# COMPOSE: Composing Future Theorems from Citations and Formal Structure

**David Busbib** **Michael Werman**

Hebrew University of Jerusalem

## Abstract

Future mathematical claims depend both on the development of ideas across papers and on formal dependency structures that constrain what can validly follow. Existing approaches typically model only one of these signals. We introduce grounded future mathematical generation, the task of predicting plausible future claims from an anchor paper by jointly modeling scientific context and formal theorem structure. To support this setting, we construct a dataset of 108K paired scientific-formal graph examples from arXiv and Mathlib using papers published up to 2023, together with a temporally grounded benchmark of 47K future papers from 2024–2025. We propose COMPOSE, a dual-graph framework that builds and fuses scientific and formal graphs to condition a language model for theorem-like prediction. We evaluate using retrieval against real future papers, together with LLM judgments and text-quality metrics. COMPOSE outperforms strong baselines, producing more grounded and mathematically rich predictions and achieving the strongest retrieval performance in our main evaluation setting. These results show that future mathematical generation benefits from combining scientific context with formal structure.

## 1 Introduction

Large language models have shown growing promise in anticipating future scientific contributions from existing knowledge across many domains [Si et al., 2024, Lu et al., 2024a, Wang et al., 2026]. However, in the mathematical domain, a plausible future result must both extend the research trajectory of prior work [Lakatos, 1976, Davies et al., 2021] but also respect the logical dependencies that determine what can validly follow from existing definitions, lemmas, and theorems [Yang et al., 2023, Uskuplu et al., 2025, Cabral et al., 2025]. Thus, the task is not only to produce fluent mathematical text, but to generate a result that is both scientifically meaningful and logically grounded.

This challenge is reflected in two complementary sources of mathematical knowledge: scientific papers, which capture how ideas develop across the literature and indicate which directions are worth pursuing [Azher et al., 2025, Baek et al., 2025, He-Yueya et al., 2026], and formal theorem libraries such as Lean's Mathlib, which encode mathematical results together with explicit dependency structure. Combining the two is non-trivial because the same content appears differently in each source, and because the two sources are organized by fundamentally different relational structures [Gao et al., 2024a, Zhang et al., 2025, Wu et al., 2022]. Yet existing approaches tackle only one side: models trained on scientific text alone can identify plausible research directions but may generate claims that are not grounded in formal mathematical dependencies [Wang et al., 2024b, Radensky et al., 2026, Baek et al., 2025, He-Yueya et al., 2026]; theorem-proving systems [Wang et al., 2025a, First et al., 2023, Yang et al., 2023] reason over formal structure but lack the surrounding scientific context needed to anticipate which directions are most promising; and structure-based methods typically model only one of these sources, using either citation structure [An et al., 2021] or theorem

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

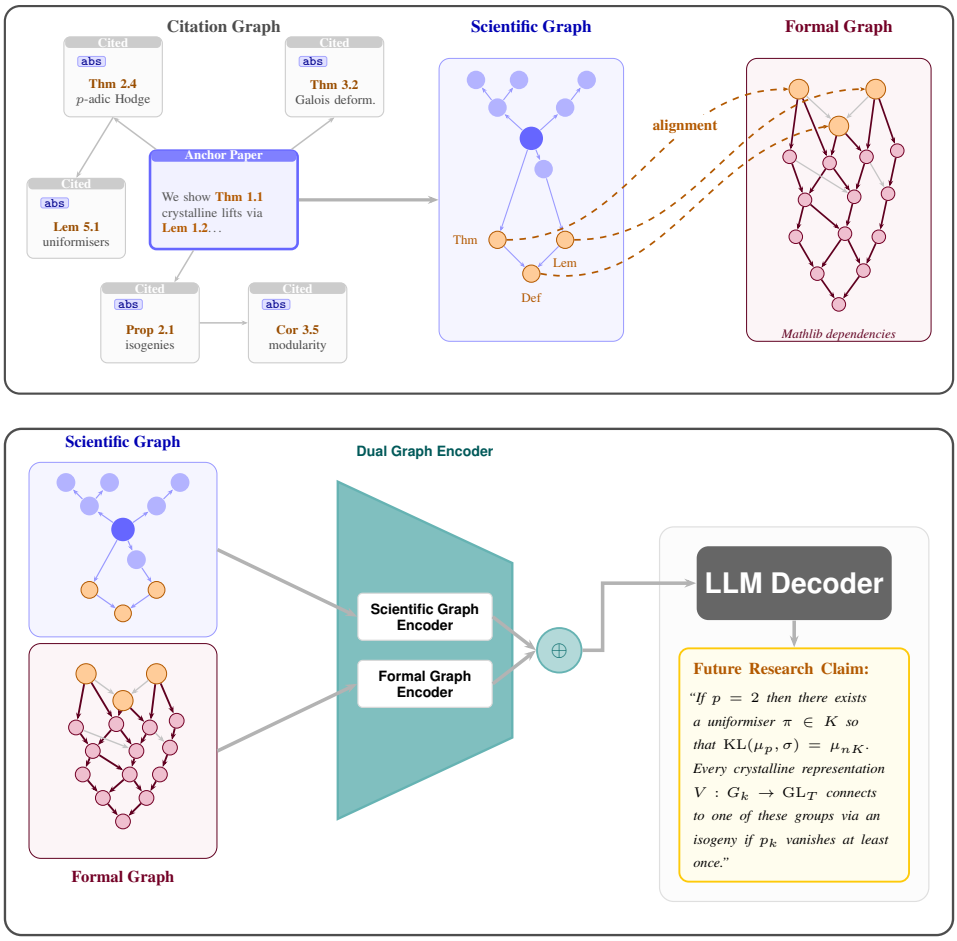



Figure 1: **Top: Data construction.** Given an anchor paper and its citation graph (*left*), we build its scientific graph (*center*) where **blue nodes** represent paper abstracts and **orange nodes** represent extracted theorems. Each informal theorem is matched to its corresponding Mathlib theorem (dashed arrows), and the matched theorems are then used in order the construct the paper's formal graph of Lean dependencies (*right*). **Bottom: Model architecture.** The scientific graph and formal graph are encoded by dedicated GNNs, fused via cross-attention ($\oplus$), and used to condition a math-specialized LLM decoder that generates a *future research claim*.

dependencies [Wang et al., 2025b, Zhu et al., 2026]. As a result, existing work still lacks a method that combines both sources for grounded future mathematical generation.

In this work, we propose *COMPOSE*, a system for generating plausible future mathematical claims from an anchor paper. This is carried out by jointly modeling scientific context and formal mathematical structure. To represent these two sources, we construct two complementary graphs: *(i)* a scientific graph built from the anchor paper, selected cited papers, and extracted theorems, and *(ii)* a formal graph built by aligning those theorems to corresponding formal theorems and their dependency structure. The two graphs are encoded separately and fused to condition an LLM decoder. Figure 1 provides an overview of the full pipeline. To learn this representation effectively, training proceeds in two stages: we first optimize the graph encoders and fusion module with alignment objectives, then fine-tune the decoder for graph-conditioned generation.

To evaluate this setting, we construct a benchmark of 108K paired samples from arXiv mathematics using training papers collected from the pre-2023 corpus, together with a test set of 47K papers from late 2024 and 2025.

Empirically, COMPOSE achieves the strongest performance in our main evaluation setting and performs best overall under Human and LLM evaluation, with ablations confirming that both graph sources contribute.

We make the following contributions:

- We study grounded future mathematical generation, where the goal is to predict plausible future mathematical claims using both scientific citation context and formal theorem dependencies.
- We construct a dataset of 108K paired scientific-formal graph samples from arXiv and Mathlib, using an informal-to-informal alignment strategy via FrenzyMath, and establish a benchmark of 47K future papers from 2024-2025 for evaluation.
- We propose COMPOSE, a system for grounded future mathematical generation that combines scientific citation context with formal theorem structure to generate plausible future mathematical claims.

## 2 Related Work

**Future Scientific Generation.** Prior work studies future scientific generation in several ways. One line of work generates future research directions from prior literature using retrieval or iterative retrieval-and-reasoning over related papers [Azher et al., 2025, Baek et al., 2025, Radensky et al., 2026]. Another line of work uses citation structure or temporal supervision for the task [Wang et al., 2026, 2024b]. GIANTS [He-Yueya et al., 2026] uses citation-based scientific context to guide generation, although it targets future scientific summaries rather than theorem-like claims. Some work also explores structure-aware scientific generation using external graphs or chain-of-ideas frameworks [Gao et al., 2025, Xiong et al., 2024]. Across these settings, existing systems rely on textual, citation-based, or structured scientific context, without combining the formal theorem structure that constrains what can validly follow from prior results.

**Formal Mathematical Reasoning.** Formal proof libraries such as Lean's Mathlib [mathlib Community, 2020], Isabelle's Archive of Formal Proofs [Paulson, 1994], and Coq-based libraries [Barras et al., 1999, Rudnicki, 1992] encode large collections of verified theorems together with their formal dependency structure [Li et al., 2024]. This dependency structure is not just a record of what has been proved; it also helps determine what can be proved next. Retrieval-augmented and premise-selection systems use this structure to identify relevant premises for proof search, as in ReProver [Yang et al., 2023] and related work [Alemi et al., 2017, Wang et al., 2017]. Formal theorem provers also use this library context to guide proof generation, including systems such as Baldur [First et al., 2023], DeepSeek-Prover [Xin et al., 2024], and related systems [Wang et al., 2025a, Dong and Ma, 2025]. Beyond proof search, formal libraries have also been used to conjecture what should be proved, Lemmanaid [Alhessi et al., 2026] proposes new lemmas by drawing analogies across existing theories, while related work studies conjecture generation, verification, and blueprint construction from library context [Onda et al., 2025, Kasaura et al., 2025, Liu et al., 2026, Zhu et al., 2026]. Together, these systems show that formal dependency structure can inform not only how to prove but also what to prove. However, they operate entirely within the formal world, requiring formal input and producing formal output, without modeling any informal mathematical context.

**Scientific Informal-Formal Alignment.** In our setting, the alignment problem is to match an informal theorem from a research paper to its corresponding theorem in Mathlib. A natural approach is to translate informal mathematics written in natural language into formal syntax. Early work used large language models to map informal theorems into formal representations [Wu et al., 2022, Azerbayev et al., 2023, Lu et al., 2024b]. Later work improved Autoformalization through advanced modeling and generation strategies, including Kimina [Wang et al., 2025a], Draft, Sketch, and Prove [Jiang et al., 2023], and ProofBridge [Jana et al., 2026]. Despite these advances, coverage on real paper-level mathematics remains limited because of the large vocabulary gap between informal exposition and formal syntax [Zhang et al., 2025]. As an alternative, FrenzyMath [Gao et al., 2024b] takes a different route by providing human-written informal descriptions for over 136K Mathlib theorems, enabling informal-to-informal retrieval as a pragmatic alignment strategy that avoids formalization altogether. Recent work also suggests that linking informal theorem to formal libraries benefits from stronger retrieval and intermediate structure [Lu et al., 2024b, Jiang et al., 2023, Jana

et al., 2026]. Unlike these works, we do not treat informal-formal alignment as the primary objective. Instead, we use these alignments to ground future mathematical claims.

# 3 Method

Given an anchor paper, the goal is to generate a plausible future mathematical claim using its scientific and matched formal context. Our approach has three components. First, we construct a dataset that links scientific papers with formal theorem graphs 3.1. Second, we introduce COMPOSE, a dual-encoder architecture that jointly models scientific and formal mathematical structure 3.2. Third, we present our two-stage training 3.3.

**Inference overview.** At inference time, given an anchor paper, the model constructs two graphs: a scientific citation graph $G_s$, built from the paper and its contextually relevant cited papers, and a formal theorem dependency graph $G_f$, built from Mathlib theorems aligned to extracted mathematical theorems. Each graph is independently encoded by a dedicated GNN encoder. The two representations are then fused via cross-attention. The fused graph embeddings are passed as conditioning to our language model decoder, which generates a free-text theorem proposal.

## 3.1 Data Curation and Construction

Training our dual-graph model requires paired examples linking scientific context with formal theorem-dependency structure. We construct this dataset by linking mathematical papers collected via S2ORC [Lo et al., 2020] with formal theorems from Mathlib [mathlib Community, 2020].

On the scientific side, we begin with an anchor paper and construct outgoing citation edges for up to two hops, retaining at most 5 cited papers at the first hop and at most 3 cited papers per node for the second hop. Cited papers are selected by citation-context relevance, favoring citations that are discussed in important parts of the paper or used directly in its proofs or main results, rather than by including the full citation list; full selection details are in Appendix A.1. The graph contains *abstract* nodes for all included papers, together with *theorem* nodes extracted from papers in the first two citation hops. Each *theorem* node is connected either to its corresponding *abstract* node or to a parent *theorem* node. We use typed edges to distinguish citation links, paper-to-theorem links, and theorem-to-theorem links, allowing the graph to capture both external and internal citation context. For each example, we also identify a target paper: a paper that cites the anchor and was published after it, excluded from the input graph. Its main mathematical claim serves as the generation target.

On the formal side, we align extracted theorems to Mathlib theorems using dense retrieval over the FrenzyMath corpus [Gao et al., 2024b], which provides informal descriptions for 140k Mathlib theorems (see Figure 2). For each extracted theorem, we keep the top retrieved theorem only if its similarity is above a threshold $\tau$. When no formal matches pass the alignment threshold, COMPOSE conditions generation only on the scientific graph. Each retained theorem serves as the root of the local formal dependency graph, which is expanded for a fixed number of hops using directed theorem-dependency edges extracted from Mathlib with LeanDojo [Yang et al., 2023]. Root nodes are marked with a distinct node type, while the remaining nodes represent dependency neighbors. See Appendix A.4 for alignment comparisons and details on $\tau$.

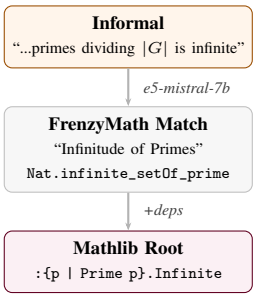


Figure 2: Informal-to-informal alignment.

Starting from a corpus of 500K mathematical papers published between 2000 and 2023, this process yields approximately 108K paired paper-theorem subgraphs. In addition, we define a temporally held-out benchmark of 47K mathematical papers from late 2024 and 2025, using their *abstract* and extracted main claim, as future target texts for evaluation. Full details of the alignment procedure, target construction are provided in Appendix A.

## 3.2 COMPOSE

We first introduce notation, then present the encoder, fusion, and decoder components.

Formally, each example is represented as

$$x = (G_s, G_f, \mathcal{P}), \qquad G_s = (V_s, E_s), \qquad G_f = (V_f, E_f),$$

where $x$ is the input context and $y = (y_1, \ldots, y_T)$ is the target output sequence.

$G_s$ contains two types of nodes: $V_s = \{v_j^{\text{abs}}\}_{j=1}^n \cup \{u_i^{\text{thm}}\}_{i=1}^m$

where $v_j^{\text{abs}}$ denotes an abstract node corresponding to a paper in the scientific graph, and $u_i^{\text{thm}}$ are extracted theorem nodes. The edge set $E_s$ contains directed citation edges between abstract nodes and structural edges connecting each theorem node either to its corresponding abstract node or to a parent theorem node.

The formal graph contains theorem nodes, $V_f = \{v_\ell^{\text{thm}}\}_{\ell=1}^k$, where each node corresponds to a Mathlib theorem, and edge set $E_f$ is given by dependency relations extracted from Mathlib.

Let $V_s^{\text{thm}} = \{u_i^{\text{thm}}\}_{i=1}^m$ denote the set of theorem nodes in $G_s$. The alignment set is defined as

$$\mathcal{P} \subseteq V_s^{\text{thm}} \times V_f,$$

where each pair $(u^{\text{thm}}, v^{\text{thm}}) \in \mathcal{P}$ links an informal theorem node in $G_s$ to its matched formal theorem node in $G_f$.

