstracts. Due to widespread paywalls and the complexity of parsing heterogeneous PDF formats, the system's reasoning capability significantly degrades to shallow inference when forced to rely on abstract-only data.
- **Inference Latency and Computational Cost:** To guarantee logical rigor, our framework requires intensive LLM operations, including multi-agent reflection loops and pairwise semantic matching. Constructing a dense evolutionary graph with hundreds of nodes incurs substantial token consumption and time latency. Consequently, EvoNarrator is currently better positioned as an in-depth offline analysis tool rather than a real-time conversational assistant.

- **Epistemic Boundaries:** The current P-M-L-F tuple structure is intrinsically optimized for "Problem-Solution" paradigms prevalent in disciplines like Computer Science and Engineering. Applying this framework to fields with distinct argumentative structuressuch as interpretative Humanities (e.g., History) or axiomatic deduction (e.g., Pure Mathematics)will require significant adaptation of the causal extraction logic to accurately capture their evolutionary essence.

Ethical Considerations

We acknowledge that AI-driven idea generation entails risks regarding scientific integrity and bias. Since our model recombines existing knowledge, it may propagate biases inherent in the source literature or pre-trained weights. Furthermore, despite RAG-based grounding, the risk of plausible hallucinations remains. We emphasize that this tool is designed to augment, not replace, human scientific reasoning. Users must strictly verify generated hypotheses and citations. Finally, we condemn the use of such technology for generating harmful or unethical research proposals and rely on underlying model safety alignment to mitigate dual-use risks.

Acknowledgments

We extend our sincere gratitude to the OpenAlex and arXiv teams for maintaining open-access data interfaces, which were instrumental in constructing the evolutionary graph presented in this study. We also thank the domain experts and doctoral researchers who participated in our double-blind "Carbon-based review" experiment; their rigorous evaluation provided essential benchmarks for assessing the feasibility of the generated hypotheses. Finally, we acknowledge the use of AI assistants (e.g., ChatGPT) for grammatical polishing and formatting assistance during the preparation of this manuscript. All scientific claims and intellectual contributions remain solely those of the authors.

This work was supported by the Open Fund of National KeyLaboratory of Parallel and Distributed Computing (PDL)(NO.2024-KJWPDL-02) and the Science and Technology Innovation-Program of Hunan Province (No.2023RC1005) .

Future Work

To transcend the aforementioned limitations and advance scientific evolutionary reasoning, future work will focus on three key directions:

- **Pre-computed Global Socket Graph:** Addressing computational costs, we plan to shift from "on-demand computation" to an "indexing paradigm." We aim to build a persistent, global-scale knowledge base where "Sockets" and "Plugs" from millions of papers are pre-extracted and vectorized. This effectively converts complex $O(N^2)$ online matching into $O(1)$ graph retrieval, enabling millisecond-level trend forecasting.
- **Multimodal Evolutionary Perception:** Scientific evolution is often encoded in non-textual modalities (e.g., new connections in model architecture diagrams or functional group replacements in chemical structures). We will integrate Vision-Language Models (VLMs) to enable EvoNarrator to "perceive" visual evolutionary deltas, aligning visual information with textual logic to capture structural innovations that are difficult to describe via text alone.
- **Interactive Human-AI Reasoning:** Current outputs are static reports. We aim to evolve the system into an interactive "Research Sparring Partner." By incorporating Reinforcement Learning from Human Feedback (RLHF), the system will dynamically prune evolutionary trees based on researcher input or pose counter-factual questions (e.g., *"Would Method A still apply if constraints were relaxed?"*) to break cognitive fixation and achieve true human-machine co-creation.

References

Aida Amini, Tom Hope, David Wadden, Madeleine van Zuylen, Eric Horvitz, Roy Schwartz, and Hannaneh Hajishirzi. 2020. Extracting a knowledge base of mechanisms from covid-19 papers. *ArXiv*, abs/2010.03824.

W. Brian Arthur. 2009. *The Nature of Technology: What It Is and How It Evolves*. Free Press, New York.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *ArXiv*, abs/2404.07738.

Xiaomei Bai, Mengyang Wang, Ivan Lee, Zhuo Yang, Xiangjie Kong, and Feng Xia. 2020. Scientific paper recommendation: A survey. *IEEE Access*, 7:9324–9339.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. *Companion Proceedings of the Web Conference 2022*.

Indra Budi and Yaniasih Yaniasih. 2022. Understanding the meanings of citations using sentiment, role, and citation function classifications. *Scientometrics*, 128:735–759.

Victor E. Cole and Mish Boutet. 2023. Researchrabbit. *The Journal of the Canadian Health Libraries Association*, 44:43 – 47.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2025. Can large language models unlock novel scientific research ideas? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33551–33575.

Avishek Lahiri, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Citeprompt: Using prompts to identify citation intent in scientific papers. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–55.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xinxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Li Bing. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *ArXiv*, abs/2410.13185.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Andrey Rzhetsky, James G Foster, Ian T Foster, and James A Evans. 2015. Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569–14574.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

sleeping ai. 2024. Iclr-dataset: A comprehensive collection of iclr papers. https://huggingface.co/datasets/sleeping-ai/ICLR-Dataset. Accessed: 2024-05-20.

