

THE-Tree: Can Tracing Historical Evolution Enhance Scientific Verification and Reasoning?

Xin Wang^{1,2} Jiyao Liu^{2,4} Yulong Xiao⁵ Junzhi Ning^{2,6}

Lihao Liu² Junjun He² Botian Shi^{2,3*} Kaicheng Yu^{1,3*}

¹Westlake University ²Shanghai Artificial Intelligence Laboratory

³Shanghai Innovation Institute ⁴Fudan University

⁵Fuzhou University ⁶Imperial College London

 [Homepage](#)  [Code](#)

{wangxin82, kyu}@westlake.edu.cn, shibotian@pjlab.org.cn

Abstract

Large Language Models (LLMs) are accelerating scientific idea generation, but rigorously evaluating these numerous, often superficial, AI-generated propositions for novelty and factual accuracy is a critical bottleneck; manual verification is too slow. Existing validation methods are inadequate: LLMs as standalone verifiers may hallucinate and lack domain knowledge (our findings show 60% unawareness of relevant papers in specific domains), while traditional citation networks lack explicit evolutionary reasoning and narrative surveys are unstructured. This underscores a core challenge: the absence of structured, verifiable, and historically grounded data tracing scientific evolution. To address this, we introduce **THE-Tree** (Technology History Evolution Tree), a computational framework that constructs such domain-specific evolution trees from scientific literature. THE-Tree employs a search algorithm to explore evolutionary paths. During its node expansion, it utilizes a novel “Think-Verbalize-Cite-Verify” (TVCV) process: an LLM proposes potential advancements and cites supporting literature. Critically, each proposed evolutionary link is then validated for logical coherence, temporal plausibility, and evidential support by a Retrieval-Augmented Natural Language Inference (RANLI) mechanism that interrogates the cited literature, ensuring that each step is historically grounded. We construct and validate 88 THE-Trees across diverse domains and release a benchmark dataset including up to 71k fact verifications covering 27k papers to foster further research. Experiments demonstrate that i) in graph completion, our THE-Tree improves Hit@1 by 8% to 14% across multiple models compared to traditional citation networks; ii) for predicting future scientific developments, it improves Hit@1 by nearly 10%; and iii) when used to augment LLM-based paper evaluation, it yields consistent improvements in

macro-averaged F1 (up to +27 points), with gains concentrated on reject discrimination. By constructing explicit, verifiable pathways of scientific progression, THE-Tree provides a robust historical foundation for evaluating new hypotheses (human or AI-generated) and enables a computable science history, fostering evidence-based AI-driven scientific discovery.

1 Introduction

Automating scientific discovery has been a long-standing goal (Langley, 1987; Hutter, 2000). The recent rise of Large Language Models (LLMs) offers new avenues, with applications from hypothesis generation (Yang et al., 2023; Boiko et al., 2023a; Baek et al., 2024) to simulating autonomous AI scientists (Bran et al., 2023; Boiko et al., 2023b). However, a critical bottleneck remains: the effective evaluation and validation of scientific ideas, whether AI or human-generated.

Current idea validation approaches face several critical challenges. First, manual verification, while ideal, is prohibitively time-consuming (Si et al., 2024). Second, automated validation using LLMs (Huang et al., 2023; Baek et al., 2024) exhibits multiple limitations: (1) potential for hallucination and incomplete domain knowledge (our findings show 60% unawareness of relevant papers in specific domains), (2) susceptibility to superficial textual features, often highly rating plausible but erroneous propositions (Si et al., 2024; Yang et al., 2023), and (3) inheritance of biases when trained on human review data (e.g., CycleReviewer (Weng et al., 2025)). Third, existing knowledge representation methods are inadequate, as (1) citation networks (Bornmann and Daniel, 2008; Hao et al., 2024) contain noise and lack explicit evolutionary reasoning links, and (2) narrative surveys remain unstructured. These challenges fundamentally stem from the absence of structured, historically grounded data tracing scientific evolution.

*Corresponding Author

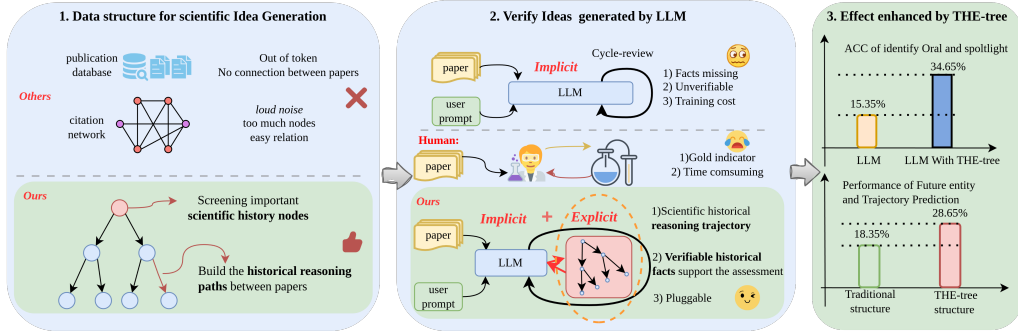


Figure 1: **Overview of THE-Tree.** (1) Limitations of existing data structures (publication databases, citation networks) for scientific idea generation versus THE-Tree’s approach of constructing historical reasoning connections by screening important scientific history nodes and building pathways between them. (2) Methods for verifying LLM-generated ideas: implicit LLM-based cycle-review, human evaluation, and explicit THE-Tree-enhanced LLM verification. (3) Comparison of verification methods, highlighting issues like fact missing in LLM-only approaches, high cost in human evaluation, and THE-Tree’s use of verifiable historical facts and scientific reasoning trajectories. (4) Performance improvements with THE-Tree in different tasks

To address this, we propose leveraging the authentic patterns and historically grounded evolutionary pathways from scientific history for more reliable assessment. We introduce **THE-Tree** (**T**echnology **H**istory **E**volution **T**ree), a computational framework to construct structured, verifiable, domain-specific technology evolution trees from scientific literature (illustrated in Figure 1). THE-Tree builds a topic’s evolution by representing individual papers as nodes and the inferential relationships between these papers, specific to the topic, as edges. This aims to provide a solid factual basis and clear historical context for evaluating new hypotheses.

THE-Tree utilizes a Self-Guided Temporal Monte Carlo Tree Search (SGT-MCTS) and a novel Think-Verbalize-Cite-Verify (TVCV) methodology for node expansion. This process prompts an LLM to generate potential evolutionary steps (Think), summarize them concisely (Verbalize), ground them in specific supporting literature (Cite), and critically, validate the proposed relationship (Verify). Crucially, the ‘Verify’ step employs a Retrieval-Augmented Natural Language Inference (RA-NLI) mechanism to assess the logical coherence, temporal plausibility, and evidential support of proposed relationships based on cited evidence, ensuring that identified evolutionary relationships are historically grounded. Tree construction often starts from human-validated knowledge like scientific surveys, providing a reliable starting point.

We demonstrate THE-Tree’s efficacy by constructing trees for 88 distinct topics across relevant scientific domains. Their quality, validated by automated metrics and human assessment, con-

firms their effectiveness in reconstructing meaningful and accurate technological trajectories. Downstream tasks, including future node prediction and graph completion, showcase THE-Tree’s potential to enhance AI-assisted scientific reasoning by anticipating subsequent developments and highlight its superiority over traditional citation networks. For instance, in graph completion, it outperforms traditional citation networks (e.g., Hit@1 improves by 8–14% across tested models); for future node and trajectory prediction, it improves Hit@1 by up to 10%; and in paper evaluation, it yields consistent improvements in macro-averaged F1 across multiple LLMs, with gains concentrated on reject discrimination. In summary, our main contributions are:

- A novel computational framework, THE-Tree, integrating SGT-MCTS with TVCV node expansion and RA-NLI validation to construct verifiable technology evolution trees from scientific literature.
- A dataset of 88 THE-Trees across AI domains and a benchmark of 71k fact verifications covering 27k papers.
- Extensive experiments demonstrating THE-Tree’s superiority over citation graphs in graph completion and future path prediction, and consistent improvements when augmenting LLM-based paper assessment.

2 Related Work

Our work intersects with several research areas, primarily AI for scientific discovery, scholarly knowledge representation, and the evaluation of scientific novelty.

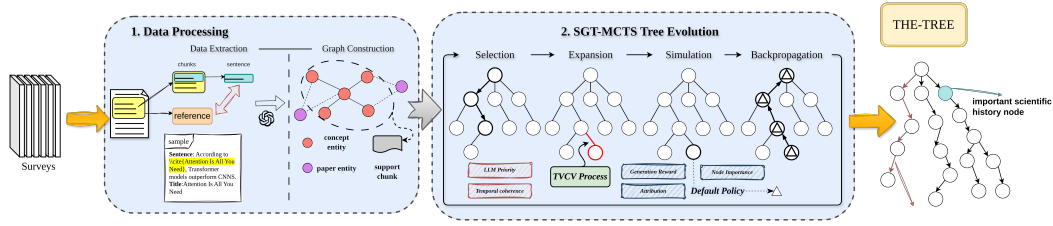


Figure 2: **THE-Tree construction overview.** (a) Extracting chunks and references from surveys to build an initial concept graph structure; (b) Generating a structured knowledge tree using the SGT-MCTS algorithm, guided by the TVCV methodology for node expansion and RA-NLI for relationship validation.

AI for Scientific Discovery and Evaluation.

The ambition to automate scientific discovery has gained momentum with LLMs (Hutter, 2001; Jansen et al., 2025), spanning hypothesis generation (Yang et al., 2023; Boiko et al., 2023a; Baek et al., 2024) and autonomous scientific agents (Bran et al., 2023; Boiko et al., 2023b). A persistent challenge is rigorous evaluation of AI-generated outputs. Manual verification remains a bottleneck (Ferdowsi et al., 2024), and emerging automation efforts primarily simulate existing review processes (Huang et al., 2023; Lu et al., 2024; Weng et al., 2025)—for instance, CycleResearcher (Weng et al., 2025) mimics peer review via an LLM-based research-review loop. While valuable, such approaches model established paradigms rather than providing evaluation grounded in verifiable scientific lineage. Our work complements them by constructing the underlying historical structure itself.

Scholarly Knowledge Representation and Analysis. Traditional bibliometric methods—citation analysis (Garfield, 1979; Small, 1973), co-word analysis (Callon et al., 1983)—and science mapping tools (VOSviewer (Van Eck and Waltman, 2010), CiteSpace (Chen, 2006)) offer insights into impact and thematic trends. However, citation networks are noisy and fail to capture explicit evolutionary dependencies signifying true intellectual inheritance (Bornmann and Daniel, 2008). Co-word analysis identifies term co-occurrence but not evolutionary links. These methods lack the directed granularity needed for reasoning about technological evolution.

Scholarly Knowledge Graphs. Large-scale scholarly knowledge graphs (Microsoft Academic KG (Wang et al., 2020), AMiner (Tang et al., 2008)) integrate diverse metadata, and KG construction techniques (Auer and Stocker, 2018) have been applied to scientific literature. However, these typically focus on entity relationships (author collaborations, affiliations) or static snapshots, without explicitly modeling the dynamic, temporal, directed

evolutionary pathways of scientific ideas. Capturing validated, directed evolution is precisely the gap THE-Tree fills.