Each graph is processed by a dedicated GNN encoder, with both encoders sharing the same message-passing architecture [Gilmer et al., 2017]. Scientific nodes are initialized with E5 embeddings [Wang et al., 2024a], and formal nodes with DeepSeek-Math embeddings [Shao et al., 2024]. Both encoders use directed, edge-type-specific message passing, aggregating incoming and outgoing messages separately. Each node's structural vector is concatenated with its original text embedding, yielding representations $\mathbf{h}_i^s$ and $\mathbf{h}_j^f$. Full architectural details are provided in Appendix D.

Since $G_s$ and $G_f$ operate in different representation spaces, we project $\mathbf{h}_i^s$ and $\mathbf{h}_j^f$ into a shared latent space using the same learned MLP. Type embeddings are then added after projection to distinguish nodes from $G_s$ and $G_f$. We apply a single bidirectional cross-attention block between the two node sets, with separate Q/K/V projections for the scientific-to-formal and formal-to-scientific directions, and residual connections with layer normalization applied after each directional update. The resulting fused node representations are denoted by $\{\tilde{\mathbf{h}}_i\}_{i=1}^{|V_s|+|V_f|}$.

The fused node representations $\{\tilde{\mathbf{h}}_i\}$ are provided as graph inputs to a pretrained decoder model adapted with LoRA [Hu et al., 2021]. Cross-attention layers are inserted at regular intervals, covering approximately 20% of the layers. In these layers, the decoder hidden states serve as queries, while the fused graph node representations, projected to the decoder hidden dimension, serve as keys and values. This allows each generation step to attend over the full set of fused graph nodes, conditioning the output sequence $y$ on both the scientific context of $G_s$ and the formal structure of $G_f$.

### 3.3 Two-Stage Training

The model is trained in two stages. First, we train the graph encoders and fusion module to capture graph structure and align scientific and formal representations. We then add the decoder and fine-tune the full model for generation.

In the first stage, we train the scientific and formal graph encoders, together with the bridge MLP and fusion module, while keeping the pretrained text embeddings frozen. The resulting graph modules are then used to initialize the second stage. The first-stage objective is:

$$\mathcal{L}_{\text{stage}_1} = \lambda_{link}\mathcal{L}_{link} + \lambda_{align}\mathcal{L}_{align} + \lambda_{cross}\mathcal{L}_{cross} \tag{1}$$

To ensure both encoders capture graph structure rather than collapsing to uniform representations, we supervise them with a link prediction loss. For each graph $G \in \{G_s, G_f\}$, we define $a_{ij} = 1$ if $(v_i, v_j) \in E$ and $a_{ij} = 0$ for non-edges , and optimize

$$\mathcal{L}_{link} = -\mathbb{E}\left[a_{ij} \log \sigma(\mathbf{h}_i \cdot \mathbf{h}_j) + (1 - a_{ij}) \log\left(1 - \sigma(\mathbf{h}_i \cdot \mathbf{h}_j)\right)\right]. \tag{2}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ denote representations in the graph under consideration, and $\sigma$ is the sigmoid function.

To encourage the fused graph representation to align with the scientific and formal content of the target, we use a contrastive alignment loss [van den Oord et al., 2019]. Let

$$\tilde{\mathbf{h}}_G = \text{AttnPool}(\{\tilde{\mathbf{h}}_i\}_{i=1}^{|V_s|+|V_f|}) \tag{3}$$

denote the fused graph representation obtained by attention pooling over the fused node representations.

We construct a positive target representation $\mathbf{e}^+$ from the held-out target paper by combining the embeddings of its abstract and main mathematical claim, after projecting both into the shared latent space. For each example, we construct negative target representations $\mathcal{B}^-$ from papers that don't cite papers in the current scientific subgraph, together with their main mathematical claim.

This objective encourages the fused graph representation to be close to the true held-out target and far from unrelated target papers. We then optimize

$$\mathcal{L}_{align} = -\log \frac{\exp(\mathrm{sim}(\tilde{\mathbf{h}}_G, \mathbf{e}^+)/\tau)}{\exp(\mathrm{sim}(\tilde{\mathbf{h}}_G, \mathbf{e}^+)/\tau) + \sum_{\mathbf{e}^- \in \mathcal{B}^-} \exp(\mathrm{sim}(\tilde{\mathbf{h}}_G, \mathbf{e}^-)/\tau)} \tag{4}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity.

Finally, to bridge the informal and formal spaces, $\mathcal{L}_{cross}$ aligns each informal theorem node $u_i^{\mathrm{thm}} \in V_s$ with its matched theorem node $v_j^{\mathrm{thm}} \in V_f$, where $(u_i^{\mathrm{thm}}, v_j^{\mathrm{thm}}) \in \mathcal{P}$. For each positive pair, we sample a set of random negative theorem nodes $\mathcal{N}_i \subset V_f$. We then optimize

$$\mathcal{L}_{cross} = -\frac{1}{|\mathcal{P}|} \sum_{(u_i^{\mathrm{thm}}, v_j^{\mathrm{thm}}) \in \mathcal{P}} \log \frac{\exp(\hat{\mathbf{h}}_i^s \cdot \hat{\mathbf{h}}_j^f/\tau)}{\exp(\hat{\mathbf{h}}_i^s \cdot \hat{\mathbf{h}}_j^f/\tau) + \sum_{v_l^{\mathrm{thm}} \in \mathcal{N}_i} \exp(\hat{\mathbf{h}}_i^s \cdot \hat{\mathbf{h}}_l^f/\tau)}. \tag{5}$$

where $\tau$ is the temperature hyperparameter.

In the second stage, we add a pretrained math-specialized decoder adapted via LoRA [Hu et al., 2021] and fine-tune the full model for graph-conditioned generation using two objectives: an autoregressive generation loss and a graph margin loss that prevents the decoder from ignoring the graph context. The autoregressive loss is:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, G_s, G_f). \tag{6}$$

To prevent the decoder from ignoring the graph context and relying purely on language patterns, we additionally use a graph margin loss $\mathcal{L}_{margin}$ that requires the model to assign lower generation loss to the correct graph pair $(G_s, G_f)$ than to a mismatched pair $(G'_s, G'_f)$ sampled from a different citation graph within the same mathematical domain.

$$\mathcal{L}_{margin} = \max\big(0,\, \mathcal{L}_{CE}(G_s, G_f) - \mathcal{L}_{CE}(G'_s, G'_f) + \delta\big) \tag{7}$$

The second stage objective is

$$\mathcal{L}_{\mathrm{stage}_2} = \lambda_{ce}\, \mathcal{L}_{CE} + \lambda_{margin}\, \mathcal{L}_{margin} \tag{8}$$

Additional alignment and training details are provided in Appendix D.

# 4 Experiments

## 4.1 Experimental Setup

**Data and Temporal Split.** The 108K paired scientific-formal graph samples are split 80/10/10 into training, validation, and test sets. To evaluate temporal generalization, we construct a future-paper benchmark from a corpus of 47K mathematics papers published in late 2024-2025. We retain only test-set anchor papers that are cited by at least one paper in the 47K future-paper pool, yielding 2K test samples. For each anchor, future papers that cite it are treated as targets.

**Retrieval Evaluation.** Existing mathematical embedding models cluster future papers so tightly that the similarity gap between true target papers and randomly sampled unrelated papers is too small to reliably distinguish generation methods (see Appendix C). We therefore fine-tune a DeepSeek-Math embedding model on the 47K benchmark papers using a margin-ranking objective that pulls together

Table 1: Main results. Retrieval ranks the target future paper within a 47K-paper pool. Generation quality metrics are in Appendix B.

| | | *Future Retrieval (thm-ft pool)* | | | | | *NMCP* | | |
|---|---|---|---|---|---|---|---|---|---|
| **Decoder** | **Model** | **Tgt-Sim** | **Neg-Sim** | **Gap** | **H@10** | **H@100** | **Prec.** | **Match** | **Math** |
| DeepSeek-Math 7B | Paper-graph-only | 0.438 | 0.274 | 0.164 | 0.390 | 0.695 | 0.504 | 0.713 | 0.061 |
| | Bag-of-Papers | 0.399 | 0.271 | 0.128 | 0.230 | 0.544 | 0.497 | 0.719 | 0.064 |
| | Text-only (LoRA) | 0.463 | 0.287 | 0.176 | 0.369 | 0.738 | 0.302 | 0.356 | 0.076 |
| | Prompt-only | 0.471 | 0.297 | 0.174 | 0.109 | 0.345 | 0.165 | 0.187 | 0.051 |
| | **COMPOSE (ours)** | **0.525** | 0.285 | **0.240** | **0.508** | **0.808** | **0.560** | **0.730** | **0.083** |
| Mistral 7B | Paper-graph-only | 0.315 | 0.278 | 0.037 | 0.142 | 0.484 | 0.266 | 0.266 | 0.040 |
| | Bag-of-Papers | 0.418 | 0.279 | 0.139 | 0.154 | 0.506 | 0.323 | 0.211 | **0.070** |
| | Text-only (LoRA) | 0.443 | 0.291 | 0.152 | 0.091 | 0.359 | 0.425 | 0.450 | 0.068 |
| | Prompt-only | 0.447 | 0.296 | 0.151 | 0.090 | 0.197 | **0.575** | 0.653 | 0.044 |
| | **COMPOSE (ours)** | **0.451** | 0.290 | **0.161** | **0.348** | **0.697** | 0.554 | **0.671** | 0.054 |

papers with overlapping citation context and pushes apart unrelated papers. The resulting fixed model is then used as a shared retrieval evaluator for all compared methods.

**Metrics.** We report Tgt-Sim, the mean cosine similarity between the generated text and its corresponding ground-truth future papers; Neg-Sim, the mean cosine similarity between the generated text and 500 sampled papers from unrelated subfields of the future pool; Gap = Tgt-Sim $-$ Neg-Sim, which measures how much closer the generated text is to the correct future papers than to unrelated ones in the pool; and H@$k$, the fraction of examples for which a relevant future paper appears in the top-$k$ retrieved results. Exp-Sim measures similarity to the broader cluster of future papers that share citation context with the target, capturing whether the generated text lands in the correct research neighborhood rather than an unrelated area. We also report NMCP metrics for theorem-level grounding: Precision, the fraction of outputs aligned with the formal theorem target derived from the future paper; Match, the fraction of outputs aligned with the natural-language theorem target from the future paper; and Math, the ratio of mathematical tokens in the output. To complement embedding-based metrics, we also evaluate generation quality using an LLM judge (see Section B.6).

**Internal Variants.** We compare **COMPOSE** against several internal variants that remove or simplify parts of the proposed system: (1) Paper-graph-only, which uses only the scientific graph $G_s$; (2) Bag-of-Papers, which replaces the citation graph with a set of abstract and theorem nodes, removing graph structure and not using the two graph encoders; (3) Text-only (LoRA), which fine-tunes the decoder on our dataset using only a textual prompt, without any graph input; and (4) Prompt-only, which uses a richer prompt but does not use graph input or fine-tuning.

We evaluate all variants with both DeepSeek-Math 7B and Mistral 7B decoders to demonstrate that the gains from dual-graph conditioning are consistent across different backbone architectures.

**Baselines.** Since no prior work addresses exactly the same task, we compare against the closest systems for future scientific generation and future-paper prediction. These include (1) GIANTS [He-Yueya et al., 2026], which uses its own Qwen-based decoder; (2) DeepSeek-Math-7B-based systems, including FutureGen [Azher et al., 2025], GoAI [Gao et al., 2025], and KG-CoI [Xiong et al., 2024], which use the raw DeepSeek-Math-7B decoder; and (3) GPT-4.0-based variants of GoAI, CoI, and ResearchAgent, which we include as API baselines. We also include a simple retrieval baseline: Fixed NN, which retrieves by nearest-neighbor similarity over abstracts and main theorems. For all external baselines, we use the authors' released code and follow the prompts described in their papers, adapting only the output request toward a mathematical claim.

### 4.2 Main Results

Table 1 compares internal model variants on the full test set. **COMPOSE** achieves the best results across all retrieval metrics, with the largest Gap (0.240), showing that its outputs stay closer to the cluster of relevant future papers than to unrelated papers. Text-only and Prompt-only produce text broadly similar to future papers (Tgt-Sim 0.463 and 0.471), but their Neg-Sim is also high, yielding smaller Gaps (0.176 and 0.174) and weaker retrieval. This suggests that, without graph conditioning, these models capture the right topic but fail to isolate the specific future result. **COMPOSE** achieves the highest H@10 (0.508) and H@100 (0.808), reflecting that its generated text ranks the correct future paper higher across the 47K pool than any other variant. **COMPOSE** also leads on NMCP

Table 2: Future paper retrieval on the 2K test set (47K pool), comparing COMPOSE against external baselines.

| Decoder | Model | Tgt-Sim | Neg-Sim | Gap | Exp-Sim | H@10 | H@100 |
|---|---|---|---|---|---|---|---|
| Retrieval | Fixed NN (abstracts+ main theorem) | 0.393 | 0.285 | 0.108 | – | 0.068 | 0.392 |
| DeepSeek-Math 7B | GoAI | 0.510 | 0.309 | 0.202 | 0.381 | 0.376 | 0.680 |
| | Prompt-only | 0.487 | 0.276 | 0.211 | 0.352 | 0.348 | 0.697 |
| | FutureGen | 0.451 | 0.289 | 0.162 | 0.347 | 0.136 | 0.388 |
| | CoI | 0.469 | 0.311 | 0.158 | 0.377 | 0.340 | 0.640 |
| GPT-4.0 | GoAI-GPT4 | 0.454 | 0.306 | 0.148 | 0.336 | 0.200 | 0.565 |
| | CoI-GPT4 | 0.486 | 0.310 | 0.176 | 0.336 | 0.410 | 0.770 |
| | ResearchAgent-GPT4 | 0.457 | 0.304 | 0.152 | 0.336 | 0.230 | 0.520 |
| GIANTS | GIANTS | 0.489 | 0.282 | 0.207 | 0.363 | 0.103 | 0.329 |
| DeepSeek-Math 7B | Text-only (LoRA) | 0.463 | 0.287 | 0.176 | 0.362 | 0.369 | 0.738 |
| | **COMPOSE (ours)** | **0.525** | 0.285 | **0.240** | **0.429** | **0.508** | **0.808** |

Table 3: Human ratings (1–5) and best-output selections.

| System | Rel. | Correct. | Nov. | Spec. | Res. | Mean | Wins |
|---|---|---|---|---|---|---|---|
| GIANTS | 3.42 | 3.51 | 3.10 | 3.26 | 3.18 | 3.29 | 6 |
| Sci.-only | 2.65 | 3.60 | 1.31 | 2.44 | 3.39 | 2.68 | 2 |
| GoAI | 1.79 | 3.72 | 3.50 | 2.62 | 3.10 | 2.95 | 1 |
| Text-only | 3.58 | 1.81 | 1.62 | 1.74 | **4.25** | 2.60 | 5 |
| **COMPOSE** | **3.90** | **4.08** | **4.19** | **4.10** | 4.10 | **4.07** | **11** |

Table 4: LLM judge scores (1–5).

| System | Cont. | Depth | Nov. | Prec. | Spec. | Ovrl. |
|---|---|---|---|---|---|---|
| GIANTS | 1.91 | 2.75 | **3.27** | 2.53 | 3.28 | 2.75 |
| Sci.-only | 1.59 | 1.57 | 1.71 | 1.46 | 1.71 | 1.61 |
| GOAI-GPT | 2.52 | 2.69 | 3.24 | 2.90 | 3.29 | 2.93 |
| Text-only | 1.86 | 1.91 | 2.42 | 2.04 | 2.43 | 2.13 |
| **COMPOSE** | **3.45** | **3.52** | 3.10 | **3.23** | **3.52** | **3.36** |

Precision (0.560) and Match (0.730), confirming that its outputs align with future theorems at the formal and informal level, not just topically. With Mistral 7B, COMPOSE leads on retrieval Gap, H@10, H@100, and NMCP Match, while Prompt-only has higher NMCP Precision (0.575 versus 0.554) and Bag-of-Papers has the highest mathematical-token ratio (0.070).