Marco Antonio Valenzuela-Escarcega, Vu A. Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Scimon: Scientific inspiration machines optimized for novelty. In *Annual Meeting of the Association for Computational Linguistics*.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024. Moosechem: Large language models for rediscovering unseen chemistry scientific hypotheses. *ArXiv*, abs/2410.07076.

A Module Experiments

To comprehensively evaluate the EvoNarrator framework, we designed two distinct experiments to verify the quality of deep information extraction and the accuracy of citation relationship classification.

A.1 Deep Information Extraction

As the cornerstone for constructing the evolutionary graph, we first assess the system's capability to parse the semantic "DNA" of individual papers.

Setup. We manually annotated a benchmark dataset comprising 79 papers enriched with deep PMLF information. We compared our approach against three baselines: **Naive LLM (GPT-4o), Standard RAG, and LLM-RAG**. To evaluate performance comprehensively and rigorously, we reported **ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), and BLEU scores (Papineni et al., 2002)**. These are standard benchmarks in the field for measuring n-gram overlap between generated and reference texts, reflecting fluency and structural integrity. Considering the limitation of n-gram metrics in capturing synonymous expressions, we introduced **BERTScore**. By leveraging pretrained contextual embeddings, it provides a more robust evaluation of deep semantic similarity between generated content and reference text. To overcome the shortcomings of automated metrics in judging factual correctness and logical coherence, we employed LLM-based evaluation methods (**LLM-Sim, LLM-Cov, LLM-Acc**). Using GPT-4o as an evaluator, we scored results based on semantic similarity, key information coverage, and factual

Table 3: **Main Results for Deep Information Extraction.** EvoNarrator achieves significant advantages in both semantic accuracy (LLM-Acc) and structural capture capability (ROUGE-2).

Method	R-1	R-2	R-L	BLEU	BERTScore	LLM-Sim	LLM-Cov	LLM-Acc
LLMRAG	0.0205	0.0031	0.0190	0.0000	0.5769	0.5375	0.4756	0.7729
PureRAG	0.2504	0.0410	0.1448	0.0112	0.8075	0.3647	0.2862	0.7856
NaiveLLM	0.3676	0.1084	0.2201	0.0344	0.8182	0.6717	0.5980	0.8595
Ours	**0.4126**	**0.1608**	**0.2549**	**0.0673**	**0.8463**	**0.7325**	**0.6576**	**0.8981**

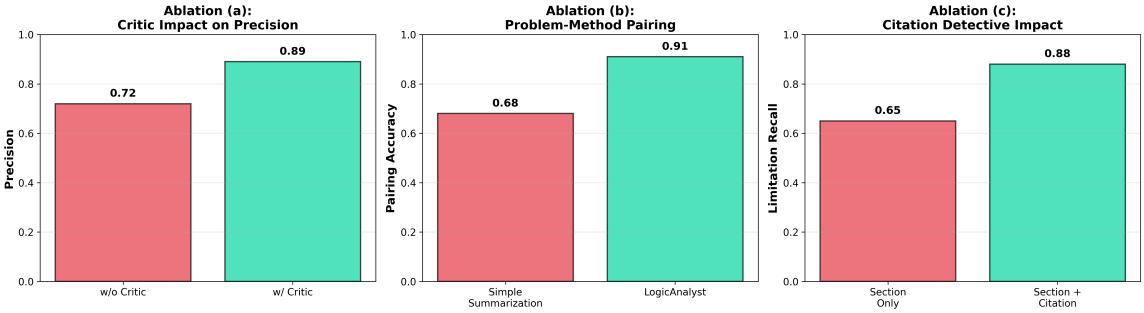

Figure 8: Ablation Study Results. (a) The **Reflection Loop** improves extraction Precision; (b) **Causal Reasoning** significantly corrects the "Problem-Method" pairing logic; (c) **Citation Analysis** (Full System) effectively excavates implicit limitations (Limitation Recall).

accuracy. This ensures our evaluation focuses not merely on "superficial resemblance" but on "truthfulness and accuracy."

Result. As shown in Table 3, EvoNarrator achieves SOTA performance across all 8 metrics. The most critical finding is that the reasoning capability introduced by the *LogicAnalyst* boosts LLM-Acc (factual accuracy) to **0.8981**, significantly outperforming retrieval-based baselines (≈0.78). This implies the system successfully overcomes logical disconnects in long documents, constructing a trustworthy structured database.

Component Ablation Study. To verify the contribution of each module, we conducted an ablation study (see Figure 8). Introducing the CriticAgent's reflection loop increased extraction Precision from 72

A.2 Citation Understanding

Building on the high-precision performance of single-paper information extraction, this section further evaluates SocketMatch's ability to capture the inter-paper Evolutionary Logic.

Setup. We constructed a manually annotated benchmark dataset containing 230 citation pairs, categorized into six evolutionary classes: Overcomes, Realizes, Extends, Alternative, Adapts_to, and Baseline. To verify effectiveness, we compared SocketMatch against a direct Large Language Model approach (Naive GPT-4o). We employed a multi-dimensional metric system: at the class level, Precision (P) and Recall (R) measure prediction accuracy and coverage respectively, with F1-Score as their harmonic mean. At the global level, in addition to overall Accuracy (Acc.), we introduced Macro F1 (Mac. F1) to evaluate class performance equally regardless of sample size, and Weighted F1 (Wgt. F1) to reflect comprehensive performance weighted by sample distribution.