3 THE-Tree as Scientific Verifier

3.1 THE-Tree: A Structured Representation of Scientific Evolution

We now formally define the THE-Tree representation. A THE-Tree is a directed graph where nodes represent papers and edges encode validated evolutionary relationships between them.

In a THE-Tree, each scientific paper is a **node** encapsulating metadata (title, abstract, authors, venue, year) and an importance score S_v reflecting its relevance within the domain (Section 3.2.1). The **edges** represent **historical, inferential, and evolutionary relationships** between paper nodes (Figure 2). Unlike citation networks where edges may merely indicate a citation without specifying its nature, THE-Tree edges signify how one paper logically enables or directly precedes advancements in a subsequent paper—reflecting substantive intellectual lineage rather than superficial connections. Our construction methodology (Section 3.2.1) validates such edges using SGT-MCTS and TVCV.

The resulting graph serves as a knowledge scaffold: new propositions can be situated within an explicit, evidence-backed evolutionary context for rigorous assessment of novelty and factual consistency.

3.2 Automated Construction of THE-Tree

The manual construction of THE-Trees for diverse domains is prohibitively laborious. We therefore formulate automated construction as an optimization problem: identify evolutionary paths through the literature that maximize a composite reward reflecting node significance and edge quality:

$$\max_{\text{Path}} \left(\sum_{v \in \text{Path}} S(v) + R_{\text{gen}} + R_{\text{attr}} \right), \quad (1)$$

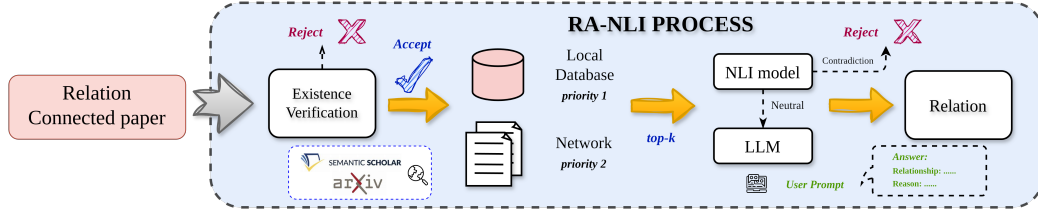


Figure 3: **Overview of RA-NLI process.** This figure illustrates the RA-NLI process, including citation existence verification, document retrieval, and semantic relation assessment using NLI and LLM, which forms the core of the *Verify* step in TVCV and the R_{attr} calculation.

Algorithm 1 SGT-MCTS Implementation

Require: Root node representing the starting technology concept
Ensure: Comprehensive technology evolution tree
1: Initialize the tree with the root node
2: **while** not converged **do**
3: **Select:** Choose promising node
4: $\text{node} \leftarrow \text{select}(\text{root})$ $\triangleright$ SGT-UCT scoring (Eq. 2)
5: **Expand:** Generate new technology nodes
6: $\text{new_nodes} \leftarrow \text{expand}(\text{node})$ $\triangleright$ TVCV methodology (Alg. 2, Sec. 3.2.1)
7: **Simulate:** Evaluate potential paths
8: $\text{reward} \leftarrow \text{simulate}(\text{new_nodes})$ $\triangleright$ Sample-based evaluation
9: **Backpropagate:** Update statistics
10: $\text{backpropagate}(\text{node}, \text{reward})$ $\triangleright$ Improve future selection
11: **end while**
12: **return** Constructed technology tree

Algorithm 2 TVCV Node Expansion

Require: Path history, domain knowledge
Ensure: Validated technology node
1: **Think:** Generate candidate technology T using LLM
2: $T \sim P_{\text{LLM}}(\cdot \mid \text{Path}, \text{domain knowledge})$
3: **Verbalize:** Summarize into concise statement
4: $s_v \leftarrow \text{Summarize}(T)$
5: **Cite:** Retrieve supporting documents
6: $D \leftarrow \text{Retrieve}(s_v)$
7: **Validate:** Verify logical and temporal coherence
8: $\text{Validate}(s_v, D) = \text{RA-NLI}(s_v, D) \wedge \text{TemporalCheck}(D)$
 $\triangleright$ detailed in Sec. 3.2.1
9: **if** Validation successful **then**
10: **return** validated node
11: **else**
12: Reject and restart expansion
13: **end if**

where $S(v)$ is the importance score of node v (Section 3.2.1), R_{gen} reflects path coherence, and R_{attr} validates evidential support for each edge.

Generate-then-verify architecture. A natural concern is that LLMs drive much of the construction (node expansion, importance assessment), yet LLM reliability is the very motivation for THE-Tree. Our resolution is architectural: LLMs operate within a *constrained generate-then-verify* pipeline. During generation (Think–Verbalize–Cite), the LLM proposes candidate nodes and relationships. During verification (Verify via RA-NLI), each proposed edge must independently pass a grounded factual check before inclusion. This separation ensures that LLM hallucinations in the generation phase are caught by the verification layer, as demonstrated by the ablation in Section 4.2: phantom citations are eliminated entirely (21.46% $\rightarrow$ 0.00%) and residual errors drop from 46.73% to 9.32%.

3.2.1 THE-Tree Construction Pipeline Overview

The construction of a THE-Tree involves several key stages, from initial data preparation to the iterative refinement of the tree structure. Figure 2 provides a schematic overview of this pipeline.

Dataset Construction. THE-Tree construction begins with scientific surveys, which encapsulate rich historical and inferential information. We transform these into a structured dataset by extracting paper nodes (from citations), concept nodes (core ideas from paragraphs), and their relationships. For each topic, we select 3–5 surveys spanning different time periods (e.g., one early foundational survey and later comprehensive reviews) to mitigate recency bias and ensure broad temporal coverage. Metadata extraction achieves >97% precision on both IEEE and APA citation formats. The resulting dataset covers 88 topics across AI domains with approximately 27k papers. Further implementation details are in Appendix E.2.

Core Component: Self-Guided Temporal Monte Carlo Tree Search (SGT-MCTS). With the structured dataset as our foundation, we employ SGT-MCTS (Browne et al., 2012) to navigate the vast space of potential evolutionary paths. It iteratively selects nodes for expansion, simulates potential paths, and backpropagates rewards (Algorithm 1), guided by the composite reward in Equation 1.

Central to SGT-MCTS guidance are:

Node Importance ($S(v)$): A weighted combination of structural significance and LLM-assessed

semantic relevance: $S(v) = \gamma \cdot S_{\text{graph}}(v) + (1 - \gamma) \cdot S_{\text{LLM}}(v)$, where S_{graph} combines multiple centrality measures (PageRank, citation count, Degree, Betweenness, Eigenvector) with dynamic weighting.

Generation Reward (R_{gen}): Encourages coherent, temporally sound paths: $R_{\text{gen}}(v|P_{\text{prev}}) = \text{DPO}(v|P_{\text{prev}}) \cdot \text{TemporalCoherence}(v|P_{\text{prev}})$, where the DPO score (Rafailov et al., 2023) evaluates trajectory continuation quality and the temporal term penalizes achronological gaps. The node selection is guided by our SGT-UCT formula:

$$\text{SGT-UCT}(v) = \left(\frac{Q(v)}{N(v)} + c \cdot \sqrt{\frac{\ln N(p)}{N(v)}} + \lambda \cdot \text{LLM}_{\text{priority}}(v) \right) \times \text{TempCoherence}(v | P_{\text{prev}}) \quad (2)$$

which combines exploitation ($Q(v)/N(v)$), exploration ($c \cdot \sqrt{\ln N(p)/N(v)}$), and LLM semantic guidance, modulated by temporal coherence to enforce chronological consistency. We also implement reward-sensitive exploration where the exploration coefficient increases with the average reward.

Node Expansion Method: Think-Verbalize-Cite-Verify (TVCV) Methodology. Node expansion within SGT-MCTS is performed using our Think-Verbalize-Cite-Verify (TVCV) methodology (Algorithm 2). Given a path history, an LLM (1) **Thinks** of candidate advancements that logically follow, (2) **Verbalizes** them as concise propositions representing new nodes, (3) **Cites** specific supporting literature that grounds each proposition, and (4) **Verifies** the proposed relationship via RA-NLI (Section 3.2.1) for logical consistency, evidential support, and temporal soundness.

Relationship Validation: Retrieval-Augmented Natural Language Inference (RA-NLI) Mechanism. The critical *Verify* step and the attribution reward (R_{attr}) are performed by our Retrieval-Augmented Natural Language Inference (RA-NLI) mechanism, which assesses logical coherence, temporal plausibility, and evidential support of each proposed edge:

$$R_{\text{attr}}(v_{\text{parent}} \rightarrow v) = \text{RA-NLI}(s_{\text{parent}}, s_v) \quad (3)$$

As illustrated in Figure 3, RA-NLI integrates embedding-based retrieval of contextual passages from cited documents, followed by fine-grained NLI (f_{NLI}) to determine entailment/contradiction/neutrality between parent and

Table 1: **Hallucination reduction via TVCV (with RA-NLI).** Residual: incorrect relations between *real* papers (phantom-free).

Mechanism	Phantom Cite (%)	Residual (%)
TVC (w/o RA-NLI)	21.46	46.73
TVCV (w/ RA-NLI)	0.00	9.32

child descriptions. An LLM verifier (f_{LLM}) refines ambiguous cases, ensuring high-fidelity evolutionary relationships.

3.3 Leveraging THE-Tree for Scientific Idea Verification

Once a THE-Tree is constructed for a scientific domain, it can be used to explicitly ground and verify a new paper/idea by retrieving its most relevant historical paths in the tree.

Pipeline. Given an input paper P_{in} (title T_{in} , abstract A_{in}), we (i) identify the most relevant topic(s) among pre-built THE-Trees, (ii) retrieve a set of recent anchor papers in the matched THE-Tree(s) based on semantic similarity, and (iii) retrospect historical paths by tracing predecessors from each anchor paper.

Predecessor selection heuristic. Starting from the current node v_{curr} on a backward trace, the predecessor v_{pred}^* is chosen by lexicographical optimization:

$$v_{\text{pred}}^* = \arg \min_{\substack{v \in \text{Pred}(v_{\text{curr}}) \\ Y_v \leq Y_{v_{\text{curr}}}}} ((Y_{v_{\text{curr}}} - Y_v), -S_v), \quad (4)$$

which prioritizes the smallest non-negative time gap and, secondarily, the highest importance score S_v .

Due to space constraints, full details (topic identification, similar-paper anchoring, and historical path retrospection) are in Appendix B.