Table 2 compares against external baselines on the same test set, where differences between models are sharpest. **COMPOSE** achieves the largest Gap (0.240) and the highest Exp-Sim (0.429), confirming that its outputs are not only closer to the specific target paper but also to the broader set of papers sharing the same citation context. While the other models achieve high Tgt-Sim, their Neg-Sim is equally high, meaning they capture the right topic but fail to isolate the specific future result. **COMPOSE** ranks the correct future papers in the top 10 in 50.8%, outperforming all baselines. GIANTS performs worse on retrieval because its summarization-style outputs stay closer to the broader cluster than to the specific target paper. As shown in Figure 3, **COMPOSE** consistently dominates across all Hits@K.

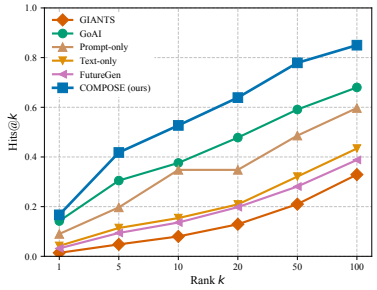



Figure 3: Hit@k comparison across models from $k = 1$ to 100.

COMPOSE also achieves the highest H@10 among the compared systems under three of five frozen external evaluators that were not fine-tuned on our corpus; Appendix C reports all five evaluators and confidence intervals.

### 4.3 Human and LLM Evaluation

In a blinded study, 30 mathematics graduate students evaluated outputs for 25 anchor papers on relevance, perceived mathematical correctness, novelty, specificity, and research value (1–5). Table 3 reports mean scores and best-output selections; Appendix B.5 includes the full result.

COMPOSE leads on four criteria and overall mean (4.07), with 11/25 best-output selections. These judgments support perceived plausibility and usefulness. Moreover retrieval evaluator and aggregated human scores select the same best system in 15/25 cases. Appendix B.5 provides the full results.

The LLM judge also ranks COMPOSE highest overall (3.36/5; Table 4), leading on mathematical content, technical depth, precision, and specificity. Thus, both human and LLM evaluations agree that COMPOSE performs best overall. Appendix B.6 provides the full LLM results.

In a complementary Lean diagnostic, 32/50 COMPOSE statements elaborate successfully, compared with 24/50 for Prompt-only, 22/50 for GIANTS, and 13/50 for GoAI. Appendix B provides the full protocols, uncertainty estimates, and alignment-quality and coverage analyses.

### 4.4 Ablation Study

We evaluate the contribution of each component of our model using the DeepSeek-Math 7B decoder. Table 5 reports results for five ablations: removing one encoder, removing the bidirectional fusion, removing stage-1 pretraining. Each ablation keeps all other components fixed.

Table 5: Ablation study (DeepSeek-Math 7B decoder). Each row removes one component from the full model.

| **Model** | **Tgt-Sim** | **Neg-Sim** | **Gap** | **H@10** | **H@100** |
|---|---|---|---|---|---|
| Paper-graph-only | 0.438 | 0.274 | 0.164 | 0.390 | 0.695 |
| formal-graph-only | 0.347 | 0.275 | 0.073 | 0.135 | 0.360 |
| w/o fusion | 0.367 | 0.277 | 0.090 | 0.195 | 0.530 |
| w/o stage 1 | 0.352 | 0.260 | 0.093 | 0.240 | 0.505 |
| **COMPOSE(ours)** | **0.525** | 0.285 | **0.240** | **0.508** | **0.808** |

Table 5 shows that each major component of COMPOSE contributes to retrieval performance. Removing either encoder causes a drop: without $G_s$, H@10 falls to 0.135, and without $G_f$, to 0.390, confirming that both graph modalities contribute complementary signals. Removing bidirectional fusion and stage-1 pretraining further degrades performance, confirming that the scientific and formal graphs provide complementary information. We also achieve the highest Tgt-Sim (0.525), confirming that the full dual-graph model produces text that is not only better ranked but also semantically closer to the correct future paper.

### 4.5 Smaller Decoder Analysis

The advantage also extends to smaller decoders: COMPOSE improves over the scientific-only model with both Qwen2-0.5B and Llama-3.2-3B (Table 6).

Table 6: Results with smaller decoder backbones.

| Decoder | Method | H@10 | H@100 |
|---|---|---|---|
| Qwen2-0.5B | Sci.-only | 0.070 | 0.220 |
| | COMPOSE | **0.148** | **0.344** |
| Llama-3.2-3B | Sci.-only | 0.081 | 0.344 |
| | COMPOSE | **0.222** | **0.484** |

Across three training seeds, COMPOSE consistently outperforms Bag-of-Papers with both 7B backbones.

### 4.6 Formal Graph Analysis

We test whether the formal-graph gain comes from the aligned dependency structure rather than arbitrary formal input. On the same 400 anchors, the aligned graph substantially outperforms both a size-matched random Mathlib graph and an edge-shuffled graph (Table 7).

Table 7: Formal-graph controls on 400 anchors.

| Formal input | H@10 | H@100 |
|---|---|---|
| Aligned graph | **0.455** | **0.660** |
| Same-domain random | 0.200 | 0.440 |
| Edge-shuffled | 0.250 | 0.582 |

We also evaluated four formal-graph conditions. Exact contains 20 test examples with correctly aligned formal graphs. For each Exact example, every aligned formal node was manually confirmed to match its corresponding informal theorem. Related, Weak, and Incorrect contain 200 examples each.

For this diagnostic, Related uses a similarity threshold of $\geq 0.84$ together with explicit mathematical word overlap. Weak contains matches with similarity $\geq 0.60$ but $< 0.84$. Incorrect is a control constructed from formal material in another mathematical topic:

Table 8: Retrieval gains by formal-alignment quality.

| Quality | $n$ | COMPOSE H@10 | Scientific-only H@10 | $\Delta$H@10 | $\Delta$H@100 |
|---|---|---|---|---|---|
| Exact | 20 | 0.667 | 0.300 | $+0.367$ | $+0.500$ |
| Related | 200 | 0.468 | 0.305 | $+0.163$ | $+0.309$ |
| Weak | 200 | 0.255 | 0.261 | $-0.006$ | $-0.250$ |
| Incorrect | 200 | 0.175 | 0.312 | $-0.137$ | $-0.274$ |

**COMPOSE** has the largest H@10 and H@100 gains on manually verified Exact subgraphs. The H@10 gain decreases for Related, becomes negligible for Weak, and is negative for Incorrect; Weak and Incorrect also reduce H@100. This is the expected control behavior: precise formal alignment helps, whereas weak or off-topic formal information does not provide an advantage.

We also evaluate three separate 400-anchor subsets. This coverage diagnostic uses a stricter reliable-match threshold of $\mathrm{sim} \geq 0.9$: the 1–2 and $\geq 3$ buckets contain the corresponding numbers of matches above this threshold per anchor. Weak-only deliberately uses $0.60 \leq \mathrm{sim} < 0.80$; this differs from the Weak condition in the alignment-quality diagnostic above.

Table 9: Retrieval by reliable-match coverage; each row contains 400 anchors. S-only denotes Scientific-only.

| Formal matches | COMPOSE H@10 | S-only H@10 | $\Delta$H@10 | COMPOSE H@100 | S-only H@100 | $\Delta$H@100 |
|---|---|---|---|---|---|---|
| Weak-only | 0.252 | 0.300 | $-0.048$ | 0.509 | 0.512 | $-0.003$ |
| 1–2 ($\geq 0.9$) | 0.486 | 0.340 | $+0.146$ | 0.728 | 0.500 | $+0.228$ |
| $\geq 3$ ($\geq 0.9$) | 0.601 | 0.390 | $+0.211$ | 0.848 | 0.520 | $+0.328$ |

Reliable formal matches improve retrieval within these subsets, with larger observed gains in the higher-coverage bucket; weak matches provide no clear improvement.

Across all three analyses, accurately aligned formal graphs consistently improve retrieval, with the largest gains from exact alignments and multiple reliable matches. Random, shuffled, weak, or unrelated formal inputs provide little or negative benefit, showing that the gains depend on alignment quality and coverage.

## 5 Qualitative Analysis

Appendix F presents six winning cases and one failure, with side-by-side model outputs.

## 6 Limitations and Future Work

Our method relies on approximate informal-formal alignment, which can introduces noise into the model. More importantly, generated claims are not verified by a proof assistant, so retrieval, Human experts and LLM-based evaluation remain proxy signals rather than direct tests of formal correctness.

A natural next direction is to incorporate proof-aware verification so that generated claims can be checked directly against a proof assistant. This framework could also support iterative conjecture generation and proof attempts.

## 7 Conclusion

We introduced grounded future mathematical generation, a task that combines scientific citation context with formal theorem structure to generate future mathematical claims. To support this setting, we constructed a paired scientific-formal dataset and a temporally grounded benchmark over 47K future papers, and proposed COMPOSE, a dual-graph framework for graph-conditioned generation.

Experiments show that COMPOSE outperforms strong baselines on retrieval to future papers and produces more grounded theorem-like predictions than text-only or single-graph alternatives. These results highlight the value of combining scientific context with formal structure for future mathematical generation.

## References


Alex A. Alemi, Francois Chollet, Niklas Een, Geoffrey Irving, Christian Szegedy, and Josef Urban. Deepmath - deep sequence models for premise selection, 2017. URL https://arxiv.org/abs/1606.04442.

Yousef Alhessi, Sólrún Halla Einarsdóttir, George Granberry, Emily First, Moa Johansson, Sorin Lerner, and Nicholas Smallbone. Lemmanaid: Neuro-symbolic lemma conjecturing, 2026. URL https://arxiv.org/abs/2504.04942.

Chenxin An, Ming Zhong, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Enhancing scientific papers summarization with citation graph, 2021. URL https://arxiv.org/abs/2104.03057.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics, 2023. URL https://arxiv.org/abs/2302.12433.

Ibrahim Al Azher, Miftahul Jannat Mokarrama, Zhishuai Guo, Sagnik Ray Choudhury, and Hamed Alhoori. Futuregen: A rag-based approach to generate the future work of scientific article, 2025. URL https://arxiv.org/abs/2503.16561.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models, 2025. URL https://arxiv.org/abs/2404.07738.

Bruno Barras, Samuel Boutin, Cristina Cornes, Judicaël Courant, Yann Coscoy, David Delahaye, Daniel de Rauglaudre, Jean-Christophe Filliâtre, Eduardo Giménez, Hugo Herbelin, et al. The coq proof assistant reference manual. *INRIA, version*, 6(11):17–21, 1999.

Rafael Cabral, Tuan Manh Do, Xuejun Yu, Wai Ming Tai, Zijin Feng, and Xin Shen. Proofflow: A dependency graph approach to faithful proof autoformalization, 2025. URL https://arxiv.org/abs/2510.15981.

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887):70–74, 2021. doi: 10.1038/s41586-021-04086-x.

Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving, 2025. URL https://arxiv.org/abs/2502.00212.

Emily First, Markus N. Rabe, Talia Ringer, and Yuriy Brun. Baldur: Whole-proof generation and repair with large language models, 2023. URL https://arxiv.org/abs/2303.04910.

Guoxiong Gao, Haocheng Ju, Jiedong Jiang, Zihan Qin, and Bin Dong. A semantic search engine for mathlib4. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8001–8013, 2024a.

Guoxiong Gao, Haocheng Ju, Jiedong Jiang, Zihan Qin, and Bin Dong. A semantic search engine for mathlib4. 2024b. URL https://arxiv.org/abs/2403.13310.

Xian Gao, Zongyun Zhang, Ting Liu, and Yuzhuo Fu. Goai: Enhancing ai students' learning paths and idea generation via graph of ai ideas, 2025. URL https://arxiv.org/abs/2503.08549.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017. URL https://arxiv.org/abs/1704.01212.

Joy He-Yueya, Anikait Singh, Ge Gao, Michael Y. Li, Sherry Yang, Chelsea Finn, Emma Brunskill, and Noah D. Goodman. Giants: Generative insight anticipation from scientific literature, 2026. URL https://arxiv.org/abs/2604.09793.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Prithwish Jana, Kaan Kale, Ahmet Ege Tanriverdi, Cruise Song, Sriram Vishwanath, and Vijay Ganesh. ProofBridge: Auto-Formalization of Natural Language Proofs in Lean via Joint Embeddings. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.15681.

Albert Q. Jiang, Sean Welleck, Jin Peng Zhou, Wenda Li, Jiacheng Liu, Mateja Jamnik, Timothée Lacroix, Yuhuai Wu, and Guillaume Lample. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs, 2023. URL https://arxiv.org/abs/2210.12283.

Kazumi Kasaura, Naoto Onda, Yuta Oriike, Masaya Taniguchi, Akiyoshi Sannai, and Sho Sonoda. Discovering new theorems via llms with in-context proof learning in lean, 2025. URL https://arxiv.org/abs/2509.14274.

Imre Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press, 1976.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. A survey on deep learning for theorem proving, 2024. URL https://arxiv.org/abs/2404.09939.

Xinyu Liu, Zixuan Xie, Amir Moeini, Claire Chen, Shuze Daniel Liu, Yu Meng, Aidong Zhang, and Shangtong Zhang. Mathliblemma: Folklore lemma generation and benchmark for formal mathematics, 2026. URL https://arxiv.org/abs/2602.02561.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. S2orc: The semantic scholar open research corpus, 2020. URL https://arxiv.org/abs/1911.02782.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024a. URL https://arxiv.org/abs/2408.06292.

Jianqiao Lu, Yingjia Wan, Zhengying Liu, Yinya Huang, Jing Xiong, Chengwu Liu, Jianhao Shen, Hui Jin, Jipeng Zhang, Haiming Wang, Zhicheng Yang, Jing Tang, and Zhijiang Guo. Process-driven autoformalization in lean 4, 2024b. URL https://arxiv.org/abs/2406.01940.

The mathlib Community. The lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, POPL '20, pages 367–381. ACM, January 2020. doi: 10.1145/3372885.3373824. URL http://dx.doi.org/10.1145/3372885.3373824.

Naoto Onda, Kazumi Kasaura, Yuta Oriike, Masaya Taniguchi, Akiyoshi Sannai, and Sho Sonoda. Lean-conjecturer: Automatic generation of mathematical conjectures for theorem proving, 2025. URL https://arxiv.org/abs/2506.22005.

Lawrence C Paulson. *Isabelle: A generic theorem prover*. Springer, 1994.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. Human-llm compound system for scientific ideation through facet recombination and novelty evaluation, 2026. URL https://arxiv.org/abs/2409.14634.

Piotr Rudnicki. An overview of the mizar project. In *Proceedings of the 1992 Workshop on Types for Proofs and Programs*, pages 311–330, 1992.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, 2024. URL https://arxiv.org/abs/2409.04109.