Result. As shown in Table 4, compared to the Abstract-Only baseline, SocketMatch improved **Macro F1** by **0.3109**. This result strongly demonstrates that superficial vocabulary in abstracts alone cannot distinguish deep scientific logic; structured deep information is essential. Specifically, the improvement of 0.2643 in the *Overcomes* category indicates the system's ability to precisely model the deep correspondence between Limitation and Method, thereby distinguishing substantive technical breakthroughs from ordinary citations. Table 5 further presents fine-grained class performance. SOCKETMATCH achieved a qualitative leap from 0 to 20% in the *Realizes* category (F1=0.0 in baseline), as the baseline could not capture cross-paper evolutionary clues not explicitly stated in abstracts. In contrast, SOCKETMATCH successfully identified this

Table 4: **Main Results for Citation Relationship Classification.** SocketMatch substantially outperforms baselines on overall metrics, achieving breakthroughs particularly in key categories representing scientific evolution.

Method	Overall Metrics			Per-Class F1-Score						Avg. F1
	Acc.	Mac. F1	Wgt. F1	Over.	Real.	Ext.	Alt.	Adpt.	Base.	
Baseline (Abstract Only)	0.2826	0.1506	0.3561	0.0690	0.0000	0.1053	0.1270	0.1519	0.4502	0.1506
SocketMatch (Ours)	**0.7174**	**0.4615**	**0.7252**	**0.3333**	**0.2069**	**0.5714**	**0.3529**	**0.4516**	**0.8529**	**0.4615**
Improvement	*+0.4348*	*+0.3109*	*+0.3691*	*+0.2643*	*+0.2069*	*+0.4661*	*+0.2259*	*+0.4027*	*+0.4027*	*+0.3109*

Table 5: **Detailed Breakdown.** Class-wise Precision (P), Recall (R), and F1 comparison. *Base*: Baseline, *Ours*: SocketMatch.

Class	Supp.	Baseline			SocketMatch (Ours)		
		P	R	F1	P	R	F1
Overcomes	25	25.0	4.0	6.9	**54.6**	**24.0**	**33.3**
Realizes	10	0.0	0.0	0.0	**15.8**	**30.0**	**20.7**
Extends	5	6.1	40.0	10.5	**44.4**	**80.0**	**57.1**
Alternative	9	7.4	**44.4**	12.7	**37.5**	33.3	**35.3**
Adapts_to	10	8.7	60.0	15.2	**33.3**	**70.0**	**45.2**
Baselines	171	86.7	30.4	45.0	**87.7**	**83.0**	**85.3**
Macro Avg.	-	22.3	29.8	15.1	**45.6**	**53.4**	**46.2**

"spatiotemporal" inheritance relationship by explicitly matching the future outlook of previous work (P_{prev}.Future) with the methodology of current work (P_{curr}.Method).

A.3 "Time-Travel" Validation against NeurIPS 2024

A key concern in hindcasting-style "time-travel" evaluation is potential contamination from model pretraining data, especially when the backbone LLM may have been trained on a portion of the designated "future" papers. To further reduce the possibility that strong performance is driven by memorization rather than genuine evolutionary reasoning, we conduct an additional experiment in the NLP domain with a stricter temporal constraint. Specifically, we construct the knowledge base strictly from papers published before 2024, and evaluate the generated ideas against accepted papers from NeurIPS 2024. We use GPT-4o-2024-11-20 with a knowledge cutoff of Oct 2023, so the model has no access to 2024 publications.

Experimental setup. Domain: Natural Language Processing (NLP). **Knowledge base:** a citation graph built strictly from pre-2024 papers (695 nodes, 757 edges). **Model:** GPT-4o-2024-11-20 (knowledge cutoff: Oct 2023). **Ground truth:** NeurIPS 2024 accepted papers.

Quantitative results. The system generated 25 ideas. When evaluated against NeurIPS 2024 accepted papers using the same retrieval-and-LLM-judge matching mechanism, 21 out of 25 ideas achieved a match score $> 5.0/10$, indicating strong alignment with frontier research trends.

Qualitative case study: Uncertainty in Transformers. Table 6 presents a representative match (score: 7.0/10). EvoNarrator identifies a concrete "pain point" (unreliable confidence and miscalibration in Transformer models) and proposes a Bayesian-inspired calibration direction. The matched NeurIPS 2024 paper further specializes this direction into the practically important setting of parameter-efficient fine-tuning (PEFT). This suggests that EvoNarrator can generate valid research prototypes that anticipate top-tier publication trajectories under a strict time cutoff.

B Visualization System

To intuitively demonstrate the evolutionary trajectories between papers and the results of generative analysis, we designed and implemented a web-based interactive visualization system. This system integrates the backend-constructed Citation Graph with RAG analysis results, rendering them in the browser via an HTML/JavaScript technology stack.