4 Experiments and Results

Our experiments are designed to answer three research questions:

- **RQ1:** Does the TVCV pipeline with RA-NLI build reliable, historically grounded evolutionary relations? (§4.2)
- **RQ2:** Does THE-Tree provide a better knowledge substrate than citation-only graphs for structural reasoning tasks? (§4.2.2, §4.3)
- **RQ3:** Does historically grounded retrieval from THE-Tree improve LLM-based scientific paper assessment? (§4.1)

For detailed information on dataset construction, statistics, and the definitions of evaluation metrics,

Table 2: Comparison of Methods for Technology Tree Relationship Verification

Method	Fact Missing Rate (%)	Accuracy (%)
Claude 3.5 Sonnet (Anthropic, 2024)	47.93	60.22
GPT-4o (OpenAI, 2024)	58.19	60.18
DeepSeek R1 (Guo et al., 2025)	48.16	76.40
Qwen 2.5-72B (Team, 2024)	58.29	53.95
DeepReviewer-7B (Zhu et al., 2025)	40.88	93.95
DeepReviewer-14B (Zhu et al., 2025)	68.84	76.41
LLaMA 3.1 (Meta AI, 2024)	42.29	60.40
Factual Supplement		
Claude 3.5 Sonnet w/ Fact	-	65.34
GPT-4o w/ Fact	-	69.40
DeepSeek R1 w/ Fact	-	81.60
Qwen 2.5-72B w/ Fact	-	66.80
DeepReviewer-7B w/ Fact	-	95.38
DeepReviewer-14B w/ Fact	-	82.09
LLaMA 3.1 w/ Fact	-	65.35
RA-NLI (Ours)	4.75	95.60

please refer to Appendix F.1 and Appendix F.6 respectively.

To mitigate potential training-data contamination, we apply a temporal cutoff: all evaluated models were released no later than mid-2024, while our paper-evaluation and trend-forecasting targets are drawn from late-2024 and 2025 papers. We summarize additional leakage-prevention measures in Appendix D.

4.1 THE-Tree Verifier for Scientific Evaluation (RQ3)

These experiments investigate whether THE-Tree can enhance the verification capabilities of Large Language Models (LLMs) for evaluating scientific claims. We augment LLM-based assessments with historically grounded evolutionary paths retrieved from THE-Trees, providing factual context that standalone LLMs lack (method details in Appendix E).

Experimental design. NeurIPS 2023 serves as an unaugmented reference baseline. THE-Tree augmentation is applied exclusively to NeurIPS 2024 papers to avoid contamination (models used were released before mid-2024; see Appendix D). We evaluate two scenarios: (1) whether THE-Tree factual grounding improves accept/reject prediction, and (2) whether it helps distinguish high-quality papers (Oral/Spotlight) among accepted submissions. The full per-model breakdown is in Table 8.

Details and a DeepReviewer-14b case study are in Appendix E and Appendix F.7. We acknowledge that conference acceptance is a noisy, subjective signal reflecting reviewer pools and quotas rather than scientific truth; we use it as a meaningful real-world proxy, and report balanced metrics (macro-F1, balanced accuracy) rather than raw accuracy to account for class imbalance and leniency bias.

Aggregate results. Since raw accept/reject accuracy can be misleading (several models exhibit strong leniency bias), we report macro-averaged F1 and balanced accuracy in Table 4. The pattern is consistent: THE-Tree augmentation primarily improves *reject discrimination* rather than uniformly boosting all sub-metrics. The largest gains are observed on DeepReviewer-7b (+27.5 macro-F1) and Claude-3.5-Sonnet (+23.5 macro-F1), while GPT-4o shows modest improvement. We note that THE-Tree augmentation is not uniformly beneficial for all model-metric combinations; accept accuracy may decrease for models that previously exhibited near-total leniency bias (e.g., Claude-3.5-Sonnet: 99.75% $\rightarrow$ 71.46% accept accuracy), as the historical grounding shifts the decision boundary toward more balanced predictions. This is a desirable correction: models with baseline reject-F1 near zero are effectively non-discriminative, and THE-Tree augmentation enables them to identify genuinely weak submissions.

To demonstrate generalizability beyond NeurIPS, we apply THE-Tree augmentation to papers from ICLR 2025, ICML 2025, and CVPR 2025. Table 9 (Appendix E) shows consistent improvements across these venues: THE-Tree augmentation improves paper evaluation accuracy for multiple models, confirming that the evolutionary knowledge encoded in THE-Trees transfers across different conference contexts without venue-specific tuning.

4.2 THE-Tree Quality Validation (RQ1)

4.2.1 Effectiveness of Technology Trajectory Reconstruction

Hallucination mitigation in TVCV. We compare a simplified “Think-Verbalize-Cite” (TVC; w/o RA-NLI) against full TVCV. Table 1 shows RA-NLI removes phantom citations (21.46% $\rightarrow$ 0.00%) and reduces residual relation hallucinations (46.73% $\rightarrow$ 9.32%; phantom-free).

Lineage fidelity beyond semantic similarity. To demonstrate that RA-NLI captures genuine intellectual lineage rather than mere semantic similarity, we measure alignment with observable citation-lineage signals. With RA-NLI filtering, the real citation-link rate (fraction of edges where the child paper actually cites the parent) rises from 37.69% to 61.14% (+23.45 pp), and the average temporal gap between connected nodes drops from 4.97 to 3.08 years (-1.89 years). While this does not prove

Table 3: **Comparison of Graph Completion Performance Between THE-Tree and Traditional Citation Graphs.** Detailed definitions of the evaluation metrics (Hit@k, MR, MedianRank, MRR) can be found in Appendix F.6.

Model	Hit@1 ($\uparrow$)	Hit@2 ($\uparrow$)	Hit@3 ($\uparrow$)	Hit@4 ($\uparrow$)	Hit@5 ($\uparrow$)	MR ($\downarrow$)	MRR ($\uparrow$)	MedianRank ($\downarrow$)
Graph built with Traditional Citations Relations								
Qwen2.5-72b	0.5854	0.7665	0.8482	0.8911	0.9113	1.8056	0.7627	1.2713
Qwen2.5-32b	0.2989	0.4833	0.6307	0.7335	0.7956	3.1271	0.5279	2.6232
Qwen2.5-7b	0.2527	0.5211	0.6786	0.7459	0.8097	3.0776	0.5109	2.7537
Gemma-7b	0.1631	0.4254	0.6273	0.7261	0.7863	3.4965	0.4356	3.4039
Gemma-2b	0.1567	0.4793	0.6425	0.7335	0.7939	3.4177	0.4433	3.3276
Graph built with Our Reasoning Relations								
Qwen2.5-72b	0.7214	0.8585	0.9058	0.9212	0.9313	1.4266	0.8593	1.0795
Qwen2.5-32b	0.3707	0.6190	0.7389	0.8109	0.8556	2.5772	0.6049	2.0296
Qwen2.5-7b	0.2586	0.5253	0.7069	0.7687	0.8149	2.9950	0.5214	2.6010
Gemma-7b	0.2405	0.4935	0.6557	0.7367	0.7985	3.2692	0.4896	3.1601
Gemma-2b	0.1564	0.3320	0.4475	0.5227	0.5772	4.7788	0.3738	4.5887

Table 4: **Macro-F1 and Balanced Accuracy for Paper Evaluation (NeurIPS 2024).** +TT denotes THE-Tree augmentation.

Model	Acc-F1	Rej-F1	Macro-F1	Bal.Acc.
DR-14b	43.31	30.25	36.78	55.45
+TT	43.54	35.57	39.56	56.16
DR-7b	41.58	31.81	36.69	53.29
+TT	54.28	74.06	64.17	69.91
Claude-3.5	42.67	0.76	21.71	50.07
+TT	40.39	49.93	45.16	54.02
GPT-4o	39.69	4.38	22.03	50.94
+TT	41.56	5.18	23.37	51.16

Table 5: **Quantitative Comparison of THE-Tree Reconstruction Quality (MCTS-generated) against Expert-Refined Ground Truth.** The methodology for constructing this expert-refined ground truth dataset is detailed in Appendix F.5. Metrics assess the ability to reconstruct entities (papers) and their evolutionary relations across the 88 topics.

Method	Entity				Relation		
	Recall	Precision	F1	Avg_Time_Diff	Recall	Precision	F1
Expert	1.00	1.00	1.00	2.93	1.00	1.00	1.00
THE-Tree	0.84	0.67	0.75	3.08	0.78	0.64	0.70

formal causality, it provides external evidence that RA-NLI substantially improves lineage fidelity.

Fact verification accuracy. We evaluate RA-NLI on a 71k fact-verification set covering 27k papers and compare against strong baselines (with/without factual supplementation). Table 2 shows our approach achieves a fact-missing rate of only 4.75% versus 40–69% for standalone LLMs, and the best overall verification accuracy (95.60%).

Reconstruction quality against expert ground truth. To validate trajectory reconstruction, we compare MCTS-generated trees against expert-refined ground truth (construction methodology in Appendix F.5). Four domain experts independently refined each tree (2–4 hours per tree), primarily by pruning noisy relations. Table 5 shows entity F1

of 0.75 and relation F1 of 0.70, with comparable temporal gaps (3.08 vs. 2.93 years). We acknowledge potential anchoring bias in this protocol and report statistics of expert edits in Appendix F.5 (Table 13). Inter-annotator agreement on a random sample of 5 trees yields Fleiss’ $\kappa = 0.74$ for node keep/remove decisions and $\kappa = 0.62$ for edge validity, with Krippendorff’s $\alpha = 0.57$, indicating substantial agreement.

4.2.2 Structural and Semantic Properties: Graph Completion (RQ2)

Table 6: **Ablation: Effect of removing path-level reward (R_{gen}) on Graph Completion.**

Model	Citation	Full	w/o R_{gen}	Δ
Hit@1				
Qwen2.5-72b	0.5854	0.7214	0.6425	+0.0789
Qwen2.5-32b	0.2989	0.3707	0.3293	+0.0414
Qwen2.5-7b	0.2527	0.2586	0.2362	+0.0224
Gemma-7b	0.1631	0.2405	0.1894	+0.0511
MRR				
Qwen2.5-72b	0.7627	0.8593	0.8061	+0.0532
Qwen2.5-32b	0.5279	0.6049	0.5438	+0.0611
Qwen2.5-7b	0.5109	0.5214	0.5132	+0.0082
Gemma-7b	0.4356	0.4896	0.4497	+0.0399

We evaluate THE-Tree’s structural quality via graph completion (Lee et al., 2023): entities from a given year are masked, and models predict them using prior history. Table 3 shows THE-Tree robustly outperforms citation graphs, particularly for larger models (Qwen2.5-72b: Hit@1 improves from 0.59 to 0.72, MRR from 0.76 to 0.86). The gains confirm that THE-Tree encodes richer latent knowledge structure than citation-only topology. We note that Gemma-2b is the only model where THE-Tree does not improve over citation graphs; this is consistent with the finding that very small models lack

Table 7: **Comparison of THE-Tree and Citation Graph on Future Path Prediction.** Metrics reported for Entity and Relation predictions (see metric definitions in Appendix F.6).