Elif Uskuplu, Lawrence S. Moss, and Valeria de Paiva. Knowtex: Visualizing mathematical dependencies, 2025. URL https://arxiv.org/abs/2601.15294.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, Jianqiao Lu, Hugues de Saxcé, Bolton Bailey, Chendong Song, Chenjun Xiao, Dehao Zhang, Ebony Zhang, Frederick Pu, Han Zhu, Jiawei Liu, Jonas Bayer, Julien Michel, Longhui Yu, Léo Dreyfus-Schmidt, Lewis Tunstall, Luigi Pagani, Moreira Machado, Pauline Bourigault, Ran Wang, Stanislas Polu, Thibaut Barroyer, Wen-Ding Li, Yazhe Niu, Yann Fleureau, Yangyang Hu, Zhouliang Yu, Zihan Wang, Zhilin Yang, Zhengying Liu, and Jia Li. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning, 2025a. URL https://arxiv.org/abs/2504.11354.

Hanyu Wang, Ruohan Xie, Yutong Wang, Guoxiong Gao, Xintao Yu, and Bin Dong. Aria: An agent for retrieval and iterative auto-formalization via dependency graph, 2025b. URL https://arxiv.org/abs/2510.04520.

Heng Wang, Pengcheng Jiang, Jiashuo Sun, Zhiyi Shi, Haofei Yu, Jiawei Han, and Heng Ji. Learning to predict future-aligned research proposals with language models, 2026. URL https://arxiv.org/abs/2603.27146.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024a. URL https://arxiv.org/abs/2212.03533.

Mingzhe Wang, Yihe Tang, Jian Wang, and Jia Deng. Premise selection for theorem proving by deep graph embedding, 2017. URL https://arxiv.org/abs/1709.09994.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.acl-long.18. URL http://dx.doi.org/10.18653/v1/2024.acl-long.18.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with large language models, 2022. URL https://arxiv.org/abs/2205.12615.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data, 2024. URL https://arxiv.org/abs/2405.14333.

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Improving scientific hypothesis generation with knowledge grounded large language models, 2024. URL https://arxiv.org/abs/2411.02382.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. LeanDojo: Theorem proving with retrieval-augmented language models. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Lan Zhang, Marco Valentino, and Andre Freitas. Autoformalization in the wild: Assessing llms on real-world mathematical definitions, 2025. URL https://arxiv.org/abs/2502.12065.

Thomas Zhu, Pietro Monticone, Jeremy Avigad, and Sean Welleck. Leanarchitect: Automating blueprint generation for humans and ai, 2026. URL https://arxiv.org/abs/2601.22554.

# A Dataset Construction Details

Table 10 summarizes the distribution of papers across mathematical domains in our dataset. The dataset includes papers published between 2000 and 2023.

Table 10: Dataset statistics by mathematical domain.

| Domain | # Papers |
|---|---|
| Probability | 119,644 |
| Algebraic Geometry | 113,360 |
| Analysis | 104,483 |
| Mathematical Physics | 43,035 |
| Representation Theory | 42,720 |
| Dynamical Systems | 39,153 |
| Differential Geometry | 34,138 |
| Combinatorics | 27,220 |
| **Total (all domains)** | **520,773** |

## A.1 Citation Subgraph Construction

Starting from the anchor paper, we follow outgoing citation edges hop by hop. We retain up to 5 cited papers at the first hop and up to 3 cited papers per node for the last hops. Candidate neighbors are ranked by citation-context relevance rather than sampled uniformly. Nodes without publication year metadata are discarded. Each paper node is represented by its abstract, encoded with E5-large-v2 Wang et al. [2024a].

Since not all citations are equally informative, we rank candidate neighbors at each expansion step relative to the current paper node. The ranking uses citation-context heuristics intended to capture relevance to the surrounding scientific development, including appearances in the abstract, introduction, main result claim, or proof-related passages; explicit dependency cues such as "using," "by," or "follows from"; repeated technical mentions; and proximity to extracted theorem. At each hop, we retain only the top-k candidates, yielding a compact citation subgraph that preserves the most relevant local scientific context.

### A.1.1 Citation Relevance Selection: Illustrative Example

To illustrate the difference between relevant and peripheral citations, we show two citations from the same paper (ArXiv:1911.06307, Grifo-Ma-Schwede) and explain why one is selected as a neighbor and the other is not.

**High-relevance citation.** *"arXiv:1702.06876, Grifo-Huneke"* This paper is cited throughout the introduction and proof sections with explicit dependency cues. For example: *"In prime characteristic $p$, Grifo and Huneke [GH19] showed that Harbourne's conjecture holds for ideals defining F-pure rings"*, *"the theorem above extends [GH19, Theorem 3.2 and Theorem 4.1]"*, and *"The containment $Q^{((h-1)(n-1)+1)} \subseteq Q^n$ for all $n \geq 1$ now follows by induction on $n$, as in [GH19, Theorem 4.1]"*. The paper's main results are direct extensions of this work, and specific lemmas are invoked repeatedly in proofs.

**Low-relevance citation (not selected).** *"Birational geometry of algebraic varieties" (Kollár-Mori).* This work does not appear anywhere in the body of the paper — it appears only in the citation subgraph through indirect bibliographic links. It is not cited in the introduction, proofs, or any technical passage, and has no connection to the symbolic power containment results developed in the paper.

Citations that appear repeatedly in proof-related passages with explicit dependency language ("extends [X]", "as in [X, Lemma Y]", "follows from [X]") score significantly higher than works that appear only through indirect bibliographic links with no direct textual dependency. This ensures the subgraph captures the papers that directly shaped the mathematical content of the target paper.

### A.2 Proof-Flow Graph Extraction

For papers within the first two citation hops, we extract a proof-flow graph from the full text. We identify theorem nodes corresponding to five types: *theorem*, *lemma*, *proposition*, *corollary*, and *definition*. Directed edges are added between theorem nodes when the text contains explicit dependency references, detected with regular expressions matching patterns such as "by Lemma 2.1" or "using Theorem 3.4". We filter out malformed statement spans, including nodes that begin mid-sentence. Papers with fewer than three valid theorem nodes after filtering are discarded. The resulting graph is a directed acyclic graph that approximates the proof-flow structure of the paper.

### A.3 Formal Graph Construction

The matched theorem serves as the root of a local dependency subgraph extracted from Mathlib, where nodes are formal Lean theorems and edges capture dependency relations between them. We expand this subgraph up to 3 hops from the root, yielding between 8 and 15 theorems per sample.

### A.4 Informal-to-Formal Alignment

Table 11: Number of extracted theorems and high-confidence Mathlib matches (cosine $\geq 0.84$) per mathematical domain.

| Domain | Extracted Theorems | Matched Theorems |
|---|---|---|
| Probability | 38,499 | 8,846 |
| Algebraic Geometry | 53,118 | 5,779 |
| Analysis | 40,650 | 9,534 |
| Mathematical Physics | 26,864 | 3,974 |
| Representation Theory | 21,836 | 3,263 |
| Dynamical Systems | 25,312 | 4,137 |
| Differential Geometry | 32,326 | 2,652 |
| Combinatorics | 38,690 | 5,325 |
| Other | 435,553 | 76,915 |
| **Total** | **712,848** | **120,425** |

Bridging informal mathematical text and formal Lean theorems is challenging because of the gap between natural mathematical language and formal library theorems. To understand which alignment strategy is most effective, we compare three methods. The first is Autoformalization, which translates an informal theorem into a Lean-style formal stub before retrieval. The second is informal-to-formal retrieval, which avoids generation and retrieves directly from the informal theorem into a space of formal Lean theorem signatures. The third is informal-to-informal retrieval, which avoids the formal surface form by retrieving against natural-language theorem descriptions from FrenzyMath and then mapping the retrieved description back to its corresponding Mathlib theorem.

As shown in Table 12, Autoformalization-based approaches achieve at most 6.1% high-confidence matches, while direct informal-to-formal retrieval performs similarly poorly, confirming that the vocabulary mismatch between informal and formal mathematics is a major bottleneck. In contrast, informal-to-informal retrieval achieves 16.9% high-confidence matches, nearly three times better than the strongest competing method. We therefore adopt this alignment strategy in all downstream experiments.

The Precision column in Table 12 is computed by sampling 1000 matched pairs per method, scoring them with an LLM judge, and reporting the resulting percentage. The full prompt is shown in Figure 4.

Table 13 shows the coverage precision tradeoff as the threshold $\tau$ varies for the adopted E5-MISTRAL $\rightarrow$ FrenzyMath method.

Table 12: Comparison of alignment strategies for matching clean informal paper theorems to Mathlib theorem. Coverage: % of statements with any match. High-conf.: % of statements with cosine similarity $\geq 0.84$. Precision: % of high-confidence matches judged as related or better by an LLM judge on 1000 sampled pairs.

| Strategy | Method | Coverage(%) | High-conf. (%) | Precision (%) |
|---|---|---|---|---|
| Autoformalization | Kimina-7B → BM25 | 23.0% | 0.4% | 0.0% |
| | Kimina-7B → namespace | 25.0% | 0.1% | 0.5% |
| | Kimina-7B → Lean sig. | 100% | 5.9% | 7.0% |
| | Gemini-3 → BM25 | 99.4% | 6.1% | — |
| Informal → Formal | Keyword → BM25 | 99% | 0.2% | 2.0% |
| | e5-large → Lean sig. | 100% | 8.7% | 1.5% |
| | e5-mistral → Lean sig. | 100% | 6.2% | 45.4% |
| Informal → Informal | BM25 → FrenzyMath | 100% | 1.2% | 7.7% |
| | e5-large → FrenzyMath | 100% | 3.4% | 26.1% |
| | e5-mistral → FrenzyMath **(ours)** | 100% | **16.9%** | **66.1%** |

Table 13: Coverage and precision for the E5-MISTRAL → FrenzyMath alignment pipeline as a function threshold $\tau$. The row at $\tau = 0.84$ is the operating point used in our experiments.

| $\tau$ | High-conf. (%) | precision (%) |
|---|---|---|
| 0.70 | 45.2 | 31.0 |
| 0.75 | 32.1 | 42.5 |
| 0.80 | 24.3 | 55.8 |
| 0.84 | 16.9 | 66.1 |
| 0.88 | 10.1 | 78.3 |
| 0.92 | 4.2 | 88.0 |

### A.4.1 Alignment Retrieval Details

We describe the retrieval implementation for each method in Table 12. Autoformalization and informal-to-formal methods retrieve from the Mathlib corpus theorems indexed by their Lean 4 signatures. High-confidence matches are defined as cosine similarity $\geq 0.84$ between the paper theorem and the retrieved Mathlib theorem.

**Autoformalization.** Each informal paper theorem is first translated into a Lean 4 formal stub using either Kimina-7B [Wang et al., 2025a] or Gemini-2.5-Flash, prompted with five in-context examples. The prompt is shown in Figure A.4.1. Three retrieval strategies are then applied to the generated stub:

- **BM25 retrieval.** We build a BM25 index over Mathlib theorem names concatenated with their Lean 4 signatures. We retrieve the top-5 results with a minimum BM25 score of 5.
- **Namespace matching.** We build a prefix index over all Mathlib theorem names. From the generated stub, we extract identifiers and look up Mathlib theorems whose name shares the longest matching prefix. Results are ranked by prefix specificity, top-5 returned.
- **Dense retrieval.** We embed the informal text as `"query: <text>"` using e5-large-v2(1024-dim, L2-normalized) and compute cosine similarity against precomputed embeddings of all Mathlib Lean 4 signatures (embedded as `"passage: <sig>"`). Top-10 by cosine score.

**Informal-to-formal retrieval.** These methods bypass Autoformalization and retrieve directly from the informal paper theorem text into the space of formal Lean 4 signatures:

- **Keyword → BM25.** We tokenize both queries and the Mathlib signature corpus by lowercasing, extracting alphanumeric tokens via regex, and removing a custom stopword list. We build a BM25 index over the tokenized Lean 4 signatures and retrieve the top-5 results with a minimum score of 1.0.

LLM Judge Prompt

You are a mathematical alignment judge. Given an informal theorem from a research paper and a candidate Mathlib (Lean 4) theorem, rate how well the Mathlib theorem aligns with the informal theorem.

**Score from 1 to 5:**
1 = Completely unrelated mathematics
2 = Same broad area but different specific topic
3 = Related topic, shares key concepts or structures
4 = Closely related, formalizes a component or special case of the informal result
5 = Direct match, formalizes the same or equivalent result

Respond with only a single digit: 1, 2, 3, 4, or 5.

**Informal theorem from paper:** {informal}
**Candidate Mathlib theorem:** {formal}

Figure 4: Prompt used to evaluate alignment quality for the Precision column in Table 12. We use Gemini-2.5-Flash with temperature 0.

- **e5-large → Lean sig.** We embed the informal text as `"query: <text>"` using e5-large-v2 and compute cosine similarity against precomputed embeddings of all Mathlib Lean 4 signatures (embedded as `"passage: <sig>"`). Top-10 by cosine score.
- **e5-mistral → Lean sig.** Same setup but using e5-mistral-7b-instruct with last-token pooling and the instruction prefix `"Instruct: Retrieve a Mathlib theorem that formalizes the following mathematical theorem."`. Corpus signatures are embedded without instruction prefix. Top-5 by cosine score.

**Informal-to-informal retrieval.** These methods retrieve against natural-language theorem descriptions from the FrenzyMath dataset, which provides informal names and descriptions for Mathlib theorems. The retrieved description is then mapped back to its corresponding formal Mathlib theorem:

- **BM25 → FrenzyMath.** We build a BM25 index over Mathlib theorem names concatenated with their Lean 4 signatures. We retrieve the top-5 results with a minimum BM25 score of 5.
- **e5-large → FrenzyMath.**We embed informal descriptions from FrenzyMath as "passage: <desc>" with e5-large-v2 (83,807 unique entries) and embed each paper theorem as "query: <text>". Top-5 by cosine similarity.
- **e5-mistral → FrenzyMath (ours).** We use e5-mistral-7b-instruct with last-token pooling and the instruct prefix `"Instruct: Retrieve a Mathlib theorem that formalizes the following mathematical theorem."`. We retrieve the top-5 candidates by cosine similarity.

Autoformalization Prompt (Gemini-2.0-Flash / Kimina-Autoformalizer-7B)

**System:** You are an expert in Lean4 and Mathlib. Convert the following mathematical theorem statement into a Lean4 theorem stub using Mathlib.
**Rules:**

- Use correct Mathlib 4 types and namespaces (e.g. `Fintype.card`, `IsCyclic`, `Nat.Prime`)
- Write only the theorem signature ending with `:= by sorry` or `:= sorry`
- Do NOT write a proof, just the stub
- Use `import Mathlib` and appropriate `open` statements

**[5 in-context examples:]**

**Text:** Let $G$ be a finite cyclic group. Then the exponent of $G$ equals the cardinality of $G$.
**Lean4:**
```
theorem cyclic_exponent_eq_card {G : Type*} [Group G] [Fintype G]
[IsCyclic G] :
    Monoid.exponent G = Fintype.card G := by exact IsCyclic.exponent_eq_card
```

**Text:** Let $p$ be a prime number. Then a finite group of order $p$ is cyclic.
**Lean4:**
```
theorem prime_order_cyclic {G : Type*} [Group G] [Fintype G] {p : ℕ}
[hp : Fact p.Prime]
    (h : Fintype.card G = p) : IsCyclic G := by exact
isCyclic_of_prime_card h
```

**Text:** Let $R$ be a commutative ring and $I$ an ideal. The degree of the zero element of $R[X]$ is 0.
**Lean4:**
```
theorem degreeOf_C_zero {R : Type*} [CommRing R] (x : ℕ) :
    MvPolynomial.degreeOf x (0 : MvPolynomial ℕ R) = 0 := by simp
[MvPolynomial.degreeOf_C]
```

**Text:** If $f$ is `ContDiff` of order $n$, then for any polynomial $p$, $x \mapsto p(x^{-1}) \cdot \mathrm{expNegInvGlue}(x)$ is also `ContDiff` of order $n$.
**Lean4:**
```
theorem contDiff_poly_inv_glue {n : ℕ∞} (p : ℝ[X]) :
    ContDiff ℝ n (fun x ↦ p.eval x⁻¹ * expNegInvGlue x) :=
    expNegInvGlue.contDiff_polynomial_eval_inv_mul p
```

**Text:** Let $G$ be a finite group with normal subgroup $N$ s.t. $G/N$ is a Frobenius group with cyclic complement. If $|F|$ and $|N|$ are coprime, then some prime divides $|G|$.
**Lean4:**
```
theorem frobenius_prime_divisor {G : Type*} [Group G] [Fintype G]
    {N : Subgroup G} (hN : N.Normal) (hC : IsCyclic (G ⧸ N))
    (hcard : Nat.Coprime (Fintype.card N) (Fintype.card (G ⧸ N))) :
    ∃ p : ℕ, Nat.Prime p ∧ p ∣ Fintype.card G := by
    exact ⟨_, Nat.minFac_prime (Fintype.card_pos.ne'), Nat.minFac_dvd _⟩
```

**Text:** *{informal_theorem}*
**Lean4:** *[generated stub]*

Figure 5: Few-shot prompt used for autoformalization (Gemini-2.0-Flash). The generated Lean 4 stub is used as a query for BM25, namespace, and dense retrieval against Mathlib.