The main interface employs a **Time-based Layout** algorithm. The X-axis represents the publication year, while node distribution along the Y-axis is algorithmically optimized to prevent overlap. The **Visual Encoding** of nodes reflects paper metadata: node size maps to citation count via logarithmic scaling, and the color gradient indicates the publication year. The visual styles of **Edges** correspond to the six citation relationship types inferred by "Socket Matching"; for instance, "Overcomes" relations are depicted with thick solid red lines to highlight the backbone paths of technological breakthroughs.

The right-hand side features a tabbed functional panel providing multi-dimensional interaction with deep information, specifically comprising the following three core modules:

Table 6: Analysis of a "Time-Travel" Match: System-generated idea (Input: Pre-2024 data) vs. Real NeurIPS 2024 Accepted Paper (Match Score: 7.0/10).

Generated Idea (Input: Pre-2024 Papers)	**Matched Real Paper (NeurIPS 2024)**
Background: Modern neural networks, including large pre-trained Transformer models, often exhibit over- and under-confidence, leading to unreliable uncertainty estimates. **Gap:** Existing Transformer pipelines lack intrinsic mechanisms for calibrated confidence scoring, especially under distribution shift and low-data regimes. **Proposed method:** Introduce an additional uncertainty calibration component, using Bayesian-inspired modeling or temperature scaling style calibration, to produce better-calibrated predictive uncertainty.	**Abstract core:** PEFT adaptation of foundation models can yield accurate but severely underconfident models, especially in few-shot learning. The paper proposes a lightweight *Bayesian Parameter-Efficient Fine-Tuning (Bayesian-PEFT)* framework that integrates PEFT with Bayesian components, supported by theoretical analysis, to achieve reliable and well-calibrated uncertainty quantification.
Evaluation reason: Both the generated idea and the real paper identify the same core failure mode, namely miscalibrated confidence and unreliable uncertainty in Transformer-based models, and both adopt Bayesian-inspired intervention as the key solution. The system proposes a general calibration direction, while the real paper instantiates it in the high-demand PEFT setting.	

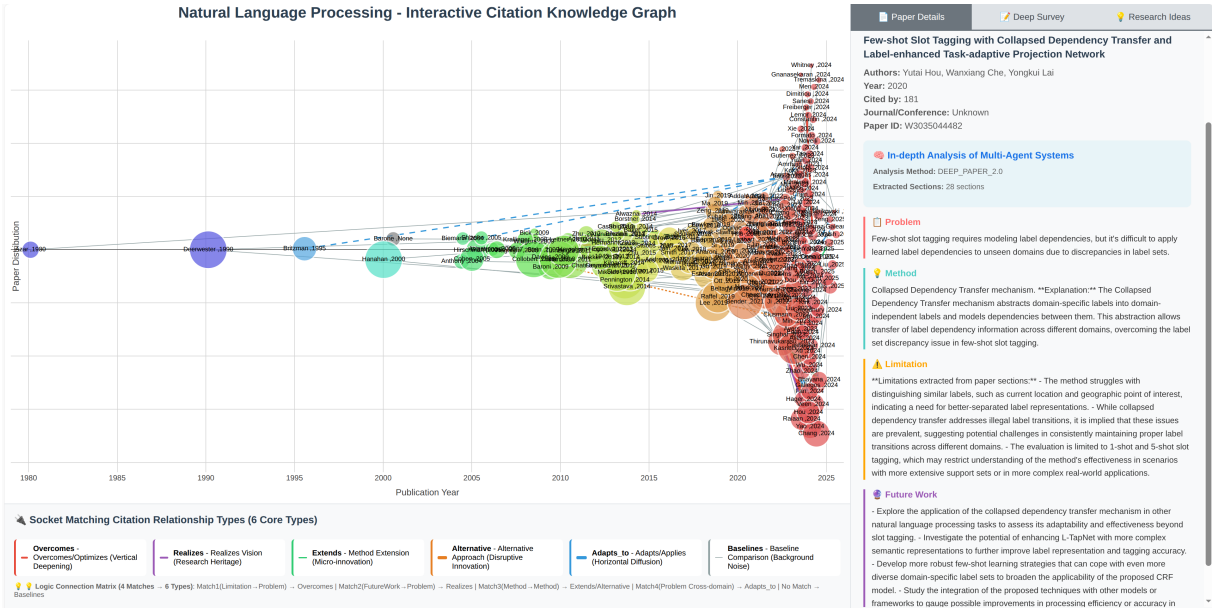

Figure 9: Paper Details Panel: Displays node metadata and structured analysis results extracted by RAG (Problem, Method, Limitation, etc.).

B.1 Paper Details Inspection

As shown in Figure 9, clicking any node in the graph instantly displays the paper's metadata and RAG-based deep analysis results in the details panel. Unlike traditional abstract displays, the system utilizes the `enriched_papers` data structure to present the "Problem," "Method," "Limitation," and "Future Work" extracted by the LLM in a structured manner. This design enables researchers to rapidly assess a node's role and contribution within the knowledge network without downloading or reading the full PDF.