Model	Graph	Entity					Relation				
		Hit@1	Hit@3	Hit@5	MR ($\downarrow$)	Med.Rank ($\downarrow$)	Hit@1	Hit@3	Hit@5	MR ($\downarrow$)	Med.Rank ($\downarrow$)
Qwen2.5-72b	Citation	0.1831	0.5421	0.7423	3.6210	3.5149	0.1262	0.2940	0.4002	4.5318	4.3731
	THE-Tree	0.2813	0.6073	0.7746	3.3961	3.2288	0.1693	0.3143	0.4154	4.2892	4.0547
Qwen2.5-32b	Citation	0.1078	0.4079	0.6189	4.5388	4.3515	0.1452	0.2779	0.3862	4.6621	4.6188
	THE-Tree	0.1500	0.4906	0.6894	4.0494	3.9064	0.1522	0.2987	0.4073	4.4674	4.3128
Qwen2.5-7b	Citation	0.1524	0.4941	0.7013	4.0054	0.3903	0.1319	0.2789	0.3859	4.6937	4.6305
	THE-Tree	0.2125	0.5675	0.7277	3.7194	3.6262	0.1437	0.2980	0.4033	4.5071	4.3812
Gemma-7b	Citation	0.1652	0.6037	0.7550	3.4445	3.4005	0.1022	0.2610	0.3790	4.8544	4.8632
	THE-Tree	0.2431	0.6250	0.7798	3.3706	3.3276	0.1243	0.2674	0.3849	4.7426	4.7069
Gemma-2b	Citation	0.1671	0.5947	0.7541	3.4951	3.4680	0.1062	0.2619	0.3747	4.8881	4.9286
	THE-Tree	0.1735	0.5894	0.7445	3.4381	3.4187	0.0896	0.2641	0.3773	4.8912	4.9113

sufficient synthesis capacity to leverage the richer relational structure.

Ablation: contribution of path-level reward.

To isolate the contribution of the DPO-style path-level generation reward (R_{gen}), we evaluate graph completion on trees constructed without it. Table 6 shows removing R_{gen} leads to partial loss of gains for most models, indicating that path-level coherence scoring contributes meaningfully to structural quality. The effect is stronger for larger models, while the smallest model (Gemma-2b) remains unstable regardless of configuration.

4.3 Future Path and Trajectory Prediction (RQ2)

We evaluate THE-Tree’s ability to capture evolutionary dynamics via future path prediction: given a tree constructed up to year Y , predict future entities and their semantic relations. Citation-only graphs show limited predictive power ($\text{Hit@1} \approx 10\text{--}18\%$), underscoring the inadequacy of citation topology alone. THE-Tree achieves substantial gains (Table 7): Hit@1 improves by 6–10 pp for entities across models, with consistent improvements in Hit@3 , Hit@5 , MR, and Median Rank, confirming that the semantic signals encoded by THE-Tree enable sharper anticipation of research trajectories.

Cross-cutting findings. Across all three research questions, the results reveal a consistent pattern: THE-Tree’s historically grounded evolutionary structure provides a substantially richer knowledge substrate than citation-only topology. The architectural separation between constrained generation (TVCV) and independent verification (RA-NLI) is the enabling mechanism: it allows LLM reasoning to propose candidate relationships while maintaining rigorous factual grounding (RQ1: 95.60% verification accuracy, 0% phantom citations). This verified structure then propagates

benefits to both structural reasoning (RQ2: up to +13.6 pp Hit@1 in graph completion, +9.8 pp in future path prediction) and downstream LLM augmentation (RQ3: up to +27.5 macro-F1 in paper assessment). The gains scale with model capacity—larger models better exploit the richer relational signals—while the smallest models (Gemma-2b) show limited or inconsistent improvement, suggesting a minimum synthesis capacity threshold for leveraging structured evolutionary knowledge.

5 Conclusion

We introduced **THE-Tree** (Technology History Evolution Tree), a computational framework for constructing structured, verifiable, and historically grounded representations of scientific evolution from literature. By moving beyond traditional bibliometrics and ungrounded LLM assessments, THE-Tree provides a principled way to trace and evaluate the evolution of ideas with greater reliability.

THE-Tree integrates self-guided temporal Monte Carlo tree search (SGT-MCTS) with a Think–Verbalize–Cite–Verify (TVCV) reasoning process and retrieval-augmented natural language inference (RA-NLI) to build evolution trees that are both interpretable and verifiable. Our main contributions are: (i) an automated construction pipeline integrating SGT-MCTS with TVCV generation and RA-NLI validation; (ii) a dataset of 88 THE-Trees spanning diverse AI topics with 71k fact verifications; and (iii) empirical evidence that THE-Tree better captures meaningful historical trajectories and supports downstream reasoning tasks, including graph completion, future path prediction, and LLM-based paper assessment.

By offering a structured and historically grounded map of knowledge development, THE-

Tree advances transparent and evidence-based evaluation of emerging scientific ideas. We defer detailed future-work directions to Appendix C.

6 Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (NSFC) under grant No. 62403389, the Provincial Natural Science Foundation of Zhejiang under grant No. QKWL25F0301, and the Zhejiang Key Laboratory of Low-Carbon Intelligent Synthetic Biology (2024ZY01025).

Limitations

Despite the promising results, our work has several limitations.

- **Coverage and bias:** THE-Tree is currently built primarily from English, peer-reviewed literature, which may under-represent non-English research, patents, and grey literature.
- **Validation and subjectivity:** Validating scientific evolution lacks a universal gold standard; some node/edge decisions can still be subjective, especially for fine-grained conceptual distinctions. Our expert-refined ground truth starts from MCTS-generated trees, so experts verify a proposed structure rather than building from scratch, which is subject to anchoring bias (though inter-annotator agreement of Fleiss' $\kappa = 0.62$ – 0.74 suggests moderate-to-substantial consistency).
- **Computation:** Constructing and validating large trees is computationally expensive due to extensive retrieval and LLM usage; Appendix Table 11 provides representative per-tree costs under different model sizes.
- **Conservative bias and uncertainty:** Any history-grounded evaluation framework may under-reward paradigm-shifting work with weak apparent predecessors, and our current system does not explicitly provide calibrated uncertainty estimates for link validity.
- **Scope of downstream evaluation:** Although source collection spans multiple databases, downstream paper-evaluation experiments are validated exclusively on CS/AI venues (NeurIPS, ICLR, ICML, CVPR). Generalization to other scientific domains remains to be demonstrated.

References

- Anthropic. 2024. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>.
- Sören Auer and Markus Stocker. 2018. Towards a knowledge graph for science. In *Proceedings of the 11th International Conference on Semantic Systems*, pages 1–8.
- 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 preprint arXiv:2404.07738*.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
- Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023a. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.
- Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023b. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.
- Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 64(1):45–80.
- Andres M Bran, Sam Schilter, Philippe Delfosse, Tobias Gensch, Anica Ivanova, Apeksha Thakkar, Yoo-Jin Jung, Philippe Schwaller, Alain C Vaucher, Kevin Maik Chang, and 1 others. 2023. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.
- Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.
- Michel Callon, John Law, and Arie Rip. 1983. Mapping the dynamics of science and technology: Sociology of science in the real world. In *Sociology of science in the real world*. Macmillan London.
- Chaomei Chen. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American society for information science and technology*, 57(3):359–377.
- Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. To cite, or not to cite? detecting citation contexts in text. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 598–603. Springer.

- Kasra Ferdowsi, Ruanqianqian Huang, Michael B James, Nadia Polikarpova, and Sorin Lerner. 2024. Validating ai-generated code with live programming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–8.
- Eugene Garfield. 1979. *Citation indexing: Its theory and application in science, technology, and humanities*. Wiley.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Qianyu Hao, Jingyang Fan, Fengli Xu, Jian Yuan, and Yong Li. 2024. Hlm-cite: Hybrid language model workflow for text-based scientific citation prediction. *arXiv preprint arXiv:2410.09112*.
- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302*.
- Marcus Hutter. 2000. A theory of universal artificial intelligence based on algorithmic complexity. *arXiv preprint cs/0004001*.
- Marcus Hutter. 2001. Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. In *Machine Learning: ECML 2001*, pages 226–238. Springer.
- Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S Weld, and Peter Clark. 2025. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation. *arXiv preprint arXiv:2503.22708*.
- Pat Langley. 1987. *Scientific discovery: Computational explorations of the creative processes*. MIT press.
- Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. Temporal knowledge graph forecasting without knowledge using in-context learning. *arXiv preprint arXiv:2305.10613*.
- Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.
- Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.
- 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*.
- Henry Small. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269.
- Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998.
- Qwen Team. 2024. **Qwen2.5: A party of foundation models**.
- Nees Jan Van Eck and Ludo Waltman. 2010. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538.
- Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, and Yuxiao Dong. 2020. Microsoft academic graph: When experts are not enough. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2846–2854.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. **CycleResearcher: Improving Automated Research via Automated Review**. In *International Conference on Learning Representations (ICLR)*.
- Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2023. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

B Using THE-Tree for Scientific Idea Verification

Once a THE-Tree, with its richly annotated nodes and semantically meaningful edges, is constructed for a specific scientific domain, it serves as a powerful instrument for the verification and contextualization of new scientific ideas or papers. We propose a straightforward yet effective methodology to utilize THE-Tree for this purpose (Figure 4), enabling the retrospection of relevant historical evolutionary paths for a given input scientific paper, P_{in} , defined by its title T_{in} and abstract A_{in} . This approach provides critical context by situating new research within established knowledge frameworks, thereby aiding in the assessment of its novelty and potential contribution.

The core steps of this verification and retrospection process are as follows:

1. **Initialization:** This methodology assumes access to:
 - A collection of pre-computed THE-Trees $\{G_k\}$, where each $G_k = (V_k, E_k)$ corresponds to a specific scientific topic $Topic_k$. Nodes $v \in V_k$ represent scientific papers with attributes such as publication year Y_v , title T_v , abstract A_v , and an importance score S_v , as defined in Section 3.1. Edges $e \in E_k$ signify directed evolutionary relationships, capturing inferential and developmental dependencies.
 - A Large Language Model (LLM) for semantic tasks such as similarity assessment and topic matching.
2. **Topic Space Identification:** For the input paper P_{in} , a set of N_T most relevant topics, $\mathcal{T}_{P_{in}} = \{Topic_1, \dots, Topic_{N_T}\}$, is identified from the available THE-Tree topics. This involves prompting an LLM with T_{in} , A_{in} , and

the list of all THE-Tree topic names to determine the most appropriate THE-Tree(s) for situating P_{in} .

3. **Candidate Path Origination in Relevant THE-Trees:** For each identified topic $Topic_k \in \mathcal{T}_{P_{in}}$ and its corresponding THE-Tree G_k :
 - (a) Identify a candidate set of M_k recent papers $\{v_{k,j}\} \subset V_k$ from G_k based on publication year $Y_{v_{k,j}}$ (e.g., papers published within a certain recent time window or up to the year of P_{in} if known).
 - (b) Calculate a semantic similarity score $Sim(P_{in}, v_{k,j}) \in [0, 1]$ between the input paper P_{in} (using its title and abstract) and each candidate $v_{k,j}$ (using its stored title and abstract) using an LLM or other embedding-based techniques.
 - (c) Select the top N_S papers from these candidates, denoted $\mathcal{S}_k = \{v_{k,1}^*, \dots, v_{k,N_S}^*\}$, that exhibit a similarity score $Sim(P_{in}, v_{k,l}^*) \geq \theta_{sim}$, where θ_{sim} is a predefined minimum similarity threshold. These papers $v_{k,l}^*$ serve as terminal nodes (or "entry points") for backward path tracing within the THE-Tree, representing the most closely related established works to P_{in} .