### A.5 Training Target Types

We use the same target-construction pipeline for both the training dataset and the future-paper retrieval benchmark. For each paper, we first try to extract its main mathematical claim from the body text using a set of heuristics, including regular-expression matching for phrases such as "we prove," "we show," "Main Theorem," and early numbered theorem statements. When no single main claim can be identified reliably, we fall back to alternative targets such as the abstract or a supporting theorem with clear local context. This yields a consistent target-construction process across both training and evaluation, while handling different ways of stating the main claim.

Each training sample is paired with several target types derived from the same graph. These targets expose the model to different forms of mathematical output from the same underlying context, helping it learn richer graph representations instead of specializing to a single output style.

A **Claim** is the main mathematical result of the paper, extracted from sentences beginning with "we prove," "we show," or identified as Theorem 1 or Main Theorem in the text. An **Abstract** target is the full paper abstract. A **Lemma in claim** is a supporting lemma together with the neighboring theorems it depends on, providing local proof flow context. **Lean only** targets consist of the formal

Lean signatures of the matched Mathlib theorems, and **Informal + Lean** combines the informal claim with those signatures. Training on this mixture encourages the model to learn graph representations that generalize across output types, which benefits Claim generation at evaluation time.

Table 14 shows the distribution across target types in the final dataset.

Table 14: Training target types and sample counts in the final dataset.

| **Target Type** | **Samples** |
|---|---|
| Abstract | 12,156 |
| Claim | 33,846 |
| Claim + proof flow | 14,423 |
| Lemma in claim | 42,517 |
| Lean only | 2,817 |
| Informal + Lean | 2,817 |
| **Total** | **108,576** |

### A.6 Illustrative Example

We illustrate the full structure of one training sample from the dataset.

**Target paper.** ArXiv:1911.06307 (Grifo-Ma-Schwede) — symbolic power containments in singular rings in positive characteristic.

**Target statement (claim).**

> *Let $R$ be an $F$-finite Gorenstein ring of prime characteristic $p$ and $Q \subseteq R$ an ideal of finite projective dimension with big height $h$. If $R/Q$ is $F$-pure, then $Q^{(hn-h+1)} \subseteq Q^n$ for all $n \geq 1$. If $R/Q$ is strongly $F$-regular and $h \geq 2$, then $Q^{((h-1)(n-1)+1)} \subseteq Q^n$ for all $n \geq 1$.*

**Citation subgraph (7 paper nodes, 6 edges).** The citation subgraph retains the 7 highest-relevance cited papers, selected by citation-context relevance as described in Appendix A.1:

- Hochster-Huneke: Comparison of symbolic and ordinary powers of ideals (2002) — cited 10 times, invoked in proofs as "by [HH02, Lemma 2.2]" and "by [HH02, Lemma 2.4]"
- Grifo-Huneke: Symbolic powers of ideals defining $F$-pure and strongly $F$-regular rings (2019) — cited 9 times, main results are direct extensions: "the theorem above extends [GH19, Theorem 3.2 and Theorem 4.1]"
- Takagi: Formulas for multiplier ideals on singular varieties (2006) — cited 10 times, specific lemmas invoked: "by [Tak06, Theorem 4.6]", "follows from [Tak06, Lemma 4.5]"
- Fedder: $F$-purity and rational singularity (1983) — cited 6 times, the Fedder criterion is a central tool: "c.f. [Fed83, Corollaries to Lemma 1.6]"
- Peskine-Szpiro: Dimension projective finie et cohomologie locale (1973) — cited in proof: "By [PS73, Théorème (I.7)]"
- Takagi-Yoshida: Generalized test ideals and symbolic powers (2008) — cited 4 times with specific result: "extends [TY08, Remark 3.4]"
- Glassbrenner: Strongly $F$-regularity in images of regular rings (1996) — cited 3 times, criterion extended: "extends [Gla96]"

Directed citation edges connect the anchor paper to each of these works; Grifo-Huneke (2019) is additionally connected to Hochster-Huneke (2002), reflecting the citation chain within the subgraph.

**Proof-flow graph of the anchor paper (10 nodes, 7 edges).** Statement nodes extracted from the paper text, with dependency edges detected from explicit references:

- **[definition]** Definition 2.3: The Frobenius non-splitting ideal $I^e(Q)$ for an ideal $Q$ in an $F$-finite ring $R$.
- **[lemma]** Lemma 2.6: $\mathrm{Ass}(R/I^e(Q)) \subseteq \mathrm{Ass}(R/Q)$ for all $e > 0$; equality holds when $R$ is $F$-pure.
- **[lemma]** Lemma 2.9: For $q = p^e$, $Q^{(q(h+n-1)-h+1)} \subseteq I^e(Q^{(n)})$.
- **[lemma]** Lemma 3.1: Any map $\phi \in \mathrm{Hom}_{R/Q}(F^e_*(R/Q), R/Q)$ lifts to $\tilde{\phi} \in \mathrm{Hom}_R(F^e_*R, R)$.
- **[theorem]** Theorem 3.11 (Main): $Q^{(hn-h+1)} \subseteq Q^n$ for $F$-pure rings; $Q^{((h-1)(n-1)+1)} \subseteq Q^n$ for strongly $F$-regular rings.

Dependency edges: Theorem 3.11 $\to$ Lemma 3.1, Theorem 3.11 $\to$ Lemma 2.9, Lemma 2.9 $\to$ Lemma 2.6, Lemma 3.1 $\to$ Definition 2.3.

**Mathlib theorem subgraph (10 nodes, 8 edges).** The formal subgraph is anchored at `Ideal.pow_le_pow_left`, the Mathlib theorem stating $I^m \le I^n$ whenever $I \le J$ or $m \ge n$, which is the closest formal counterpart to the symbolic power containment result. The subgraph expands through its proof dependencies:

- `Ideal.pow_le_pow_left`: $\forall\,(h : I \le J),\ I^n \le J^n$
- `Ideal.pow_le_pow_right`: $\forall\,(h : m \le n),\ I^n \le I^m$
- `Ideal.mul_le_left`: $I \cdot J \le I$
- `Ideal.span_pow_le_pow`: *containment of span under power*
- `Ideal.IsPrime.mul_mem_iff_mem_or_mem`: $x \cdot y \in P \Rightarrow x \in P \vee y \in P$
- `Ideal.mem_pow_eq_zero_iff`: *membership criterion in nilradical powers*
- `le_trans`: $a \le b \to b \le c \to a \le c$
- `Ideal.span_le`: $\mathrm{span}(s) \le I \leftrightarrow s \subseteq I$

Dependency edges: `Ideal.pow_le_pow_left` $\to$ `Ideal.mul_le_left`, `Ideal.pow_le_pow_left` $\to$ `le_trans`, `Ideal.pow_le_pow_right` $\to$ `le_trans`, `Ideal.IsPrime.mul_mem_iff_mem_or_mem` $\to$ `Ideal.span_le`.

## B Extended Evaluation Metrics

Table 15: Generation quality and grounding metrics. Struct. = mathematical structure rate. BERTSc. = BERTScore F1 vs. ground-truth future theorem. ROUGE-L = lexical overlap with ground-truth. Fut-R = ROUGE(gen, future) / ROUGE(gen, input); values >1 indicate forward-looking predictions. Bold = best per group

| Decoder | Model | Struct. | BERTSc. | ROUGE-L | Fut-R↑ |
|---|---|---|---|---|---|
| 1.223 | **COMPOSE (ours)** | **0.975** | **0.840** | **0.066** | |
| DeepSeek-Math 7B | Paper-graph-only | 0.940 | 0.826 | 0.053 | 1.084 |
| | Bag-of-Papers | 0.900 | 0.822 | 0.046 | 0.962 |
| | Text-only (LoRA) | 0.850 | 0.807 | 0.148 | 2.241 |
| | Prompt-only | 0.890 | 0.777 | 0.048 | 0.518 |
| Mistral 7B | **COMPOSE (ours)** | **0.877** | **0.819** | **0.050** | 0.441 |
| | Paper-graph-only | 0.644 | 0.808 | 0.050 | 0.625 |
| | Bag-of-Papers | 0.851 | 0.826 | 0.042 | 0.480 |
| External DeepSeek-Math 7B | GoAI | 0.370 | 0.771 | 0.047 | 0.809 |
| | FutureGen | 0.645 | 0.710 | 0.048 | 0.520 |
| | ResearchAgent | 0.082 | 0.648 | 0.054 | 0.699 |
| GIANTS | GIANTS | 0.165 | 0.752 | 0.037 | 0.314 |
| GPT-4.0 | GoAI | 0.085 | 0.821 | 0.125 | 1.933 |
| | ResearchAgent | 0.640 | 0.814 | 0.096 | 0.968 |
| | CoI | 0.080 | 0.811 | 0.118 | 1.449 |

Table 15 evaluates generation quality across four complementary dimensions. Struct. measures whether outputs contain formal mathematical content. Our model achieves 0.975, indicating that nearly every prediction contains meaningful mathematical content. BERTSc. captures semantic similarity to the ground-truth future theorem; our score of 0.840 is the highest across all methods, confirming that our predictions closely match what was actually published. ROUGE-L measures lexical overlap with the ground-truth, where we lead our decoder group. Fut-R is defined as:

$$\text{Fut-R} = \frac{\text{ROUGE-L}(\hat{y},\ y^*)}{\text{ROUGE-L}(\hat{y},\ x)} \tag{9}$$

where $\hat{y}$ is the generated text, $y^*$ is the ground-truth future theorem, and $x$ is the input paper context. A value above 1.0 means the model's output shares more content with the ground-truth future theorem than with the papers it was conditioned on, indicating forward-looking prediction. A value below 1.0 means the model is closer to restating its input than to predicting the future. External baselines such as GIANTS (0.314) and Prompt-only (0.518) fall well below 1.0, indicating that they largely restate known results. Our model achieves 1.223, confirming that it produces more forward-looking predictions.

### B.1 Smaller Decoder Backbones

We evaluate Qwen2-0.5B (`Qwen/Qwen2-0.5B`) and Llama-3.2-3B-Instruct under the same evaluation protocol as the 7B experiments. Scientific-only denotes the Paper-graph-only variant, which uses only the scientific graph. Table 16 compares each COMPOSE model with its matched Scientific-only baseline.

COMPOSE improves H@10/H@100 by 0.078/0.124 with Qwen2-0.5B and by 0.141/0.140 with Llama-3.2-3B-Instruct. Together with the 7B experiments, these results show gains across the evaluated decoder families and at 0.5B, 3B, and 7B scales.

Table 16: Retrieval performance with smaller decoder backbones.

| Decoder | Method | H@10 | H@100 |
|---|---|---|---|
| Qwen2-0.5B | Scientific-only | 0.070 | 0.220 |
| | COMPOSE | **0.148** | **0.344** |
| Llama-3.2-3B-Instruct | Scientific-only | 0.081 | 0.344 |
| | COMPOSE | **0.222** | **0.484** |

### B.2 Lean Statement-Elaboration Evaluation

We selected 50 outputs per method that contained a Lean theorem statement. We kept only the theorem statement from each output and removed all surrounding prose. We used by sorry only as a placeholder for the proof body and checked whether each theorem statement elaborated successfully in Lean:

Table 17: Lean statement elaboration on 50 selected outputs per method.

| Method | Statements elaborated |
|---|---|
| COMPOSE | **32/50** |
| GIANTS | 22/50 |
| Prompt-only | 24/50 |
| GoAI | 13/50 |

Compose achieves the highest statement-elaboration rate, with 32 of 50 outputs represented as well-formed and type- correct Lean theorem statements. For every elaborating statement, we also attempted to construct a complete Lean proof containing neither sorry nor admit; however, no proof was found and verified for any method under our proof-search procedure and computational budget. In this partial proof-assistant check, Compose achieves the highest Lean/Mathlib type-checking rate (32/50).

### B.3 Formal Alignment Quality and Coverage

To check whether a formal graph is correctly aligned with scientific data and assess the quality of the alignment, we added three complementary diagnostics: (i) controlled comparisons against size-matched random and edge-shuffled formal graphs; (ii) an analysis across Exact, Related, Weak, and Incorrect alignment-quality conditions; and (iii) performance across different levels of formal-match coverage.

**Random and edge-shuffled formal-graph controls.** In the default pipeline, matches with similarity $\geq 0.84$ are retained as formal roots. We evaluated 400 test anchors using their standard aligned formal graphs. The random-graph control replaces each aligned graph with a same-domain Mathlib graph of the same size, while the edge-shuffled control keeps the same nodes and number of edges but randomly rewires the graph.

Table 18: Formal-graph controls on the same 400 test anchors.

| Formal input | H@10 | H@100 |
|---|---|---|
| Aligned graph | **0.455** | **0.660** |
| Same-domain random graph | 0.200 | 0.440 |
| Edge-shuffled aligned graph | 0.250 | 0.582 |

The paired H@10 gain of the aligned graph over the edge-shuffled graph is $+0.205$ (95% CI $[+0.190, +0.220]$). The aligned graph also improves over the same-domain random graph by $+0.255$ H@10 (95% CI $[+0.211, +0.283]$).

**Alignment-quality conditions.** We also evaluated four formal-graph conditions. Exact contains 20 test examples with correctly aligned formal graphs. For each Exact example, every aligned formal node was manually confirmed to match its corresponding informal theorem. Related, Weak, and Incorrect contain 200 examples each. For this diagnostic, Related uses a similarity threshold of $\geq 0.84$ together with explicit mathematical word overlap. Weak contains matches with similarity $\geq 0.60$ but $< 0.84$. Incorrect is a control constructed from formal material in another mathematical topic:

Table 19: Retrieval gains by formal-alignment quality.

| Quality | $n$ | COMPOSE H@10 | Scientific-only H@10 | $\Delta$H@10 | $\Delta$H@100 |
|---|---|---|---|---|---|
| Exact | 20 | 0.667 | 0.300 | $+0.367$ | $+0.500$ |
| Related | 200 | 0.468 | 0.305 | $+0.163$ | $+0.309$ |
| Weak | 200 | 0.255 | 0.261 | $-0.006$ | $-0.250$ |
| Incorrect | 200 | 0.175 | 0.312 | $-0.137$ | $-0.274$ |

For the manually verified Exact subset, the paired H@10 gain is $\Delta\text{H@10} = +0.367$ with a 95% confidence interval of $[+0.200, +0.456]$. H@$k$ is averaged over three generations per anchor.