B.2 Evolutionary Review Navigation

Figure 10 illustrates the deep review report generated based on graph pruning algorithms. This module displays not only filtering statistics from the original paper set to core paths (e.g., retention rate, seed paper count) but, more importantly, visual **"Evolutionary Storylines."** Each card represents a technological evolution trajectory identified via Chain or Star patterns, containing narrative text generated by the LLM. The module supports bidirectional interaction: clicking a specific storyline card automatically highlights relevant nodes and edges in the main graph view while graying out irrelevant ones, clearly presenting the origin and development of specific techni-

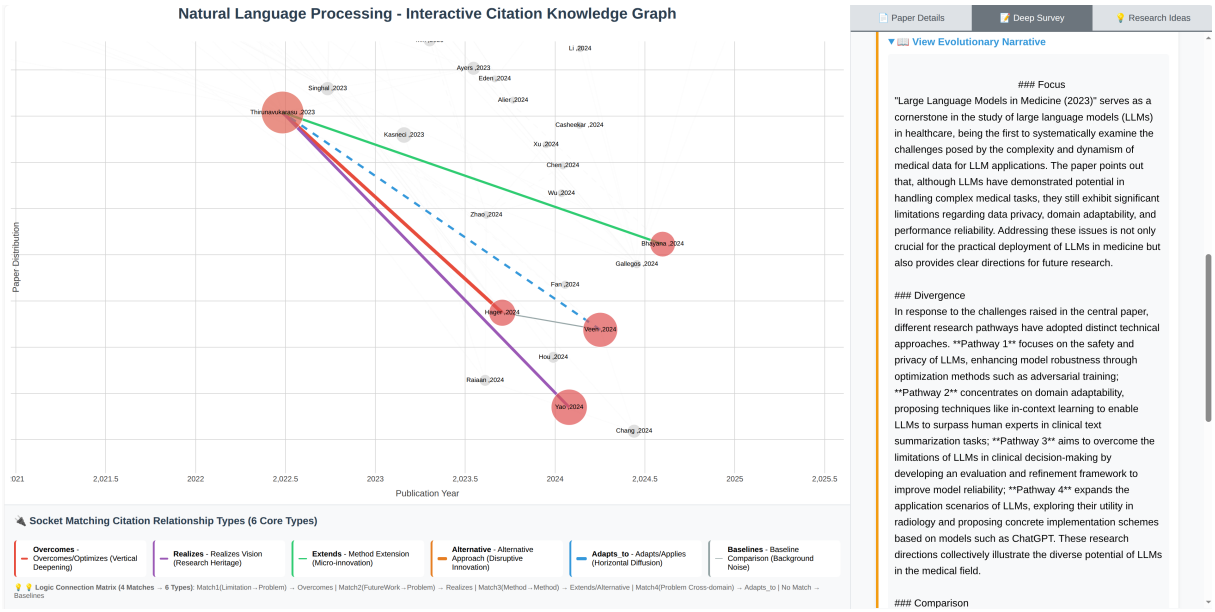

Figure 10: Evolutionary Review Panel: Displays automatically generated evolutionary storylines and graph pruning statistics, allowing users to click cards to highlight evolutionary paths in the graph.

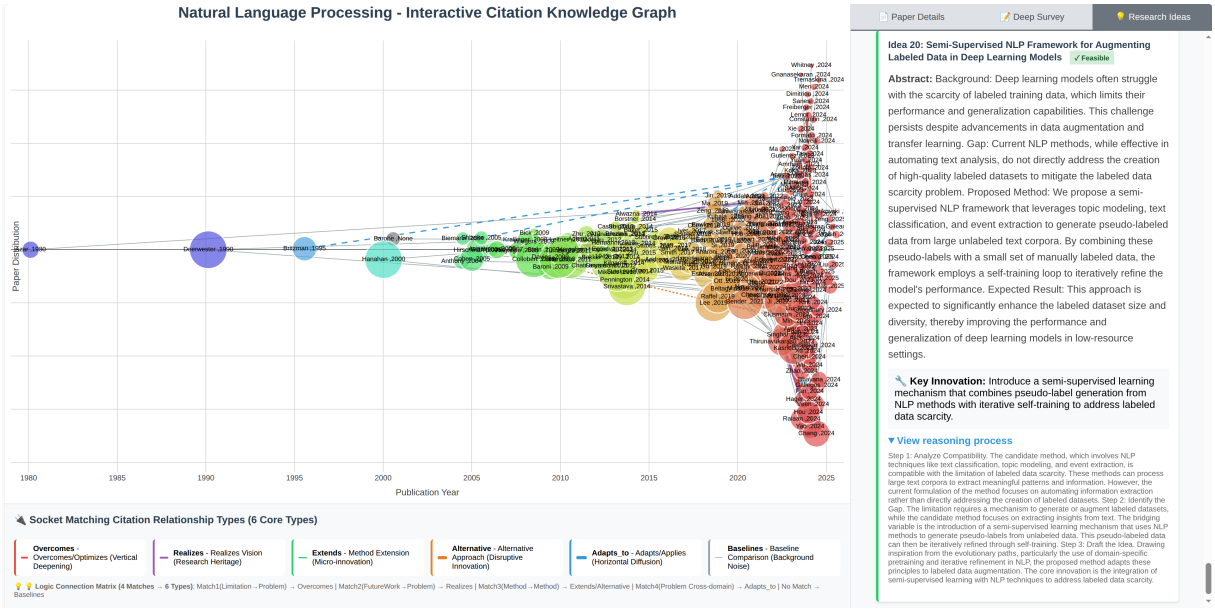

Figure 11: Scientific Idea Panel: Displays innovations generated by the system based on the collision of limitations and method pools, including feasibility assessments and reasoning chains.

cal branches.