4. **Historical Path Retrospection:** For each selected terminal paper $v_{term}^* \in \mathcal{S}_k$ in graph G_k :
 - A historical path $Path_{k,l} = (v_1, v_2, \dots, v_n = v_{term}^*)$ is constructed by iteratively selecting predecessors. This path traces the lineage of v_{term}^* backward through the THE-Tree.
 - Starting from $v_{curr} = v_{term}^*$, the predecessor v_{pred}^* is chosen from the set of all predecessors $Pred(v_{curr})$ (i.e., nodes from which an edge points to v_{curr}) according to the lexicographical optimization that prioritizes strong evolutionary links and historical relevance:

$$v_{pred}^* = \arg \min_{\substack{v \in Pred(v_{curr}) \\ Y_v \leq Y_{v_{curr}}}} ((Y_{v_{curr}} - Y_v), -S_v)$$

This heuristic prioritizes the predecessor with the smallest non-negative time

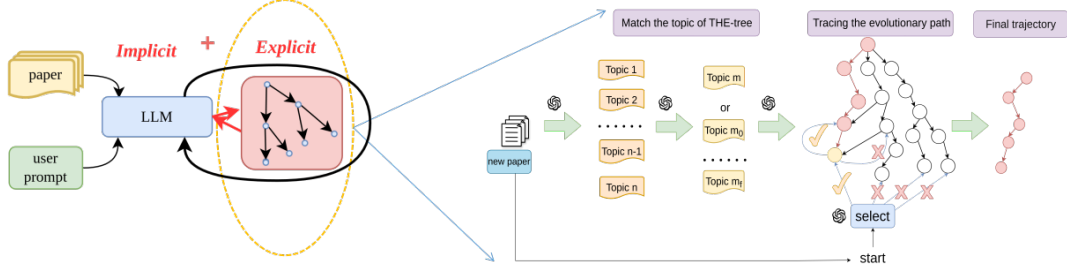


Figure 4: A simple way to explicitly use THE-tree

difference ($Y_{v_{curr}} - Y_v$), reflecting direct chronological succession, and secondarily, the one with the highest importance score S_v , indicating significant contributions. The use of node attributes Y_v (year) and S_v (importance) from the THE-Tree structure is central here.

- Traversal continues backward until no valid predecessor (respecting chronological order and edge semantics) is found, a predefined maximum path length is exceeded, or a root node of the THE-Tree (a node with no predecessors in the context of this path’s history) is reached. The resulting path is then presented in chronological order from the earliest paper to v_{term}^* .

5. **Path Aggregation and Presentation:** All generated paths $\{Path_{k,l}\}$ are collected. They can be ranked based on various criteria, such as the semantic similarity score $Sim(P_{in}, v_{term}^*)$ of their terminal nodes to the input paper P_{in} , the cumulative importance of nodes in the path, or the overall coherence of the path. The top N_P unique paths are selected and formatted for presentation. This presentation details the sequence of papers, their key attributes (title, year, summary of contribution derived from abstract/node info), and the nature of the connecting evolutionary relationships.

This methodology provides a simple yet powerful way to leverage the structured knowledge within THE-Trees to verify a new scientific idea by exploring its historical context and connections to established research. The retrieved paths can highlight the foundations upon which P_{in} builds, identify potentially overlooked prior art, or help assess its incremental novelty versus a more radical departure from existing trajectories. Users can adapt

this general approach based on specific analytical needs, such as modifying the LLM prompts, similarity thresholds, path selection heuristics, or the depth of retrospection.

C Additional Discussion: Future Work

We outline several directions that can extend THE-Tree beyond the current scope:

- **Broader coverage and representations:** incorporate non-English literature, patents/technical reports, finer temporal granularity, and more diverse research outcomes (e.g., negative results).
- **Stronger validation and reasoning:** integrate richer inferential reasoning and uncertainty estimation for evolutionary links, and improve robustness on underexplored or abandoned research paths.
- **Usability and applications:** build interactive visualization tools and integrate THE-Tree into scientific workflows for verification, exploration, and trajectory analysis.

D Temporal Cutoff and Leakage Prevention Protocol

We employ a simple time-threshold protocol to reduce training-data contamination and inadvertent future information leakage in our LLM-based experiments.

- **Model release-date cutoff.** We only evaluate models released no later than mid-2024, so they are unlikely to have been trained on late-2024–2025 papers.
- **Held-out targets.** Our paper-evaluation and trend/trajectory prediction experiments focus on late-2024 and 2025 papers as prediction/analysis targets.

- **No external browsing.** Models are not given web access or external tools during evaluation; they operate only on the provided inputs and our fixed resources.
- **Temporal filtering of evidence.** Whenever grounding/evidence retrieval is used, the candidate evidence pool is restricted to publications not later than the specified cutoff year Y of the experiment.

This protocol is complementary to our use of a fixed, pre-constructed THE-Tree in downstream evaluation (i.e., no online updates during testing), which further reduces the chance of leaking post-cutoff information.

E Additional Results: THE-Tree Verifier for Scientific Evaluation

To reduce potential data contamination, we run THE-Tree-augmented scientific evaluation on NeurIPS 2024 submissions. Table 8 provides the full per-model breakdown. Here we include supplementary protocol details and additional qualitative analysis.

E.1 Data Collection and Processing

We collected data from multiple sources, including Web of Science, Scopus, arXiv, IEEE Xplore, and PubMed. The data processing pipeline consisted of the following steps:

1. **Metadata extraction:** We extracted titles, abstracts, authors, venues, and publication dates using custom parsers for each data source.
2. **Citation network construction:** We built a directed graph where nodes represent papers and edges represent citation relationships.
3. **Text preprocessing:** We applied standard NLP preprocessing techniques, including tokenisation, stopword removal, and lemmatisation.
4. **Entity recognition:** We used a combination of dictionary-based and machine learning approaches to identify technical terms and concepts.
5. **Temporal alignment:** We aligned papers along a timeline, accounting for publication delays and citation patterns.

To evaluate the accuracy of our metadata extraction component, we measured its precision on two commonly used formatting styles: 98.29% for IEEE format and 97.30% for APA format. These high precision scores demonstrate the robustness of our system in handling different citation conventions.

E.2 Detailed Dataset Construction from Surveys for THE-Trees

The processing pipeline applied to these curated surveys involves the following key steps designed to extract structured information:

1. **Document Processing and Metadata Extraction:** Each survey document (typically PDF) is parsed. Metadata for the survey itself is extracted. Crucially, its reference list is parsed to create initial 'paper nodes' for each cited work, populated with available metadata (title, authors, year, etc.). The survey text is segmented into paragraphs and sentences.
2. **Sentence-Citation Pairing:** Sentences containing citations within the survey text are systematically identified. For sentences in the paper that originally contain citations, we do not directly use them as citation sentences; instead, we perform a series of post-processing steps on their factual content, including the removal of invalid facts. For each processed sentence, its textual content is extracted and explicitly linked to the corresponding cited paper node(s) (identified via citation markers like ((Färber et al., 2018))). This step generates numerous '<citing sentence, cited paper>' pairs. These pairs form the basis for two critical downstream tasks: (a) creating a dataset for subsequent causal relationship validation (e.g., using NLI to assess if the citing sentence is supported by the cited paper's content); and (b) creating a dataset for evaluating the model's ability to accurately extract citation context.
3. **Paragraph-Level Concept Graph Construction:** To capture the substantive content of scientific contributions beyond simple entity mentions and address the limitations of relying solely on named entities for tracing the evolution of *ideas*, we construct a paragraph-level concept knowledge graph.

Table 8: **Impact of THE-Tree Augmentation on LLM-based NeurIPS Paper Evaluation.** The table compares the performance of various LLMs in predicting paper acceptance/rejection and status (Poster, Spotlight, Oral) for NeurIPS 2023 and NeurIPS 2024. For the NeurIPS 2024 dataset, results are presented with and without THE-Tree augmentation to demonstrate its effect on evaluation accuracy. Performance is evaluated using standard metrics (see Appendix F.6).

Model	NeurIPS 2023							NeurIPS 2024						
	Accuracy of accept and reject			Accuracy of Status				Accuracy of accept and reject			Accuracy of Status			
	Acc%	Rej%	Total%	Poster%	Spot%	Oral%	Total	Acc%	Rej%	Total%	Poster%	Spot%	Oral%	Total%
Qwen2.5-72b-instruct	99.48	0.52	26.34	3.59	100	0	3.38	99.63	0	25.70	1.24	100	0	2.42
Deep-Reviewer-14b	93.70	16.06	36.32	73.65	31.82	0	29.59	92.61	18.28	37.45	74.03	22.77	6.94	31.03
Deep-Reviewer-7b	83.94	18.75	35.76	57.49	18.18	0	27.54	86.79	19.78	37.07	60.42	19.05	0	28.9
GPT4-O	99.48	2.59	27.88	30.54	90.91	0	10.95	99.63	2.24	26.02	29.46	86.96	0	10.26
Claude-3.5-Sonnet	99.75	0.52	26.42	6.27	55	44.16	3.38	99.75	0.38	27.28	10.90	77.96	18.57	4.59
Deepseek-R1	99.57	3.42	28.51	54.49	52.66	0	16.27	100.0	2.00	26.03	54.13	53.31	0	15.00
With THE-tree Augmentation														
Qwen2.5-72b-instruct_tree	-	-	-	-	-	-	-	99.84	0.37	26.03	1.76	97.38	0	2.76
Deep-Reviewer-14b_tree	-	-	-	-	-	-	-	89.93	22.39	39.82	63.2	34.55	34.72	32.08
Deep-Reviewer-7b_tree	-	-	-	-	-	-	-	76.12	63.69	66.90	57	24.99	2.33	60.84
GPT4-O_tree	-	-	-	-	-	-	-	99.66	2.66	27.69	34.45	72	2.6	11.41
Claude-3.5-Sonnet_tree	-	-	-	-	-	-	-	71.46	36.57	45.57	28.72	38.13	36.57	34.81
Deepseek-R1_tree	-	-	-	-	-	-	-	99.57	3.42	28.23	56.49	56.66	2.6	16.68

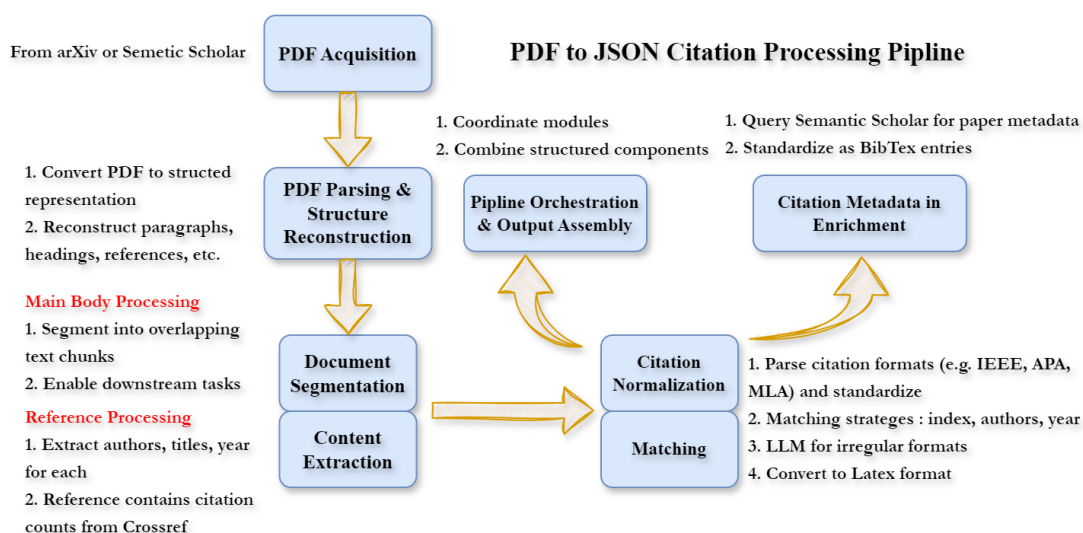


Figure 5: Document processing and metadata extraction pipeline applied to curated surveys for THE-Tree dataset construction.