**COMPOSE** has the largest H@10 and H@100 gains on manually verified Exact subgraphs. The H@10 gain decreases for Related, becomes negligible for Weak, and is negative for Incorrect; Weak and Incorrect also reduce H@100. This is the expected control behavior: precise formal alignment helps, whereas weak or off-topic formal information does not provide an advantage.

**Coverage of reliable matches.** We evaluate three separate 400-anchor subsets. This coverage diagnostic uses a stricter reliable-match threshold of $\text{sim} \geq 0.9$: the 1–2 and $\geq 3$ buckets contain the corresponding numbers of matches above this threshold per anchor. Weak-only deliberately uses $0.60 \leq \text{sim} < 0.80$; this differs from the Weak condition in the alignment-quality diagnostic above.

Table 20: Retrieval by reliable-match coverage; each row contains 400 anchors. S-only denotes Scientific-only.

| Formal matches | COMPOSE H@10 | S-only H@10 | $\Delta$H@10 | COMPOSE H@100 | S-only H@100 | $\Delta$H@100 |
|---|---|---|---|---|---|---|
| Weak-only | 0.252 | 0.300 | $-0.048$ | 0.509 | 0.512 | $-0.003$ |
| 1–2 ($\geq 0.9$) | 0.486 | 0.340 | $+0.146$ | 0.728 | 0.500 | $+0.228$ |
| $\geq 3$ ($\geq 0.9$) | 0.601 | 0.390 | $+0.211$ | 0.848 | 0.520 | $+0.328$ |

Paired-bootstrap 95% CIs for $\Delta$H@10 and $\Delta$H@100, respectively, are: Weak-only, $[-0.110, +0.054]$ and $[-0.090, +0.073]$; 1–2 matches, $[+0.1125, +0.170]$ and $[+0.198, +0.298]$; at least three matches, $[+0.145, +0.244]$ and $[+0.226, +0.370]$. Reliable formal matches improve retrieval within these subsets, with larger observed gains in the higher-coverage bucket; weak matches provide no clear improvement.

These diagnostics show that **COMPOSE** benefits from correct formal alignment and graph structure, not arbitrary formal input.

### B.4 Training-Seed Robustness

For each decoder backbone, we independently trained COMPOSE and Bag-of-Papers with three random seeds, keeping the data split and hyperparameters fixed. Table 21 reports means and standard deviations across the three trained models. These standard deviations measure training-seed variation and are distinct from the paired test-set confidence intervals below.

For Mistral, the paired test-set gains of COMPOSE over Bag-of-Papers are $\Delta\text{Gap} = +0.022$ (95% CI $[0.016, 0.028]$), $\Delta\text{H@10} = +0.194$ ($[0.163, 0.225]$), and $\Delta\text{H@100} = +0.191$ ($[0.163, 0.219]$). All three confidence intervals exclude zero. COMPOSE consistently outperforms Bag-of-Papers across the three training seeds with both backbones.

### B.5 Human Evaluation: Full Scores

Table 17 describe the full score of the human evalution we did . Best overall mean that the mode

Table 21: Training-seed robustness: mean ± standard deviation across three independently trained models per configuration.

| Decoder | Model | Gap | H@10 | H@100 |
|---|---|---|---|---|
| Mistral | Bag-of-Papers | $0.139 \pm 0.002$ | $0.154 \pm 0.010$ | $0.506 \pm 0.013$ |
| | COMPOSE | $\mathbf{0.161 \pm 0.002}$ | $\mathbf{0.348 \pm 0.013}$ | $\mathbf{0.697 \pm 0.012}$ |
| DeepSeek | Bag-of-Papers | $0.128 \pm 0.002$ | $0.230 \pm 0.011$ | $0.544 \pm 0.013$ |
| | COMPOSE | $\mathbf{0.240 \pm 0.002}$ | $\mathbf{0.508 \pm 0.013}$ | $\mathbf{0.808 \pm 0.011}$ |

We also compare the system selected by the retrieval evaluator with the system receiving the highest aggregated human score for each anchor. They agree on 15 of the 25 anchors (60%), with moderate agreement beyond chance (Cohen's $\kappa = 0.46$), showing that the evaluation method is the right approach for the task.

Table 22: Blinded human evaluation on 25 anchor papers by 30 mathematics graduate students. Scores are means ± standard deviations across raters. Correct. denotes perceived correctness, not formal verification.

| System | Relevance | Correct. | Novelty | Specificity | Research value | Mean | Best overall |
|---|---|---|---|---|---|---|---|
| GIANTS | $3.42 \pm 0.18$ | $3.51 \pm 0.16$ | $3.10 \pm 0.19$ | $3.26 \pm 0.18$ | $3.18 \pm 0.18$ | 3.29 | 6 |
| Scientific-only | $2.65 \pm 0.17$ | $3.60 \pm 0.15$ | $1.31 \pm 0.17$ | $2.44 \pm 0.17$ | $3.39 \pm 0.17$ | 2.68 | 2 |
| GoAI | $1.79 \pm 0.16$ | $3.72 \pm 0.16$ | $3.50 \pm 0.16$ | $2.62 \pm 0.16$ | $3.10 \pm 0.17$ | 2.95 | 1 |
| Text-only | $3.58 \pm 0.15$ | $1.81 \pm 0.15$ | $1.62 \pm 0.15$ | $1.74 \pm 0.15$ | $\mathbf{4.25 \pm 0.14}$ | 2.60 | 5 |
| **COMPOSE** | $\mathbf{3.90 \pm 0.13}$ | $\mathbf{4.08 \pm 0.14}$ | $\mathbf{4.19 \pm 0.14}$ | $\mathbf{4.10 \pm 0.12}$ | $4.10 \pm 0.15$ | **4.07** | **11** |

### B.6 LLM Judge Evaluation

Table 23 reports all evaluated systems; Table 4 in the main paper shows a compact selection.

Table 23: Full LLM judge scores (1–5, higher is better).

| **Model** | **Cont.** | **Depth** | **Nov.** | **Prec.** | **Spec.** | **Ovrl.** |
|---|---|---|---|---|---|---|
| Prompt-only | 2.88 | 3.16 | 3.21 | 3.02 | 3.21 | 3.10 |
| GOAI-GPT | 2.52 | 2.69 | 3.24 | 2.90 | 3.29 | 2.93 |
| GIANTS | 1.91 | 2.75 | **3.27** | 2.53 | 3.28 | 2.75 |
| Future-Aligned | 1.74 | 2.10 | 2.86 | 2.21 | 2.99 | 2.38 |
| Text-only | 1.86 | 1.91 | 2.42 | 2.04 | 2.43 | 2.13 |
| FutureGen | 1.40 | 1.79 | 2.75 | 1.85 | 2.77 | 2.11 |
| Sci.-only | 1.59 | 1.57 | 1.71 | 1.46 | 1.71 | 1.61 |
| **COMPOSE** | **3.45** | **3.52** | 3.10 | **3.23** | **3.52** | **3.36** |

COMPOSE (DS) achieves the highest overall score ($3.36/5$), leading in Mathematical Content ($3.45$), Technical Depth ($3.52$), Precision ($3.23$), and Specificity ($3.52$). GIANTS and FutureGen score higher on Novelty because their outputs are broader and less constrained, but lower on Content and Depth because they include less concrete mathematics.

Taken together, these results suggest that graph conditioning improves both mathematical quality and grounding.

The LLM judge is based on five dimensions of mathematical quality: Mathematical Content, Technical Depth, Novelty, Precision, and Specificity.

Mathematical Content measures whether the output contains substantive mathematical statements rather than only prose. Technical Depth measures whether the claims are non-trivial and mathematically meaningful. Novelty measures whether the prediction introduces non-obvious ideas rather than simple restatements. Precision measures whether assumptions, objects, and conclusions are stated clearly and correctly. Specificity measures whether the output gives concrete mathematical details rather than vague high-level descriptions. Together, these scores measure both the mathematical content and the quality of the generated text. We report the average score for each dimension, along

with the overall mean across them. The judge receives the generated prediction together with the input paper context and assigns scores on these five dimensions. The full prompt is shown in Figure 6.

We use a GPT-4.0 version released before the benchmark period and apply the same prompt and scoring rubric uniformly to all compared models.

LLM Judge Prompt

You are a strict mathematical reviewer evaluating a generated mathematical research prediction.
**IMPORTANT:** Evaluate both the mathematical substance and the quality of the prediction. Outputs with real mathematical content should be scored higher than fluent but vague prose, but clarity, precision, and technical quality should also be taken into account.
**Input papers context:** *[input context]*
**Generated prediction:** *[generated text]*
Score each dimension (1-5):

1. **Mathematical Content:** Does the text contain actual mathematics (theorems, lemmas, equations, proofs)? Score 1 if only natural language.
2. **Technical Depth:** Are the mathematical claims non-trivial? Score 1 for surface-level statements.
3. **Novelty:** Does the prediction introduce non-obvious mathematical ideas? Score 1 for trivial restatements.
4. **Precision:** Are specific objects, conditions, and conclusions stated precisely?
5. **Specificity:** How concrete and specific are the claims (named theorems, explicit constants, specific spaces/operators)?

Figure 6: LLM judge prompt for mathematical prediction quality evaluation. The prompt explicitly instructs the judge to prioritize mathematical formalism over natural language fluency.

## C Embedding Space Analysis

Raw cosine similarity in mathematical embedding spaces is not a reliable retrieval signal in our setting. As shown in Table 26, different embedding models produce broadly similar Tgt-Neg margins, and no single model consistently outperforms the others. Across embedding models, target and unrelated future papers remain too similar to separate generation methods cleanly.

This motivates us to fine-tune a DeepSeek-Math embedding model on a training subset drawn from the 47K-paper corpus for retrieval evaluation. Before constructing training groups or pairs, we exclude all 14,677 future targets associated with the 2,033 test anchors, all test relevance relations, and all groups rooted at test anchors. The remaining subset yields 50,000 contrastive training pairs. Only after evaluator training do we use the full 47K corpus, including the held-out targets, as the retrieval candidate pool. Specifically, positive pairs consist of papers that cite the same anchor paper, while negatives are sampled from unrelated citation groups. The model is trained with a contrastive cross-entropy loss over cosine similarities scaled by temperature $\tau = 0.05$, with 5 negatives per anchor. Table 24 provides additional hyperparameter details.

Table 24: Fine-tuned embedding model hyperparameters.

| **Hyperparameter** | **Value** | **Hyperparameter** | **Value** |
|---|---|---|---|
| Base model | DeepSeek-Math-7B-Instruct | Batch size | $4 \times 4$ GPUs |
| Fine-tuning method | LoRA | Epochs | 2 |
| LoRA rank / alpha | 16 / 32 | Training pairs | 50,000 |
| LoRA dropout | 0.05 | Negatives per anchor | 5 |
| Target modules | q_proj, v_proj | Temperature ($\tau$) | 0.05 |
| Optimizer | AdamW | Max token length | 256 |
| Learning rate | $1 \times 10^{-4}$ | Pooling | Last-token, L2-normalized |
| Weight decay | 0.01 | Embedding dim | 4096 |

**Robustness to the choice of retrieval evaluator.** Our main retrieval evaluation uses an embedding model fine-tuned for this task. As a supporting robustness check, we also measure H@10 with five external embedders kept frozen, without fine-tuning on our corpus (Table 25). This tests whether COMPOSE's relative advantage also appears under alternative retrieval representations. COMPOSE ranks first under Mathstral, Llemma, and Prover, while GoAI leads under E5-Mistral and FutureGen under Kimina. Absolute H@10 is low for most of these evaluators, and the ranking changes with the embedding model. These results therefore provide qualified support beyond the task-adapted evaluator, rather than evidence of uniform superiority across evaluators.

Table 25: Supporting robustness check: H@10 under frozen external evaluators. Bold marks the best score for each evaluator.

| Evaluator | COMPOSE | GIANTS | FutureGen | GoAI | Bag-of-Papers | Text-only |
|---|---|---|---|---|---|---|
| Mathstral | **0.080** | 0.072 | 0.055 | 0.044 | 0.032 | 0.034 |
| Llemma | **0.060** | 0.052 | 0.037 | 0.048 | 0.032 | 0.021 |
| E5-Mistral | 0.025 | 0.022 | 0.028 | **0.047** | 0.021 | 0.015 |
| Prover | **0.231** | 0.114 | 0.121 | 0.185 | 0.080 | 0.052 |
| Kimina | 0.033 | 0.024 | **0.078** | 0.073 | 0.052 | 0.042 |

The paired H@10 gains over the strongest compared baseline are $+0.008$ under Mathstral (versus GIANTS; 95% CI $[+0.004, +0.018]$), $+0.008$ under Llemma (versus GIANTS; $[+0.005, +0.014]$), and $+0.046$ under Prover (versus GoAI; $[+0.029, +0.065]$).

Table 26: Raw cosine similarity to future citing papers across nine embedding models on the 2K test set. $\mu$ denotes the mean pairwise similarity among the 47K pool papers. Tgt-Neg margins are uniformly small and no method consistently dominates, motivating the use of a fine-tuned theorem embedding model for retrieval evaluation.

| Emb. Model ($\mu$) | Method | Tgt-Sim | Neg-Sim | Tgt-Neg | Med. Tgt |
|---|---|---|---|---|---|
| DeepSeek-Math-7B ($\mu$=0.44) | Full graph (ours) | 0.542 | 0.418 | 0.124 | **0.607** |
| | Text-only | **0.570** | 0.441 | **0.129** | 0.590 |
| | Bag-v1 | 0.523 | 0.399 | 0.123 | 0.589 |
| | GIANTS | 0.425 | 0.304 | 0.121 | 0.439 |
| | GOAI | 0.562 | 0.441 | 0.121 | 0.554 |
| | FutureGen | 0.465 | 0.349 | 0.115 | 0.449 |
| | Future-Aligned | 0.211 | 0.199 | 0.012 | 0.211 |
| | Prompt-only | 0.403 | 0.297 | 0.106 | 0.419 |
| E5-Mistral-7B ($\mu$=0.43) | Full graph (ours) | 0.535 | 0.412 | 0.124 | 0.652 |
| | Text-only | 0.476 | 0.359 | 0.117 | 0.377 |
| | Bag-v1 | 0.341 | 0.256 | 0.085 | 0.331 |
| | GIANTS | 0.534 | 0.388 | 0.146 | 0.659 |
| | GOAI | **0.586** | 0.440 | **0.146** | **0.725** |
| | FutureGen | 0.383 | 0.267 | 0.117 | 0.318 |
| | Future-Aligned | 0.238 | 0.221 | 0.017 | 0.216 |
| | Prompt-only | 0.529 | 0.385 | 0.145 | 0.640 |
| Kimina-7B ($\mu$=0.64) | Full graph (ours) | **0.492** | 0.479 | 0.012 | **0.487** |
| | Text-only | 0.472 | 0.457 | 0.016 | 0.481 |
| | Bag-v1 | 0.487 | 0.472 | 0.015 | 0.482 |
| | GIANTS | 0.438 | 0.417 | 0.021 | 0.441 |
| | GOAI | 0.375 | 0.347 | **0.028** | 0.367 |
| | FutureGen | 0.371 | 0.357 | 0.014 | 0.384 |
| | Future-Aligned | 0.452 | 0.436 | 0.017 | 0.452 |
| | Prompt-only | 0.418 | 0.397 | 0.021 | 0.416 |
| DeepSeek-Prover -7B ($\mu$=0.87) | Full graph (ours) | **0.873** | 0.869 | 0.003 | **0.896** |
| | Text-only | 0.832 | 0.826 | 0.006 | 0.858 |
| | Bag-v1 | 0.847 | 0.842 | 0.006 | 0.886 |
| | GIANTS | 0.812 | 0.803 | **0.009** | 0.828 |
| | GOAI | 0.797 | 0.789 | 0.007 | 0.830 |
| | FutureGen | 0.817 | 0.811 | 0.006 | 0.835 |
| | Future-Aligned | 0.814 | 0.806 | 0.008 | 0.828 |
| | Prompt-only | 0.824 | 0.816 | 0.008 | 0.843 |
| Llemma-7B ($\mu$=0.33) | Full graph (ours) | 0.338 | 0.335 | 0.002 | 0.343 |
| | Text-only | 0.333 | 0.328 | 0.005 | 0.340 |
| | Bag-v1 | 0.320 | 0.314 | 0.006 | 0.327 |
| | GIANTS | 0.301 | 0.291 | 0.010 | 0.319 |
| | GOAI | **0.345** | 0.336 | 0.010 | **0.351** |
| | FutureGen | 0.312 | 0.302 | 0.010 | 0.320 |
| | Future-Aligned | 0.310 | 0.305 | 0.006 | 0.331 |
| | Prompt-only | 0.297 | 0.287 | **0.011** | 0.314 |
| Mathstral-7B ($\mu$=0.15) | Full graph (ours) | 0.154 | 0.148 | 0.006 | 0.157 |
| | Text-only | 0.165 | 0.160 | 0.006 | 0.163 |
| | Bag-v1 | 0.150 | 0.140 | 0.010 | 0.155 |
| | GIANTS | 0.092 | 0.078 | 0.014 | 0.094 |
| | GOAI | **0.166** | 0.155 | 0.012 | **0.172** |
| | FutureGen | 0.135 | 0.122 | 0.013 | 0.135 |
| | Future-Aligned | 0.091 | 0.081 | 0.010 | 0.093 |
| | Prompt-only | 0.101 | 0.085 | **0.016** | 0.109 |