B.3 Scientific Idea Generation

As shown in Figure 11, the system integrates a scientific idea generation module based on graph reasoning. This module showcases research ideas generated by the system through the collision of the "Unsolved Limitations Pool" and the "Candidate Methods Pool." Idea cards explicitly indicate feasibility status (e.g., SUCCESS or INCOMPATIBLE) and detail the idea abstract, key modifications, and the complete Chain of Thought (CoT) reasoning process. This not only exhibits the current state of knowledge but also provides researchers with potential research directions logically deduced from existing literature.

Regarding technical implementation, the backend uses **NetworkX** to construct the graph structure and serialize all analysis data into JSON. The frontend utilizes the **Plotly.js** engine for high-performance SVG/Canvas rendering, ensuring a fluid interactive experience even when processing complex networks containing hundreds of nodes and edges.

C Prompts

C.1 Evolutionary Graph Construction

We employ Large Language Models as a domain knowledge base to perform semantic-level expansion and deconstruction of the user's initial query intent. This process generates an extended query vector containing synonyms, hyponyms, and domain-specific terminology. The corresponding prompt is presented below:

expand semantic concepts

System Prompt: You are a domain expert in computer science research. Your task is to expand research topics and keywords by providing semantically related concepts, synonyms, subfields, and alternative terminology. Focus on: Computer Science, AI, and Machine Learning domains; academic and technical terminology; both broad and specific related concepts.

User Prompt: Research Topic: {topic}
Current Keywords: {keywords}
Please expand this research area by providing:

1. Related Topics: 2-{max_topics} semantically similar or overlapping research topics
2. Expanded Keywords: 5-{max_keywords} additional relevant technical terms, methods, or concepts
3. Synonyms: 2-4 alternative terms or abbreviations for the main topic
4. Subfields: 2-3 more specific subfields or applications within this area

Important:

- Focus on computer science and AI-related terms
- Use technical/academic terminology
- Keep each item concise (1-5 words)
- Avoid generic terms like "research", "study", "analysis"

Output: Output ONLY valid JSON in this exact format (no markdown, no code blocks):
{ "expanded_topics": ["topic1", ...], "expanded_keywords": ["keyword1", ...], "synonyms": ["synonym1", ...], "subfields": ["subfield1", ...] }

C.2 Deep Semantic Gene Extraction

The system prompts for Deep Semantic Gene Extraction are presented below:

Logic Analyst Agent

Role: You are a professional expert in scientific paper logic analysis.
Task:
1. Identify the **core pain point** (The Lock/Problem) the authors aim to solve.
2. Identify the **core mechanism/solution** (The Key/Solution) designed by the authors.
3. Explain **specifically how** this mechanism addresses the pain point.

Output Requirements:
- Format: "Problem-Solution Pairs".
- Specificity: Problem descriptions must be precise; avoid generalizations.
- Focus: Solutions must focus on the core mechanism, not technical trivia.
- Logic: Clearly explain the causal relationship ("how it solves").
- Quantity: Focus on the top 1-3 core pairs.

Notes: Do not simply repeat the abstract. Distill the logic. Focus on "why this design solves this problem".
Output Format: JSON only: { "pairs": [{ "problem": string, "solution": string, "explanation": string, "confidence": float, "evidence": string }] }

Limitation Extractor

Role: You are a professional expert in scientific paper analysis.
Task: Extract the **limitations** of the method proposed in this paper.

Important Distinction:
✓ **Extract:** Limitations of "this paper" / "our method".
× **Ignore:** Criticisms of "prior work" or "baselines".
× **Ignore:** Weaknesses mentioned in "Related Work".
Identification Techniques:
- Look for: However, Unfortunately, Limitation, still, yet.
- Focus on self-references: our method, our approach.
- Check the end of Discussion/Conclusion for honest disclosures.

Output Requirements:
- List 2–4 specific limitations using bullet points.
- 1–2 sentences per limitation. Direct output only (no meta-talk).

Future Work Extractor

Role: You are a professional expert in scientific paper analysis.
Task: Extract the **future work directions** proposed in this paper.
Identification Techniques:
- Look for keywords: future, next, further, explore, plan, will.
- Focus on explicit mentions of "future work".
- Infer improvements from acknowledged limitations.
- Check the outlook at the end of the Conclusion or Discussion.

Output Requirements:
- List 2–4 specific directions using bullet points.
- 1–2 sentences per direction. Direct output only (no meta-talk).

Citation Detective Agent

Role: You are a professional expert in scientific literature analysis.
Task: Extract **genuine, specific limitations** from citation contexts.
Input: A citation context (text where a citing paper references the target).
Output Requirements:
1. Determine if the citation contains **critical evaluation**.
2. Extract specific limitation descriptions (avoid generalizations).
3. Summarize using clear, objective language.

Notes:
- Extract only **actual criticisms**; ignore positive/neutral feedback.
- Avoid vague phrases like "has certain limitations".
- If neutral, return has_limitation=false.