Traditional entity-based KGs can face challenges in citation alignment for our task:

- (a) **Granularity Mismatch:** A citation often supports a broader claim or methodology within a sentence/paragraph, not just a specific entity mentioned nearby.
- (b) **Synonymy/Paraphrasing:** The core scientific concept might be described using different terminology or entities across papers (or even within the same paper), potentially breaking entity-based links.
- (c) **Implicit Concepts and Alignment Omission:** Important ideas (e.g., a novel argument, a methodological variant) may be difficult to represent as standard named entities, or relevant entities might

not appear immediately adjacent to the citation marker, leading to missed connections by proximity-based alignment methods, as you noted.

To overcome these issues, we employ NLP techniques (e.g., keyphrase extraction, relation extraction, or summarisation) to extract core scientific/technical concepts from each paragraph. A ‘concept’ here refers to a core idea, method, or finding, often represented as a phrase or concise statement, rather than just an isolated named entity. A ‘concept node’ is created for each extracted concept. Bidirectional indexing is established between these concept nodes and their source paragraph text

chunks, facilitating traceability to the original text for verification. Crucially, if a concept is derived from a sentence that contains a citation, this ‘concept node’ is linked to the ‘paper node(s)’ referenced by that sentence. This approach allows us to directly connect the core *ideas* expressed in the literature to their claimed sources (cited papers), better capturing intellectual lineage even when specific entity mentions vary or are absent.

This pipeline yields a structured dataset comprising surveys, cited papers (with metadata), concepts, sentences, paragraphs, and explicit links representing relationships such as ‘<citing sentence, cited paper>’, ‘<concept, paragraph>’, and ‘<concept, cited paper>’. This rich dataset covers content related to up to 27k papers, with as many as 71k entries used for factual verification evaluation. It forms the foundation for the subsequent construction of detailed technology evolution histories using the THE-Tree framework (integrating SGT-MCTS and TVCV).

E.3 Model Architecture Details

Construction Model Implementation: Our technology tree construction model uses a multi-layer graph neural network with the following architecture:

- Node embedding layer: 768-dimensional vectors initialised from SciBERT (Beltagy et al., 2019)
- Graph attention layers: 3 layers with 8 attention heads each
- Temporal encoding: Sinusoidal position encoding based on publication year
- Edge type encoding: Learned embeddings for different relationship types

Training was performed using Adam optimizer with a learning rate of 0.0001 and batch size of 32. We used early stopping with a patience of 10 epochs based on validation loss.

NLI Model Implementation: To assess the textual entailment within scientific citations in RA-NLI, we employ a fine-tuned DeBERTa model specifically adapted for inference tasks in the scientific domain. The model categorizes the relationship between a citation claim (hypothesis) and its associated source content (premise) into one of the following three classes:

- **Entailment:** The premise logically supports or implies the claim.
- **Contradiction:** The premise directly contradicts the claim.
- **Neutral:** There is insufficient evidence to determine a clear inferential relationship.

In cases where the NLI model outputs a neutral label—indicating ambiguity or lack of strong inferential evidence—we introduce a secondary validation step using a large language model (Qwen2.5-72B-Instruct). This LLM-based evaluator conducts a more nuanced assessment by considering broader contextual and semantic factors, further determining whether the cited relationship constitutes a direct citation, a paraphrase, or no meaningful connection.

Our validation process can be formalized as follows:

For a given pair of technology nodes (v_i, v_j) where v_i is claimed to influence v_j , we compute:

$$R_{\text{attr}}(v_i \rightarrow v_j) = \alpha \cdot \text{NLI}(s_i, s_j) + (1 - \alpha) \cdot \text{LLM}_{\text{eval}}(s_i, s_j, C), \quad (5)$$

where s_i and s_j are the textual descriptions of nodes v_i and v_j respectively, C represents the retrieved context from the literature, and α is a weighting parameter determined empirically (set to 0.7 in our experiments).

The NLI function returns a normalised score in the range $[0, 1]$ based on the entailment probability, while LLM_{eval} returns a similar score based on the LLM’s assessment of the citation validity.

E.4 The data structure of THE-tree

Figure 6 shows the data structure of THE-tree

E.5 Cross-Conference Validation Results

To demonstrate the generalizability of THE-Tree augmentation, Table 9 presents cross-conference validation results across ICLR2025, ICML2025, and CVPR2025, showing consistent improvements in paper evaluation accuracy across different venues and models.

F Supplementary Details for Experiments

F.1 Dataset Construction and Statistics

Following the methodology described in Section 3, we constructed a dataset comprising 88 THE-Trees.

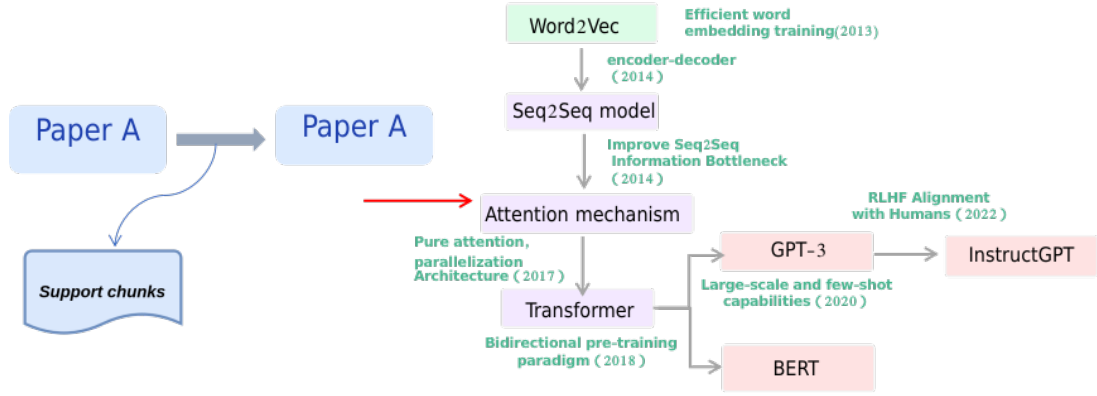


Figure 6: The data structure of THE-tree

Table 9: Cross-Conference Validation of THE-Tree Augmented Paper Evaluation

Conference	Model	Setting	Accept	Reject	Poster	Spotlight	Oral
ICLR2025	GPT-4o	THE-tree	99.84%	12.76%	28.23%	75.94%	16.90%
	GPT-4o	without_history	99.97%	4.40%	37.13%	64.74%	5.63%
	Qwen2.5-72B	THE-tree	99.89%	1.08%	0.77%	96.70%	1.41%
	Qwen2.5-72B	without_history	99.92%	0.20%	0.58%	99.21%	0.47%
	Claude-3.5-Sonnet	THE-tree	100.00%	20.68%	5.41%	77.36%	33.80%
	Claude-3.5-Sonnet	without_history	100.00%	10.08%	3.41%	64.10%	15.36%
ICML2025	GPT-4o	THE-tree	99.23%	17.23%	31.05%	62.86%	5.56%
	GPT-4o	without_history	99.55%	15.95%	48.46%	48.00%	0.00%
	Qwen2.5-72B	THE-tree	99.55%	1.37%	12.09%	96.81%	2.78%
	Qwen2.5-72B	without_history	99.82%	0.54%	0.57%	98.22%	0.93%
	Claude-3.5-Sonnet	THE-tree	100.00%	34.16%	69.39%	73.33%	5.56%
	Claude-3.5-Sonnet	without_history	99.97%	24.67%	5.66%	93.78%	16.67%
CVPR2025	GPT-4o	THE-tree	98.97%	-	29.71%	72.09%	20.00%
	GPT-4o	without_history	99.97%	-	55.33%	51.85%	0.00%
	Qwen2.5-72B	THE-tree	100.00%	-	1.86%	98.72%	0.00%
	Qwen2.5-72B	without_history	100.00%	-	0.38%	99.74%	0.00%
	Claude-3.5-Sonnet	THE-tree	100.00%	-	9.10%	94.44%	40.00%
	Claude-3.5-Sonnet	without_history	100.00%	-	84.22%	18.10%	13.33%

These trees cover distinct technological topics primarily within core AI and its applications across diverse scientific domains (e.g., Computer Science, Biomedicine, Materials Science).

Table 10 provides enhanced statistical summary of our dataset construction, including detailed metrics on processed topics, paper nodes, and edges across different aggregation levels. The 88 THE-Trees represent validated evolutionary trajectories within this broader knowledge space.

F.2 Model Efficiency Analysis

We evaluate the computational efficiency of different model sizes for THE-Tree construction, as shown in Table 11. The analysis demonstrates the trade-off between model performance and computational cost across various Qwen2.5 model variants.

Table 10: Enhanced Statistical Summary with Standardization

Metric	Total	Avg/Topic	Avg/Tree	Avg/Human
Processed topics	88	-	-	-
Paper nodes	35,392	402.18	103.14	46.49
Paper edges	140,616	1597.91	255.57	204.67

Note: All metrics calculated over 1950–2023 temporal scope.

Avg/Topic values computed using harmonic mean.