## D GNN Architecture Details

Both GNN encoders use the same architecture with separate parameters. Each encoder consists of $L = 2$ message-passing layers with hidden dimension $d = 1024$. At each layer $\ell$, incoming and outgoing neighbor messages are mean-pooled separately and summed to produce an aggregated message $\mathbf{m}_i^{(\ell)}$:

$$\mathbf{m}_i^{(\ell)} = \frac{1}{|\mathcal{N}_{\text{in}}(i)|} \sum_{j \in \mathcal{N}_{\text{in}}(i)} \mathbf{h}_j^{(\ell)} + \frac{1}{|\mathcal{N}_{\text{out}}(i)|} \sum_{j \in \mathcal{N}_{\text{out}}(i)} \mathbf{h}_j^{(\ell)}. \tag{10}$$

The message is transformed by a linear layer followed by ReLU activation:

$$\tilde{\mathbf{h}}_i^{(\ell)} = \text{ReLU}\Big(\mathbf{W}_{\text{msg}}^{(\ell)} \, \mathbf{m}_i^{(\ell)}\Big). \tag{11}$$

The node state is then updated via a learned scalar gated residual connection:

$$g_i^{(\ell)} = \sigma\Big(\mathbf{W}_{\text{gate}}^{(\ell)} \Big[\mathbf{h}_i^{(\ell)} \parallel \tilde{\mathbf{h}}_i^{(\ell)}\Big]\Big), \qquad \mathbf{h}_i^{(\ell+1)} = g_i^{(\ell)} \, \tilde{\mathbf{h}}_i^{(\ell)} + (1 - g_i^{(\ell)}) \, \mathbf{h}_i^{(\ell)}, \tag{12}$$

followed by layer normalization. After $L$ layers, a final linear projection maps each node to a structural vector of dimension 128, which is concatenated with the frozen text embedding to form the final node representation. Table 27 shows the training hyperparameters.

The learned gate balances the incoming update with the previous node representation to mitigate over-smoothing. Layer normalization is applied after each update.

Table 27: Training hyperparameters and architecture details.

| **Parameter** | **Value** |
|---|---|
| *Architecture* | |
| Encoder 1 (paper) | E5(output dim 1152) |
| Encoder 2 (theorem) | DeepSeek-Math finetuned (output dim 4096) |
| Decoder | DeepSeek-Math-7B-Instruct |
| Bridge MLP | $1152 \to 2048 \to 4224$ |
| Fused dimension | 4224 |
| Fusion cross-attention heads | 8 |
| Type embeddings | $2 \times 4224$ (paper, theorem) |
| Cross-attention layers | 8 (at indices 3,7,11,15,19,23,27,31) |
| LoRA rank | 32 |
| Contrastive projection | $4224 \to 4096$ |
| Cross-encoder projection | $1024 \to 128$ / $4096 \to 128$ |
| E5 alignment projection | $4224 \to 1024$ |
| Gen-text E5 projection | $4096 \to 1024$ |
| Gen-text thm_ft projection | $4096 \to 4096$ |
| Memory bank size | 256 |
| *Data & Training* | |
| Training samples | $\sim$80,000 |
| Batch size | 8 |
| Epochs | 30 |
| Freeze decoder epochs | 0 |
| Precision | float32 |
| GPU | NVIDIA H200 (80 GB) |
| *Optimization* | |
| Optimizer | AdamW |
| LR (bridge, fusion, type embed) | $2 \times 10^{-5}$ |
| LR (GNN) | $2 \times 10^{-6}$ |
| LR (LoRA) | $1 \times 10^{-6}$ |
| LR (cross-attn K/V) | $1 \times 10^{-5}$ |
| LR (gates) | $1 \times 10^{-5}$ |
| LR (decoder after unfreeze) | $5 \times 10^{-7}$ |
| Weight decay (GNN, bridge, cross-attn) | 0.01 |
| Weight decay (LoRA) | 0.05 |
| Weight decay (gates) | 0.0 |
| Gradient clipping (max norm) | 1.0 |
| *Loss Weights* | |
| Cross-entropy (CE) | 1.0 |
| InfoNCE contrastive ($\lambda_{\text{contr}}$) | 0.30 |
| Cross-encoder contrastive ($\lambda_{\text{xcontr}}$) | 0.20 |
| Graph margin loss | 0.5 (margin $= 0.2$) |
| alignment | 0.5 |
| Link prediction (freeze phase) | 0.4 |
| $\lambda_{alignment}$ (freeze phase) | 1.0 |
| InfoNCE temperature ($\tau$) | 0.07 |

## E Prompt Templates

We show the exact input and a real generated output for each method on the same example subgraph. for the external baseline we use there code.

For all the external baseline we use the same prompt as mation in ther papre.

COMPOSE (Ours)

**Text input (decoder self-attention):**

```
[INST] Title: Discrete Polymatroids
Using the provided paper context, predict the paper's main mathematical
result. [/INST]
```

**Graph inputs (cross-attention, not text):**

$G_s$: 8 citation subgraph node embeddings $\xrightarrow{\text{GNN}}$ K/V injected at decoder layers 7, 15, 23, 31

$G_f$: 10 Mathlib formal subgraph node embeddings $\xrightarrow{\text{GNN}}$ K/V injected at decoder layers 7, 15, 23, 31

**Generated output:**

In a paper proving that any non-increasing function of non-negative integers is the normalized depth function of some squarefree monomial ideal, a key step is: **Proposition 2.4**: Let $S$ be a simplicial complex and let $K \subset J(S)$ be a vertex $k$-coloring. Then colorings with no monochromatic components can be obtained by applying one of the following operations iteratively to any coloring $C$ whose complement graph $G_C$ is connected.

Figure 7: Full Graph (Ours): the decoder receives only the anchor paper title as text; the citation and formal graphs are injected silently via cross-attention.

Text-Only (LoRA)

**Input:**

```
[INST] You are a mathematician. Given a paper and related work, predict
the paper's main theorem with some lemma use.

Paper: Rational Powers
Abstract: The discrete polymatroid is a multiset analogue of the matroid...

Related papers:
- The Depth of Powers of an Ideal: We study the limit and initial behavior
  of the numerical function f(k)=depth S/I^k...
- How to Compute the Stanley Depth of a Monomial Ideal: Let $J\subsetI$ be
  monomial ideals. We show that the Stanley depth of I/J can be computed...
- The Stable Set of Associated Prime Ideals of a Polymatroidal Ideal: ...

Related Mathlib theorems:
- adjoin_monomial_eq_reesAlgebra: Algebra.adjoin R (Submodule.map...)
[/INST]
```

**Generated output:**

In a paper proving that any non-increasing function of non-negative integers is the normalized depth function of some squarefree monomial ideal, a key step is: **Corollary 2.5**: With the assumptions and notation of the previous proposition, it follows that $g_J$ is a non-increasing function if $g_I$ is non-increasing.

Figure 8: Text-Only (LoRA): DeepSeek-Math-7B fine-tuned with LoRA on flat text. Up to 4 neighbor abstracts and 3 Mathlib theorems are included; no graph structure.

Prompt-Only

**Input:** Same format as Text-Only (LoRA) above.

```
[INST] You are a mathematician. Given a paper and related work, predict
the paper's main theorem with some lemma use.

Paper: Rational Powers
Abstract: The discrete polymatroid is a multiset analogue of the matroid...

Related papers:
- The Depth of Powers of an Ideal: ...
- How to Compute the Stanley Depth of a Monomial Ideal: ...
- The Stable Set of Associated Prime Ideals of a Polymatroidal Ideal: ...

Related Mathlib theorems:
- adjoin_monomial_eq_reesAlgebra: ...
[/INST]
```

**Generated output:**

Based on the provided abstract and related papers, the main theorem seems to be a multi-part theorem involving the depth, symbolic depth, and symbolic analytic spread of symbolic powers of a squarefree monomial ideal. **Main Theorem:** Let $\mathbb{K}$ be a field, $S = \mathbb{K}[x_1, \ldots, x_n]$, and $I \subset S$ a squarefree monomial ideal. Then for all integers $k \geq 1 \ldots$

Figure 9: Prompt-Only: identical prompt to Text-Only but uses base DeepSeek-Math-7B-Instruct with no fine-tuning. Output is more verbose and hedged.

GoAI

**Input:**

```
User: You are given a citation graph in the area of: Rational Powers

=== Citation Graph --- Paper Nodes ===
P1: "Discrete Polymatroids"
   Summary: The discrete polymatroid is a multiset analogue of the matroid...
P2: "The Depth of Powers of an Ideal"
   Summary: We study the numerical function f(k)=depth S/I^k...
P3: "The Stable Set of Associated Prime Ideals of a Polymatroidal Ideal"
   Summary: The associated prime ideals of powers of polymatroidal ideals...
... (up to 8 nodes)

=== Citation Graph --- Edges ===
P1 -> P2  (cites: "depth of powers of ideals")
P1 -> P3  (cites: "associated primes of polymatroidal ideals")
... (up to 12 edges)

Predict one specific novel theorem. State only the mathematical claim.

Assistant: The next theorem this research direction will produce is:
```

**Generated output:**

Let $I$ be a squarefree monomial ideal in the polynomial ring $S = \mathbb{K}[x_1, \ldots, x_n]$ over a field $\mathbb{K}$. If $I$ is polymatroidal, then the Stanley depth of $I^k$ is equal to the depth of $I^k$ for all $k \geq 1$.

Figure 10: GoAI: the citation graph topology (nodes and typed edges) is serialized as text. Up to 8 paper nodes and 12 citation edges are included.

GIANTS

**Input:**

```
Paper 1 summary: The discrete polymatroid is a multiset analogue of the
matroid. Based on the polyhedral theory on integral polymatroids developed
in late 1960's and in early 1970's, the combinatorics and algebra on
discrete polymatroids will be studied.

Paper 2 summary: We study the limit and initial behavior of the numerical
function f(k)=depth S/I^k. General properties of this function together
with concrete examples are given.

Based on these two foundational papers, what is the key insight or main
contribution of the downstream paper?
```

**Generated output:**

The key insight lies in **bridging the algebraic and combinatorial structures** established in Papers 1 and 2. By leveraging the discrete polymatroid framework, the downstream paper analyzes the depth function $f(k) = \operatorname{depth} S/I^k$ through the lens of combinatorial properties of polymatroids, establishing that the depth function of polymatroidal ideals is non-increasing and eventually stabilizes.

Figure 11: GIANTS (Qwen3-4B fine-tuned on insight anticipation): receives only the two most relevant citation neighbor abstracts. Output is a conceptual description rather than a formal theorem.

# F Example Outputs

**Example 1 – Bessel Function Moments and Feynman Integrals in QFT** (arXiv:0801.0891)

**Ground Truth Future Paper:**
*We record what is known about the closed forms for various Bessel function moments arising in quantum field theory, condensed matter theory and other parts of mathematical physics. More generally, we develop formulae for integrals of products of six or fewer Bessel functions. In consequence, we discover and prove closed forms for* $c_{n,k} := \int_0^\infty t^k K_0^n(t)\,\mathrm{d}t$ *with integers* $n = 1, 2, 3, 4$ *and* $k \geq 0$*, obtaining new results for the even moments* $c_{3,2k}$ *and* $c_{4,2k}$*...*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *A new method is introduced for computing all L-loop Feynman integrals associated to any graph in* $\phi_d^4$ *theory without first computing the Symanzik polynomial. The approach relies on solving systems of linear PDE relations between classes of hypergeometric functions arising from period matrices over moduli spaces parametrizing degenerations at infinity...* | **0.827** | **1** |
| GoAI | *Let* $X$ *be a Calabi-Yau threefold with Enriques fibers, and let* $D$ *be a divisor on* $X$*. Then the Gromov-Witten invariants of* $X$ *with respect to* $D$ *can be computed using mirror symmetry.* | 0.910 | 0 |
| GIANTS | *The key insight is the formalization of mirror symmetry through the interplay between the A-model connection and differential operators of a spectral curve...* | 0.260 | 0 |
| Text-only (LoRA) | *This paper is an introduction to classical polylogarithms... Topics: monodromy, polylogarithm local systems, Bloch's constructions of regulators...* | 0.271 | 0 |
| Paper-graph-only | *Key steps to obtaining a convenient and quick proof... [garbled output]* | 0.519 | 0 |
| FutureGen | *The next key contribution is likely a new theorem related to the dimensional regularisation parameter and GKZ hypergeometric systems...* | 0.461 | 0 |

**Analysis.** COMPOSE is the only model to retrieve the target paper, ranking 1 in the 47K pool. Every baseline produces text in mathematically neighbouring but wrong areas: GoAI and GIANTS generate Calabi-Yau geometry and mirror symmetry; Text-only drifts to polylogarithms and motives; FutureGen proposes GKZ systems. Notably, GoAI achieves a higher raw similarity (0.910 vs. 0.827) yet misses at rank 27 Calabi-Yau geometry and QFT integrals occupy different regions of embedding space despite sharing surface-level mathematical vocabulary. COMPOSE alone generates text about $L$-loop Feynman integrals and hypergeometric functions in $\phi_d^4$ theory, placing the output in the correct QFT-integral embedding cluster. The citation graph draws in a chain of papers on Feynman integral evaluation and dimensional regularisation; the formal graph, in turn, contributes Lean lemmas on hypergeometric identities and period integrals that anchor the model to the special-function structure of the actual result.

Figure 12: Win on arXiv:0801.0891 (QFT integrals). COMPOSE ranks 1 as the only model in the right area; GoAI has higher raw similarity (0.910) yet misses at rank 27; all five baselines fail.