Output Format: JSON only:
{ "has_limitation": bool, "limitation": string, "confidence": float, "reasoning": string }

Section Locator Agent

Role: You are an expert in analyzing the structure of academic papers.
Task: Accurately locate specific information within the paper.
Input Description:
- Paper structure: Multiple sections with title and section_type.
- Logic: Judge based on headers and content preview.

Key Distinctions:
- **Limitation:** Authors admit faults of their own method (not others).
- **Future Work:** Future directions proposed by the authors.

Notes:
- Check the end of Discussion or Conclusion.
- Keywords: However, Unfortunately, remains to be.

Output Format: JSON only:
{ "target_sections": [indices], "reasoning": "...", "confidence": float }

C.3 SocketMatch

The system prompts for SocketMatch are presented below:

match limitation problem

Role: Expert in AI citation analysis.
Task: Determine if Paper B (Citing) aims to overcome a limitation of Paper A (Cited).

[Input Information]

- **Paper A:** Title: {cited_title}; Self-reported limitation: {cited_limitation}
- **Paper B:** Title: {citing_title}; Research Goal: {citing_problem}; Citation Context: {citation_context}

[Analysis Logic]

1. **Context Check:** Does B use negative markers (e.g., "However", "fails to", "drawback")?
2. **Alignment Check:** Does B's goal specifically address the critique?

[Judgment Criteria]

- **True:** B explicitly criticizes A AND B's goal resolves it; OR A reports a flaw AND B resolves it.
- **False:** B cites A only as background/baseline.

[Output JSON] { "is_match": bool, "confidence": float, "reasoning": str, "evidence": str }

match futurework problem

Role: Expert in NLP academic paper relationship analysis.
Task: Detect whether Paper B implements a specific plan envisioned in the Future Work of Paper A.

[Core Challenge] Distinguish "True Realization" from generic statements. If A's future work is boilerplate (e.g., "improve accuracy"), B constitutes a Method Extension, not Realizes.

[Input Information]

- **Paper A:** Title: {cited_title}; Future Work: {cited_future_work}
- **Paper B:** Title: {citing_title}; Goal: {citing_problem}; Context: {citation_context}

[Judgment Criteria]

- **Specific (High Confidence):** A proposes a concrete direction (e.g., new modality, specific algorithm integration) and B implements it.
- **Generic (Low Confidence):** A only mentions general performance boosts.

[Output JSON] { "is_match": bool, "specificity": "high/low", "confidence": float, "reasoning": str, "evidence": str }

match method extension

Role: Expert in academic paper relationship analysis.
Task: Analyze whether Paper B extends Paper A (Inheritance) or proposes an alternative paradigm (Substitution).

[Core Pre-condition] Do A and B address similar research problems? If no, output "none".

[Category Definitions]

- **Extension:** B retains A's core backbone with incremental mods (e.g., "based on...", "we extend...").
- **Alternative:** B solves the same problem via a fundamentally different route, explicitly rejecting A's approach (e.g., "Unlike [A]...", "Instead of RNN...").
- **None:** B uses A only as a baseline or background.

[Input Data]

- **Paper A:** {cited_title}, Method: {cited_method}
- **Paper B:** {citing_title}, Method: {citing_method}
- **Context:** {citation_context}

[Output JSON] { "relationship_type": "extension/alternative/none", "confidence": float, "reasoning": str, "evidence": str }

match problem adaptation

Role: Expert in interdisciplinary technology transfer.
Task: Detect if Paper B adapts Paper A's technology to a significantly different domain.

[Core Criterion: Cross-Domain Reusability]

1. **Method Reuse:** B must adopt A's core algorithm/architecture.
2. **Significant Shift:** The task or modality must change fundamentally.

[Judgment Logic]

- **True (Adaptation):** Cross-Modality (e.g., NLP → CV), Cross-Mechanism (e.g., Classification → Generation), or Cross-Discipline (e.g., Text → Biology).
- **False (Extension):** Same task with new data (e.g., English → Chinese), or simple dataset swaps (e.g., IMDb → Yelp).

[Input Data]

- **Paper A & B:** Titles, Problem Domains, and Core Methods.
- **Context:** {citation_context}

[Output JSON] { "is_adaptation": bool, "domain_shift_type": str, "reasoning": str }

C.4 Pattern-Driven Generation

The system prompt for hypothesis generation is presented below:

idea generation

Role: Senior Principal Researcher.
Task: Evaluate if a Candidate Method can solve a Limitation via Chain of Thought.

[Reasoning Process]

1. **Compatibility Analysis:** Check mathematical/algorithmic alignment with the limitation's constraints. If fundamental conflicts exist, output INCOMPATIBLE.
2. **Gap Identification:** Identify the "Bridging Variable"the specific modification needed to adapt the method to the new problem context.
3. **Idea Drafting:** Generate a title and

structured abstract (Background → Gap → Method → Result).

[Evolutionary Context Integration] Utilize provided evolutionary paths (Chain, Divergence, Convergence) to identify successful adaptation patterns and apply them to the current pair.
[Output JSON] { "status": "SUCCESS/INCOMPATIBLE", "rationale": "Complete narrative of the decision chain and evidence.", "idea": { "title": str, "abstract": str, "core_innovation": str } }

C.5 NaiveLLM

The prompt for hypothesis generation with Naive LLM is presented below:

NaiveLLM

System Prompt: You are an expert research scientist with deep knowledge in academic research methodology and scientific innovation. Your task is to generate novel and feasible research ideas based on a given research topic. You must provide a comprehensive research background explaining the current state of the field and a novel research idea that addresses an important gap. Ensure ideas are novel, feasible, well-grounded in theory, and clearly articulated.