Table 11: Model Efficiency Comparison for THE-Tree Construction, Larger models yield higher reconstruction F1 but require more compute (all runs used $8 \times A100$)

Model Size	Entity F1	Relation F1	Time (hours)	GPU-Hours
Qwen2.5-7B	0.65	0.58	2.42	19.4
Qwen2.5-14B	0.69	0.63	3.15	25.2
Qwen2.5-32B	0.72	0.66	3.89	31.1
Qwen2.5-72B	0.75	0.70	4.73	37.8

F.3 Expert Annotation Protocol

Our expert annotation protocol involves domain experts (with PhD-level expertise in the respective fields) who: (1) Review automatically generated evolutionary pathways for semantic coherence, (2) Validate or correct temporal relationships between nodes, (3) Remove spurious connections that lack genuine evolutionary influence, and (4) Add missing critical evolutionary steps. This expert validation process typically requires 2-4 hours per tree.

Table 12 presents detailed statistics on expert editing across different scientific domains, showing the extent of modifications required and the resulting coherence improvements.

F.4 RA-NLI Technical Details

Our RA-NLI mechanism operates through a three-stage pipeline: (1) *Retrieval*: Using sentence-transformers with cosine similarity ≥ 0.75 , we extract top-k=5 relevant passages from the target paper; (2) *Inference*: A fine-tuned RoBERTa-large model predicts entailment, contradiction, or neutrality with confidence scores; (3) *Aggregation*: Multiple evidence pieces are combined using weighted voting, where $\tau_{i,j} = \frac{1}{K} \sum_{k=1}^K w_k \cdot p_k^{entail}$, with weights w_k based on passage relevance scores. The final confidence threshold $\theta_{min} = 0.7$ was determined through extensive validation on 10k expert-annotated relationship pairs.

F.5 Expert-Refined Ground Truth Construction Methodology

Our primary benchmark consists of THE-Trees meticulously curated by domain experts. This process involved two main stages: 1) *Initial Tree Generation by MCTS*: Our self-guided temporal Monte Carlo Tree Search (MCTS) algorithm, incorporating the Think-Verbalize-Cite-Verify (TVCV) methodology with Retrieval-Augmented Natural Language Inference (RA-NLI), first generated initial THE-Trees for each topic. This ensures that MCTS primarily proposes semantically and temporally plausible connections. 2) *Expert Refinement*

and Augmentation: Domain experts then reviewed these MCTS-generated trees, performing comprehensive modifications. This included validating, correcting, or removing paths and nodes; augmenting trees with crucial missing links, milestone papers, or overlooked developmental trajectories; and ensuring overall semantic coherence, evolutionary validity, and accurate representation of the field’s historical evolution. The resulting expert-curated THE-Trees form the ground truth dataset used for validation as described in the main experimental sections. In practice, each THE-Tree was refined by a team of typically four domain experts; for interdisciplinary topics, we expanded to six experts (e.g., three from each field). Experts first conducted independent reviews and then held a consensus meeting to resolve disagreements.

Table 13 shows the expert refinement edits.

F.6 Evaluation Metrics Definitions

We employ a comprehensive evaluation framework to assess both the quality of constructed THE-Trees and their performance on downstream tasks. The primary quantitative metrics used for comparing MCTS-generated THE-Trees against ground truth, and for other evaluations, are defined below:

- **Node Recall**: The proportion of nodes from the expert-refined ground truth trees successfully identified by the MCTS process. $\text{Node Recall} = \frac{|\text{MCTS-identified Nodes} \cap \text{Ground Truth Nodes}|}{|\text{Total Nodes in Ground Truth}|}$.

- **Node Precision**: The proportion of nodes in MCTS-generated trees that are present in the expert-refined ground truth.

$$\text{Node Precision} = \frac{|\text{MCTS-identified Nodes} \cap \text{Ground Truth Nodes}|}{|\text{Total Nodes in MCTS Tree}|}$$

- **Edge Recall**: The proportion of evolutionary connections from the expert-refined ground truth trees successfully identified by the MCTS process.

$$\text{Edge Recall} = \frac{|\text{MCTS-identified Edges} \cap \text{Ground Truth Edges}|}{|\text{Total Edges in Ground Truth}|}$$

- **Edge Precision**: The proportion of evolutionary connections in MCTS-generated trees that are present in the expert-refined ground truth. $\text{Edge Precision} = \frac{|\text{MCTS-identified Edges} \cap \text{Ground Truth Edges}|}{|\text{Total Edges in MCTS Tree}|}$.

- **F1 Score**: The harmonic mean of precision and recall, calculated separately for nodes and edges. $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

Table 12: Expert Editing Statistics and Cross-Domain Analysis

Domain	Trees	Avg Edit Time (hrs)	Nodes Added	Edges Modified	Coherence Gain
Computer Science	32	2.4	12.3%	18.7%	+16.2%
Biomedical	28	3.1	15.8%	22.4%	+14.9%
Materials Science	18	2.8	11.7%	16.3%	+15.8%
Physics	10	3.4	14.2%	20.1%	+13.7%
Overall	88	2.9	13.5%	19.4%	+15.3%

Table 13: **Expert refinement edits (avg per THE-Tree)**. These statistics summarize how experts prune and augment the initial MCTS-generated drafts to form the final ground truth.

Edit type	Nodes	Edges
Added by experts	6.24	37.45
Removed by experts	62.89	88.35

- **Average Temporal Interval:** Given that our MCTS ensures chronological validity (parent node year $\leq$ child node year), this metric calculates the mean time difference in publication years between directly connected parent (v_p) and child (v_c) nodes: $\text{AvgInterval} = \frac{1}{|E|} \sum_{(v_p, v_c) \in E} (\text{Year}(v_c) - \text{Year}(v_p))$. It characterizes the typical evolutionary pace captured.
- **Metrics for Future Node Prediction and Graph Completion (e.g., Hits@k, MR, MRR):** These standard link prediction metrics are used as described in the main text when evaluating future node prediction (Section 4.3) and graph completion (Section 4.2.2). Their standard definitions apply.
- **Overall Accuracy Metrics in NeurIPS Paper Evaluation (Total% in Table 8):** The experiment table includes two "Total%" overall accuracy metrics, calculated as explained below:

- **"Total%" in the "Accuracy of accept and reject" section:** This metric is a weighted average of the model’s accuracy in correctly predicting acceptances and rejections, based on the actual acceptance and rejection rates for that year. The formula is:

$$\text{Total}\%_{\text{acc/rej}} = P(\text{Accept}) \cdot \text{Acc}\% + P(\text{Reject}) \cdot \text{Rej}\%.$$

Here, $P(\text{Accept})$ and $P(\text{Reject})$ are the true proportions of accepted and rejected papers in the evaluation set.

- **"Total%" in the "Accuracy of Status" section:** This metric comprehensively evaluates

the model’s overall accuracy in predicting all specific paper statuses (Poster, Spotlight, Oral, Reject). The formula is:

$$\text{Total}\%_{\text{status}} = P(\text{Accept}) \sum_{s \in \{\text{Pos}, \text{Spo}, \text{Oral}\}} P(s \mid \text{Accept}) \cdot \text{Acc}_s + P(\text{Reject}) \cdot \text{Rej}\%.$$

where $P(s \mid \text{Accept})$ is the proportion of accepted papers with status s , and Acc_s corresponds to the per-status accuracy columns ("Poster%", "Spot%", "Oral%"). This metric is a weighted average of the accuracies for all final statuses based on their actual proportions.

F.7 Case Study: THE-Tree Augmentation Impact on DeepReviewer-14b for NeurIPS Paper Evaluation

Further detailed analysis of the DeepReviewer-14b model from the NeurIPS 2024 paper evaluation task (see Table 8 in the main text) provides a clear illustration of THE-Tree’s impact. Figure 7 specifically highlights the performance differences when DeepReviewer-14b is augmented with THE-Tree versus its standalone performance. The augmen-

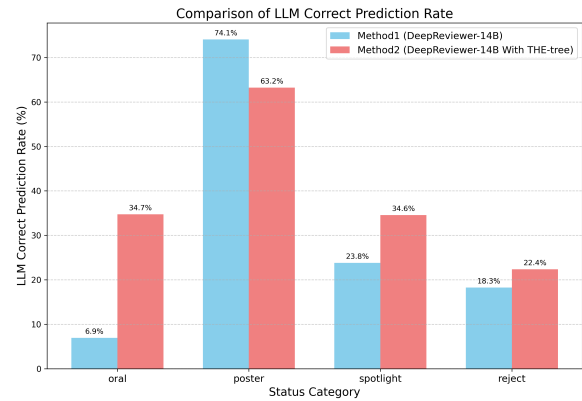


Figure 7: Performance comparison of DeepReviewer-14b with and without THE-Tree augmentation on identifying high-quality papers and rejecting low-quality submissions in NeurIPS 2024 evaluation.

tation demonstrably enhances the model’s ability to discern paper quality. Notably, THE-Tree improves DeepReviewer-14b’s capacity to correctly identify high-quality submissions, such as those ultimately designated as Oral or Spotlight presentations. Concurrently, it significantly boosts the model’s effectiveness in rejecting papers that do not meet the acceptance criteria, thus reducing the likelihood of erroneously endorsing lower-quality work.

Figure 8 delves into the prediction distribution for papers whose ground truth status was Oral or Spotlight. When THE-Tree augmentation is applied, the predictions made by DeepReviewer-14b for these high-impact papers shift more decisively towards categories indicating higher quality (e.g., predicting them as Oral or Spotlight with greater confidence or frequency). This contrasts with the standalone model, which may exhibit a more dispersed or less accurate prediction pattern for these important papers. This case study underscores THE-Tree’s role as a powerful enhancement for LLM-based scientific evaluation. By providing structured historical context and verifiable evolutionary pathways, THE-Tree equips models like DeepReviewer-14b with a more robust foundation for assessing scientific contributions, leading to more accurate identification of impactful research and more reliable filtering of less meritorious submissions.

F.8 Detailed Case Study: LLM Evaluation of "Neural Pfaffians" with and without THE-Tree Augmentation

To further illustrate the impact of THE-Tree augmentation on the LLM’s evaluation capabilities, we present a detailed comparison of an LLM’s assessment of the paper titled "Neural Pfaffians: Solving Many Many-Electron Schrödinger Equations." The ground truth status for this paper was **Oral**.