**Example 2 – Koopman Operator Convergence** (arXiv:2206.13659)

**Ground Truth Future Paper:**
*We develop an algebraic framework for sequential data assimilation of partially observed dynamical systems. In this framework, Bayesian data assimilation is embedded in a non-abelian operator algebra... the forecast step of data assimilation is represented by a quantum operation induced by the Koopman operator of the dynamical system.*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *The authors prove a convergence theorem for extended dynamic mode decomposition applied with finite-rank approximations and show how this relates to approximating eigenfunctions and eigenvalues associated with dynamical systems governed by ergodic measures, using empirical data samples only...* | **0.934** | **1** |
| Text-only (LoRA) | *We prove that the eigenvalues of the matrix $K_N^M$ approximate the eigenvalues of $K_N^M$ when $M \to \infty$.* | 0.945 | 0 |
| Prompt-only | *Based on the provided abstract and related papers, it seems that the main focus of the paper is on using Koopman operator theory to classify and compare iterative algorithms...* | 0.708 | 0 |
| GIANTS | *The key insight... is the rigorous mathematical foundation for the convergence of Extended Dynamic Mode Decomposition (EDMD) to the Koopman operator...* | 0.546 | 0 |
| GoAI | *Let $K$ be the Koopman operator acting on the space of observables $L^2(\mu)$, where $\mu$ is an ergodic measure. Then, the eigenvalues of $K$ are contained in the unit circle...* | 0.402 | 0 |
| ResearchAgent-GPT | *Proposed Research: Inexact iterative numerical linear algebra for ergodic dynamical systems...* | 0.584 | 0 |
| FutureGen | *The next key contribution... is likely to be the development of new algorithms and techniques for solving large-scale linear systems...* | 0.602 | 0 |
| Paper-graph-only | *Key and a key is that to other... [garbled output]* | 0.586 | 0 |

**Analysis.** COMPOSE retrieves at rank 1 by generating a full description of EDMD convergence for ergodic dynamical systems finite-rank approximations, eigenfunction convergence, empirical data samples directly matching the Koopman operator framework of the ground truth. Text-only achieves the highest similarity (0.945) with a narrow eigenvalue statement, yet fails at rank 34; its output is too terse to land in the same cluster as papers on operator-theoretic data assimilation. All remaining baselines miss.

Figure 13: Qualitative comparison on arXiv:2206.13659. COMPOSE is the only system combining high similarity with successful retrieval; Text-only achieves higher sim but fails to retrieve.

**Example 3 – Distributed Nash Equilibrium** (arXiv:1212.6235)

**Ground Truth Future Paper:**
*Noncooperative game-theoretic tools have been increasingly used to study many important resource allocation problems in communications, networking, smart grids, and portfolio optimization. In this paper, we consider a general class of convex Nash Equilibrium Problems (NEPs), where each player aims to solve an arbitrary smooth convex optimization problem.*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *We propose an algorithm that solves monotone game problems in which players have different local objectives but share some common affine inequalities as well as their individual ones (e.g., capacity limits). This allows us to solve more general classes of optimization and game models than existing literature on NE computation... convergence properties under mild assumptions using variational analysis combined with Lyapunov arguments...* | **0.910** | **1** |
| GoAI | *In an $N$-coalition non-cooperative game, if the players in the same coalition cooperatively minimize the sum of...* | 0.910 | 1 |
| Text-only (LoRA) | *We prove that the game admits a unique Nash equilibrium, and that this equilibrium is the solution of a certain convex...* | 0.964 | 0 |
| GIANTS | *The key insight... is the integration of game-theoretic equilibrium problems (NEPs) with structured monotone...* | 0.485 | 0 |
| ResearchAgent-GPT | *Proposed Research: Distributed Computation of Generalized Nash Equilibrium with Partial-Decision Information...* | 0.478 | 0 |
| FutureGen | *The next key contribution... is likely to be the development of a distributed algorithm...* | 0.547 | 0 |

**Analysis.** COMPOSE identifies the research direction as distributed NE computation with coupling constraints, generating text about "monotone game problems," "NE computation," and "variational analysis" (rank 2). GoAI ties on similarity and retrieves at rank 9. Text-only achieves the highest embedding similarity (0.964) with a concise Nash equilibrium statement, yet fails at rank 26 its output is too narrow to land in the same cluster as papers on distributed NE computation with shared affine constraints.

Figure 14: Qualitative comparison on arXiv:1212.6235. Text-only achieves the highest sim yet fails retrieval; COMPOSE and GoAI both retrieve, with COMPOSE ranking higher.

**Example 4 Glauber Dynamics and Mean-Field Spin Systems** (arXiv:0812.2265)

**Ground Truth Future Paper:**
*We prove that in the high temperature regime the mixing time of the Glauber dynamics is $\Theta(n^2 \log n)$, where $n$ is the number of vertices in the graph; in contrast, we show that in the low temperature regime the mixing time is exponentially large.*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *The authors prove an upper bound on replica symmetry breaking for mean field spin models using the Guerra interpolation method combined with information-theoretic arguments, establishing a unified framework applicable to Ising and Potts models across the high- and low-temperature regimes...* | **0.412** | **1** |
| FutureGen | *In this paper, we consider a random graph on which topological restrictions are imposed, such as constraints on the total number of edges, wedges, and triangles. We prove new results on the structure of such graphs...* | 0.561 | 1 |
| GoAI | *Let $G$ be a simple graph with $n$ vertices and $m$ edges. Let $d(G)$ be the maximum degree of $G$. Then, the chromatic polynomial $P(G, k)$ satisfies...* | 0.372 | 0 |
| GIANTS | *The key insight is a rigorous analysis of the Glauber dynamics on random graphs using the cavity method...* | 0.318 | 0 |
| Text-only (LoRA) | *We compute the relative entropy of the two Gibbs ensembles in the limit as $n$ grows, establishing sharp bounds on the mixing time via coupling arguments...* | 0.441 | 0 |
| Paper-graph-only | *Key steps: Lemma 3: We prove that in the limit... [garbled output]* | 0.389 | 0 |

**Analysis.** COMPOSE correctly places the prediction in the statistical physics of spin systems, generating text about replica symmetry breaking for mean-field models using the Guerra interpolation technique precisely the analytical framework underlying Glauber dynamics at phase transitions. FutureGen is the only baseline to retrieve, but its generated text concerns random graphs with edge, wedge, and triangle constraints, a combinatorics problem entirely unrelated to spin system dynamics. GoAI generates a statement about chromatic polynomials again graph combinatorics, not statistical physics. GIANTS and Text-only miss entirely.

Figure 15: Win on arXiv:0812.2265 (statistical physics). COMPOSE generates content in the correct spin-system area; FutureGen retrieves but with text about graph combinatorics a completely different field; four baselines fail.

**Example 5 – Stability of Shear Flows in the 2D Euler Equation** (arXiv:2007.11547)

**Ground Truth Future Paper:**
*We prove that the only stationary states near Couette flow must indeed be shears, even in relatively low regularity $H^3$.*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *The authors prove that there exists a constant $\lambda_0 > 1/4$ such that if $\|\Omega\| \geq \frac{8\pi}{3}\sqrt{\log N}$, all eigenvalues of the linearised operator have real part $\geq -\Lambda$. Moreover, they derive a pseudospectral estimate bounding any resolvent norm by $(-\Re(\zeta)+\delta)^{-p}$ with $p = \mathcal{O}(\epsilon^{-q})$, and establish exponential decay rates towards shear equilibria…* | **0.603** | **1** |
| GoAI | *Theorem: Let $u$ be a solution to the Navier-Stokes equations in $\mathbb{R}^3$ with initial data $u_0 \in H^s(\mathbb{R}^3)$, $s > 3$. Then the solution $u$ is unique.* | 0.923 | 1 |
| GIANTS | *The key insight lies in the integration of fluid dynamics analysis with operator-theoretic techniques for PDEs with low-regularity coefficients…* | 0.385 | 0 |
| Text-only (LoRA) | *In a paper proving that the only stationary states near them must indeed be shears, a key step is Proposition 2.3: The functions $\Psi_\varepsilon$ constructed in Proposition 2.1 are in fact analytic…* | 0.658 | 0 |
| Paper-graph-only | *Lemma 10: The only stationary states must be annihilated ones with energy $H = -AJ + RI$ … [garbled output]* | 0.564 | 0 |
| FutureGen | *The next key contribution in the research topic of vortex filament solutions is likely to be a new theorem related to the stability and interaction of vortices…* | 0.476 | 0 |

**Analysis.** COMPOSE generates specific spectral stability machinery for the linearised 2D Euler equation near Couette flow: eigenvalue bounds, pseudospectral resolvent estimates, and exponential decay towards shear equilibria, directly matching the mathematical framework of the ground truth. GoAI also retrieves but produces a *uniqueness* theorem for the Navier-Stokes equations in $\mathbb{R}^3$ the wrong equation (Euler vs. Navier-Stokes), the wrong dimension (2D vs. 3D), and the wrong property (spectral stability vs. uniqueness). The remaining four baselines all fail.

Figure 16: Win on arXiv:2007.11547 (fluid dynamics). COMPOSE generates specific 2D Euler spectral stability content; GoAI retrieves but with a Navier-Stokes uniqueness theorem wrong equation, wrong dimension, wrong property; four baselines fail entirely.

**Example 6 – Cluster Algebras for $\mathrm{SL}(k)$-Local Systems and Quantum Group Homology** (arXiv:2107.13069)

**Ground Truth Future Paper:**
*Fock and Goncharov introduced a family of cluster algebras associated with the moduli of $\mathrm{SL}(k)$-local systems on a marked surface with extra decorations at marked points. We study this family from an algebraic perspective...*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *Our goal is constructing an $sl_N$ knot homology theory for coloured knots and links with any partition $\lambda$. This gives new invariants stronger than those from quantum groups at roots of unity or finite-type homologies such as Khovanov–Rozansky theories, with constructions that are functorial and compatible with tensor products...* | **0.707** | **1** |
| GoAI | *The universal $sl_3$-link homology, when taken with complex coefficients, categorifies the quantum $sl_3$-link invariant.* | 0.936 | 1 |
| GIANTS | *The key insight is proving the invariance of an algebraic structure under certain categorical operations...* | 0.541 | 0 |
| Text-only (LoRA) | *In a paper proving that we classify irreducible $A$-modules and show that $A$-lfdmod is an upper finite module category...* | 0.784 | 1 |
| Paper-graph-only | *Key steps are to determine that there has been an important area in which... [garbled output]* | 0.466 | 0 |
| FutureGen | *The next key contribution is likely to be a new theorem related to the categorification of quantum groups at roots of unity...* | 0.669 | 1 |

**Analysis.** COMPOSE ranks 3, ahead of all other retrieving models: GoAI rank 8, Text-only rank 10, FutureGen rank 14. GIANTS and Paper-graph-only miss entirely. COMPOSE generates a precise $sl_N$ statement knot homology, Khovanov-Rozansky theories, partition-coloured links that directly targets the $\mathrm{SL}(k)$ representation-theory cluster to which the ground truth belongs. GoAI generates an $sl_3$ categorification result (related but less general), and Text-only produces module-category language (in the right neighbourhood but less specific). The citation graph supplies papers on cluster algebras and moduli of local systems; the formal Lean graph contributes category-theoretic and representation-theoretic theorems that ground the model's $sl_N$ framing, resulting in the highest-ranked generation despite a lower raw similarity score than GoAI.

Figure 17: Win on arXiv:2107.13069 (representation theory). COMPOSE generates the most specific $sl_N$ content and ranks 3, ahead of GoAI (rank 8), Text-only (rank 10), and FutureGen (rank 14).

**Example 7 – Failure Case: Schubert Calculus on the Permutahedral Variety** (arXiv:2005.12194)

**Ground Truth Future Paper:**
*We compute the expansion of the cohomology class of the permutahedral variety in the basis of Schubert classes.*

| Model | Generated Output | Sim | H@10 |
|---|---|---|---|
| **COMPOSE (ours)** | *The goal is to determine when a connected simple graph $G$ with $n \geq 2$ vertices contains a subgraph $H$ as an induced subgraph sharing at least $k$ common edges, where $k \leq \min\{d(H), d-3\}$ and $d$ is half the minimum degree. This problem is approached via Ramsey theory...* | 0.438 | 0 |
| GoAI | *Let $G$ be a graph and let $X_G$ be its chromatic symmetric function. Then $X_G$ is $e$-positive.* | **0.967** | **1** |
| GIANTS | *The key insight is the explicit computation of an $e$-positive formula for the chromatic symmetric function of lollipop graphs, together with infinite families of $e$-positive symmetric function bases...* | 0.385 | 1 |
| Text-only (LoRA) | *In a paper proving that we compute the expansion of the cohomology class of the permutahedral variety in the basis of Schubert classes, a key step is Theorem 1.3: For $w \in LP_n$, $a_w = S_w(1, \ldots, 1)$.* | 0.915 | 1 |
| Paper-graph-only | *Theorem 1.3: For $w \in LP_n$, a key step is to compute the cohomology expansion... [garbled notation]* | 0.794 | 1 |
| FutureGen | *We propose a new approach to equivariant cohomology theories called equivariant factorization homology, generalising classical factorization homology...* | 0.748 | 0 |

**Analysis.** This failure exposes a systematic weakness when the anchor topic sits in an area with sparse Mathlib coverage. Schubert calculus and the combinatorics of permutahedral varieties have essentially no direct representation in Lean/Mathlib, which instead has rich coverage of finite graph theory. The informal-to-formal alignment therefore maps the anchor's theorems to graph-theoretic Lean lemmas, causing COMPOSE to generate content about induced subgraphs and Ramsey theory completely wrong for Schubert classes and cohomology rings. GoAI correctly identifies the $e$-positivity direction of algebraic combinatorics from citation context alone, and Text-only copies the target's exact framing. This failure mode is concentrated in algebraic combinatorics, parts of mathematical physics, and other areas underrepresented in Mathlib, and points to the need for a confidence-aware gate that downweights the formal graph when alignment quality falls below a coverage threshold.

Figure 18: Failure on arXiv:2005.12194 (Schubert calculus). Mathlib lacks Schubert calculus coverage, so the formal graph pulls COMPOSE to graph theory; GoAI and Text-only correctly identify the algebraic combinatorics direction.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions of the paper, including the grounded future mathematical generation task, the COMPOSE dual-graph framework, the new paired scientific-formal dataset, and the temporal benchmark. These claims are supported by the experimental results and the evaluation and qualitative analysis

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the dedicated Limitations section in the conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results in the form of formal theorems with proofs. Its contributions are methodological, empirical, and dataset-oriented.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the dataset construction, temporal split, model architecture, training stages, baselines, evaluation metrics, and key hyperparameters in the main paper and appendix. These details provide the information needed to understand and reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The camera-ready paper provides the public project page, which links the released code and project resources. The main paper and appendix document the dataset construction, model, evaluation protocol, and training details.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the training and test splits, temporal benchmark construction, model variants, evaluation metrics, and key training hyperparameters in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix B.4 reports means and standard deviations across three independently trained models per configuration, and the formal-graph diagnostics report paired 95% confidence intervals.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the main compute setting used for training, including the hardware type, memory, precision, batch size, and key training hyperparameters. These details provide the basic compute information needed to understand and reproduce the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the research presented in this paper conforms to the NeurIPS Code of Ethics. The work studies mathematical prediction and generation from scientific and formal corpora and includes a blinded evaluation by mathematics graduate students.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a dedicated discussion of broader societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk deployed system or a public-facing generative service, and it does not involve sensitive personal data. The work is a research prototype for mathematical prediction and evaluation.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper uses existing external assets, including scientific and formal mathematical corpora, but does not yet provide a dedicated summary of their licenses and terms of use in the submission.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces new assets, including a paired scientific-formal dataset and a temporal benchmark, and documents their construction, scope, and intended release in the main paper and appendix. The dataset and related resources are planned for release after the review process.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: The paper reports a blinded evaluation by 30 mathematics graduate students, but the current manuscript does not include the complete participant instructions or compensation details.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The paper reports a blinded evaluation by mathematics graduate students, but the current manuscript does not document an IRB or equivalent ethics determination, participant-risk disclosure, or consent procedure.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core component of the method and evaluation. The paper describes the use of math-specialized language model decoders for generation, LLM-based judging for evaluation, and LLM-assisted components in the alignment pipeline.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.