User Prompt: Research Topic: {topic} Please generate a novel research idea in the field of **{topic}** meeting the following requirements:

1. **Research Background:** Provide a comprehensive background (2–3 paragraphs) that describes the current state of research, identifies key challenges or gaps, and explains the importance of addressing these issues.

2. **Research Idea:** Propose a novel research idea (1–2 paragraphs) that addresses the identified gap, provides a clear technical approach, and explains expected contributions and novelty compared to existing work.

Output Format: Please provide your response in the following JSON format:
{ "background": "<comprehensive research background in 2-3 paragraphs>", "idea": "<novel research idea in 1-2 paragraphs, including technical approach>" }

C.6 StandardRAG

The prompt for hypothesis generation with Standard RAG is presented below:

StandardRAG

System Prompt: You are an expert research scientist with deep knowledge in academic research methodology and scientific innovation. Your task is to analyze a research paper and generate a novel research idea that builds upon or extends the work presented. The idea must be innovative, feasible, well-grounded, and clearly articulate the technical approach and expected outcomes as a concrete follow-up study.

User Prompt: Please read the following research paper content and generate a novel research idea that builds upon this work.
Paper Content:
{paper_text}
Requirements: Generate a research idea (1–2 paragraphs) that:

- Identifies a specific gap, limitation, or unexplored direction in the paper
- Proposes a novel approach or extension to address this gap
- Explains the technical methodology and expected contributions
- Demonstrates how this idea advances the field beyond the current paper

Output Format: Please provide your response in the following JSON format (no markdown, no code blocks):
{ "idea": "<your research idea in 1-2 paragraphs, including technical approach and expected outcomes>" }

C.7 Hindcasting

The system prompt for Hindcasting experiment is presented below:

Hindcasting

Role: You are a Senior Research Reviewer.
Task: I will provide you with two research ideas: one is a system prediction, and the other is a real published paper. Please determine if their core innovations are essentially the same.
Idea A (System Prediction): idea_text
Idea B (Real Paper): paper_abstract
Criteria: 1. Is the problem addressed consistent? 2. Are the proposed core methods/technical routes highly similar? 3. Are the application scenarios and goals the same?
Output Format (Please strictly follow this JSON format): "match_score": <Integer between 0-10, where 10 means almost identical>, "problem_consistency": <Integer 0-10, score for problem consistency>, "method_similarity": <Integer 0-10, score for method similarity>, "application_similarity":

```
<Integer 0-10, score for application similarity>, "reason": "<Brief explanation for the given scores>" Please output ONLY the JSON, with no additional text.
```

D Evaluation Rubric

To strictly standardize the assessment of generated research ideas, we employed a comprehensive quantitative rubric, as detailed in Table 7. The evaluation covers four critical dimensions: Novelty, Feasibility, Theoretical Support, and Logical Alignment.

Table 7: Evaluation Rubric and Quantitative Criteria for Research Ideas

Dimension	Score	Specific Criteria & Gatekeepers
1. Novelty	5.0	**Truly Revolutionary**: Paradigm-shifting innovation.
	4.7 – 4.9	**Highly Novel**: Innovation with significant advancement and unique perspective.
	4.0 – 4.6	**Clear/Moderately Novel**: Meaningful innovation, or containing some incremental elements.
	< 4.0	**Limited/Minimal**: Primarily incremental improvements with insufficient novelty.
2. Feasibility	5.0	**Perfect Path**: All implementation details are complete and impeccable.
	4.7 – 4.9	**Highly Feasible**: Clear and comprehensive path with detailed specifics.
	4.0 – 4.6	**Very Feasible/Feasible**: Good outline but may lack some details.
	< 4.0	**Moderate/Difficult**: Severe lack of details, or hard to implement with existing resources.
3. Theoretical Support	**4.5 – 5.0**	**[Condition: Must contain "Step 1/2/3" structural tags]** **5.0**: Tags + Outstanding argumentation (>300 words) + Profound insight **4.9**: Tags + Comprehensive argumentation (>250 words) + In-depth analysis **4.8**: Tags + Solid argumentation (>200 words) + Clear analysis **4.7**: Tags + Sufficient argumentation (>150 words)
	$\leq$ **4.4**	**[Hard Cap: No Step Tags]** Even with well-explained solid reasoning, if systematic step tags are missing, the score cannot exceed 4.4.
4. Logical Alignment	**4.5 – 5.0**	**[Condition: Must contain a complete multi-step logic chain]** (Limitation $\rightarrow$ Compatibility $\rightarrow$ Gap $\rightarrow$ Solution $\rightarrow$ Result) **5.0**: Complete chain + Outstanding modification details + Perfect result prediction **4.7 – 4.9**: Complete chain + Comprehensive/Solid modification details **4.5 – 4.6**: Chain exists, details need supplementation
	$\leq$ **4.4**	**[Hard Cap: No Complete Chain]** Explains how the solution works clearly, but lacks a systematic logical deduction chain; score cannot exceed 4.4.