F.8.1 Case 1: LLM Evaluation *with* THE-Tree Augmentation

- **Paper Title:** Neural Pfaffians: Solving Many Many-Electron Schrödinger Equations
- **Original Output (Ground Truth):**
 - Status: Oral
- **LLM Evaluation (with THE-Tree):**
 - Decision: accept
 - Rating: **oral**

- Reason: *"The paper introduces a novel approach to solving the many-electron Schrödinger equation using a learnable Pfaffian-based neural wave function. This is a significant departure from traditional methods that rely on Slater determinants and non-learnable orbital selection, addressing a key limitation in the field of computational quantum chemistry. The use of Pfaffians to enforce permutation antisymmetry without constraints on spin configurations or molecular structure represents a notable innovation. The empirical results, demonstrating chemical accuracy on a range of systems and outperforming the CCSD(T) CBS reference on the TinyMol dataset, suggest a substantial improvement over existing generalized neural wave function methods. The paper’s claim of reducing energy errors by up to an order of magnitude compared to previous methods, if substantiated, indicates a high impact potential for the field. However, the abstract lacks detailed information on the specific neural network architecture and training procedure, which would be crucial for a full assessment of the technical soundness. The claim of a single neural Pfaffian calculating both ground and ionization energies warrants further scrutiny regarding the methodology used for ionization energy calculations."*

F.8.2 Case 2: LLM Evaluation *without* THE-Tree Augmentation

- **Paper Title:** Neural Pfaffians: Solving Many Many-Electron Schrödinger Equations
- **Original Output (Ground Truth):**
 - Status: Oral
- **LLM Evaluation (without THE-Tree):**
 - Decision: accept
 - Rating: **poster**
 - Reason: *"The paper introduces a novel approach to learning generalized wave functions for many-electron systems by utilizing Pfaffians instead of Slater determinants, addressing the challenge of enforcing permutation antisymmetry in a fully learnable manner. This represents a significant advancement in the field of neural wave functions,*

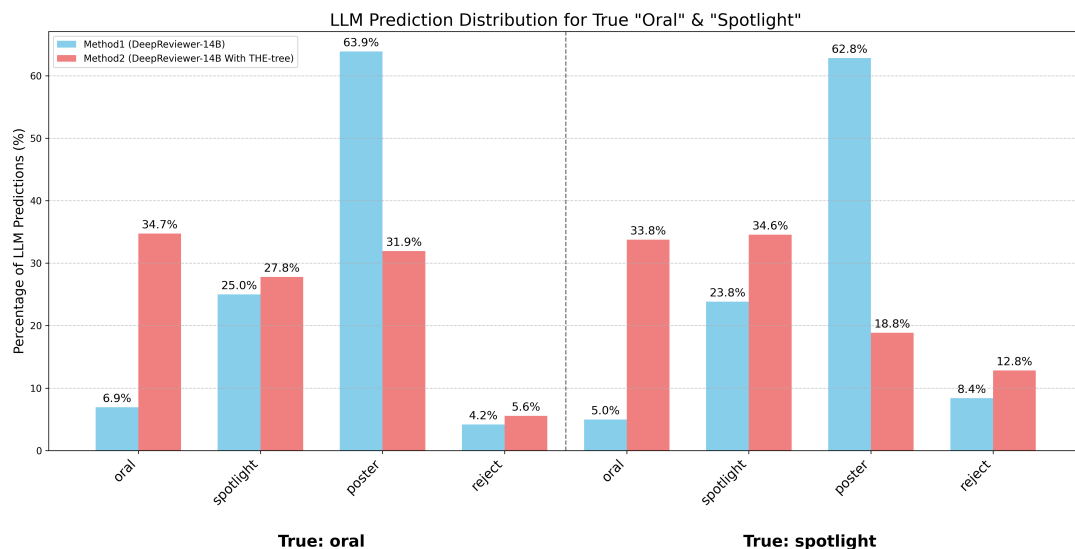


Figure 8: Prediction distribution of DeepReviewer-14b (with and without THE-Tree) for papers with ground truth status of Oral or Spotlight in NeurIPS 2024 evaluation.

potentially reducing the computational cost associated with solving each system independently. The empirical results demonstrate that the proposed 'neural Pfaffian' achieves chemical accuracy across various systems and outperforms the CCSD(T) CBS reference energies on the TinyMol dataset. The claim of reducing energy errors compared to previous generalized neural wave functions by up to an order of magnitude, if substantiated, indicates a substantial improvement in accuracy. The method's ability to generalize across molecules without constraints on electronic spin configurations or molecular structure further enhances its potential impact. However, the paper's clarity could be improved by providing more details on the specific architecture of the overparametrized neural network and the implementation of the Pfaffian. Additionally, a more detailed comparison with existing methods, including a discussion of their limitations and how the proposed approach overcomes them, would strengthen the paper's contribution. The significance of this work lies in its potential to revolutionize the way we approach quantum chemistry calculations, making them more efficient and accurate. The technical soundness appears promising, but a more thorough description of the methodology and a more detailed

analysis of the results are needed to fully assess its robustness. Overall, the paper presents a compelling idea with strong empirical support, suggesting that it would be a valuable addition to the conference."

F.8.3 Comparative Analysis

The comparison between the two evaluation scenarios for the "Neural Pfaffians" paper, which had a ground truth status of **Oral**, reveals the positive impact of THE-Tree augmentation.

- Rating Accuracy:** The LLM augmented with THE-Tree correctly predicted the paper's status as "**oral**", aligning with the ground truth. In contrast, the standalone LLM, while still deciding to "accept" the paper, rated it as "**poster**", underestimating its eventual impact tier.
- Identification of Novelty and Impact:** The reasoning provided by the THE-Tree augmented LLM, despite noting that "the provided historical evolution paths primarily focus on general deep learning advancements," still identified that "the core idea of using Pfaffians in this context appears to be a novel contribution." This suggests that even with somewhat general historical context, THE-Tree helped the LLM anchor the paper's specific contribution within an evolutionary landscape, allowing it to better discern the novelty. It also highlights the "high impact potential" based on the claimed error reduction.

The standalone LLM also recognized the significance ("potential to revolutionize"), but its reasoning carried more reservations regarding the need for "a more thorough description of the methodology and a more detailed analysis of the results ... to fully assess its robustness." While these are valid points for any review, the overall tone and the resulting "poster" rating suggest a slightly diminished confidence in the paper's immediate high-tier impact compared to the augmented evaluation.

- **Confidence in Assessment:** THE-Tree augmentation appears to provide the LLM with additional contextual evidence that strengthens its assessment. By situating the paper's contributions within a (even broadly) relevant historical trajectory, the LLM can form a more nuanced understanding of its specific advancements and their potential significance. This leads to an evaluation ("oral") that more closely mirrors the expert consensus reflected in the ground truth. The non-augmented LLM, lacking this explicit historical grounding, relies more heavily on the paper's intrinsic claims and presentation, which, while positive, led to a more conservative (and in this case, less accurate) impact assessment.

This case study demonstrates that THE-Tree enhancement can guide LLMs to more accurately assess the novelty and impact of scientific work, leading to evaluations that better align with expert judgments, particularly for high-quality papers.

F.9 Detailed Case Study: LLM Evaluation of "Exploitation of a Latent Mechanism in Graph Contrastive Learning: Representation Scattering" with and without THE-Tree Augmentation

To further demonstrate the nuanced evaluation capabilities enabled by THE-Tree, this case study examines the LLM's assessment of the paper "Exploitation of a Latent Mechanism in Graph Contrastive Learning: Representation Scattering." The ground truth for this paper was also **Oral**.

F.9.1 Case 1: LLM Evaluation *with* THE-Tree Augmentation

- **Paper Title:** Exploitation of a Latent Mechanism in Graph Contrastive Learning: Representation Scattering
- **Original Output (Ground Truth):**

– Status: Oral

• LLM Evaluation (with THE-Tree):

- Decision: accept
- Rating: **oral**
- Reason: *"The paper introduces a novel perspective on Graph Contrastive Learning (GCL) by identifying a common mechanism, termed 'representation scattering,' that underlies the effectiveness of various GCL frameworks. This unifying concept is a significant contribution, offering a deeper understanding of why different GCL approaches work. The proposed Scattering Graph Representation Learning (SGRL) framework, which incorporates a center-away strategy to enhance representation diversity and a topology-based constraint to prevent excessive scattering, appears to be a technically sound and innovative approach. The paper's motivation is clear, and the proposed method is well-articulated. The experimental results, though not detailed in the provided text, are claimed to demonstrate the efficacy and superiority of SGRL over existing GCL methods, which, if substantiated, would indicate a significant impact on the field. The paper effectively bridges the gap between different GCL paradigms by highlighting their shared mechanism, and the proposed SGRL framework builds upon this understanding to improve performance. The technical evolution paths provided highlight the progression of GNNs and contrastive learning separately, and this paper attempts to synthesize these advancements in the context of GCL, which is a valuable contribution."*

F.9.2 Case 2: LLM Evaluation *without* THE-Tree Augmentation

- **Paper Title:** Exploitation of a Latent Mechanism in Graph Contrastive Learning: Representation Scattering
- **Original Output (Ground Truth):**
 - Status: Oral
- **LLM Evaluation (without THE-Tree):**
 - Decision: accept
 - Rating: **poster**

- Reason: *"The paper introduces a novel perspective on Graph Contrastive Learning (GCL) by identifying a common mechanism, termed 'representation scattering,' across different GCL frameworks. This unifying concept is a significant contribution, offering a deeper understanding of why these methods work. The proposed Scattering Graph Representation Learning (SGRL) framework, which explicitly leverages this mechanism with a center-away strategy and a topology-based constraint, demonstrates a clear technical innovation. The paper's claim of SGRL's efficacy and superiority is supported by evaluations on benchmark datasets, although the specific details of these evaluations are not provided in the abstract. The clarity of the abstract is good, outlining the problem, the proposed solution, and the key results. The potential impact of this work is substantial, as it not only provides a theoretical insight into GCL but also proposes a practical framework that could advance the field of graph representation learning."*

F.9.3 Comparative Analysis

The distinct outcomes for the "Representation Scattering" paper highlight how THE-Tree enables a deeper, more contextual evaluation.

- **Rating Accuracy:** The THE-Tree augmented LLM correctly identified the paper as "oral," matching the ground truth. The standalone LLM, while positive, assigned a "poster" rating, failing to capture its full impact.
- **Contextual Understanding of Contribution:** Both evaluations acknowledge the novel "representation scattering" concept. However, the reasoning from the augmented LLM is more insightful. It explicitly references the historical context provided by THE-Tree, stating, "The technical evolution paths provided highlight the progression of GNNs and contrastive learning separately, and this paper attempts to synthesize these advancements..." This demonstrates that the LLM used the evolutionary context to understand how the paper unified two distinct research threads, a key factor in its high impact. The standalone LLM's analysis, lacking this context, remains more superficial, focusing only on the paper's

self-described contributions without appreciating its role in synthesizing prior work.

- **Assessment Confidence and Nuance:** The augmented evaluation confidently points to the paper's value as a "synthesis of advancements." The standalone LLM, while acknowledging the "substantial" potential impact, gives a more standard review focused on the abstract's contents. The ability to place the work within its historical and technical lineage allowed the augmented LLM to make a more decisive and accurate judgment, mirroring the expert consensus of an "oral" presentation.

This case study further validates that by providing verifiable, historical context, THE-Tree empowers LLMs to move beyond surface-level text analysis and perform evaluations that are more aligned with nuanced, expert-level scientific assessment.

G The Use of Large Language Models

We used large language models (GPT-4o and Claude Sonnet 4.0) only as general-purpose writing assistants to refine the clarity and fluency of the manuscript, such as improving grammar, style, and consistency of academic tone. All scientific ideas, methodological designs (including the THE-Tree framework, algorithms, and experiments), data collection, and analysis were conceived and executed entirely by the authors. Any LLM-generated or suggested text was critically reviewed and edited to ensure accuracy and faithfulness to our original contributions.

H Code and Data Availability

Our code and reconstructed data will be released upon paper acceptance.