Retrieval-Augmented Language Models are Mimetic Theorem Provers

Wenjie Yang[1*], Ruiyuan Huang[1], Jiaxing Guo[1*], Zicheng Lyu[1], Tongshan Xu[1], Shengzhong Zhang[1], Lun Du[2], Da Zheng [2†], Zengfeng Huang[1,3†]

[1]Fudan University, [2]Ant Group, [3]Shanghai Innovation Institute

Abstract

Large language models have shown impressive capabilities in various mathematical tasks, yet they often struggle with rigorous, proof-based reasoning required for research-level mathematics. Retrieval-augmented generation presents a promising path to address this limitation. This paper systematically explores RAG for natural language theorem proving. We reveal that LLMs, when augmented with retrieved proofs, can function as powerful *mimetic* theorem provers: models can effectively generalize proof techniques from unstructured retrieved contexts to construct correct proofs for novel theorems. Building upon this insight, we introduce Dual RAG, a simple yet effective RAG framework that retrieves both relevant theorems and proof techniques. Dual RAG uses LLMs to identify the underlying reasoning challenges posed by theorems, enhancing both queries and document contexts to improve retrieval quality. Our experiments show that Dual RAG achieves substantial improvements in retrieval performance, with gains of up to 34.19%. Expert evaluations further confirm that these retrieval enhancements directly translate into higher quality proof generation. Notably, when integrated with the arXiv API, Dual RAG demonstrates the ability to prove research-level theorems in theoretical machine learning, highlighting its strong potential as a foundational element of a practical mathematical copilot.

1 Introduction

Recent large language models (LLMs), such as OpenAI-o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025), have demonstrated strong performance on a variety of mathematical benchmarks. However, their capabilities in rigorous, proof-based reasoning remain underexplored. Early studies have revealed substantial limitations of LLMs in tackling graduate-level theorem proving (Frieder et al., 2023), and more recent works show that even the most advanced reasoning mod-

Target theorem: Let X_n be independent Poisson r.v.'s with $EX_n = \lambda_n$, and let $S_n = X_1 + \cdots + X_n$. Show that if $\sum \lambda_n = \infty$ then $S_n/ES_n \to 1$ a.s.
DeepSeek-R1 w/o RAG: ... The series $\sum_{n=1}^{\infty} e^{-c\epsilon^2\mu_n}$ converges because: $\mu_n \geq \mu_{n-1} + \lambda_n \geq \mu_{n-1}$, so μ_n is non-decreasing. For large n, $\mu_n \geq Cn$ for some $C > 0$ (if λ_n does not decay too rapidly). Even if μ_n grows sublinearly, $e^{-c\epsilon^2\mu_n}$ is still summable... **Expert evaluation:** Wrong
Retrieved documents: [Document 3] Theorem 2.3.9. If $A_1, A_2, \ldots$ Proof: ... To get almost sure convergence, we have to take subsequences...
DeepSeek-R1 w/ RAG: To prove that $S_n/ES_n \to 1$ almost surely when $S_n = X_1 + \cdots + X_n$ (with independent Poisson X_i such that $\sum \lambda_n = \infty$), we adapt the strategy from Theorem 2.3.9 in Document 3... **Step 2: Subsequence for Almost Sure Convergence**... **Expert evaluation:** Correct

Table 1: A demonstration of the mimetic theorem proving capability of retrieval-augmented language models. When addressing the example theorem, the raw LLM fails, whereas the RAG variant successfully adopts the subsequence-based approach from Theorem 2.3.9 in a retrieved document to achieve a correct proof. The proof technique in the document is underlined. Critical steps of the mimetic proof are shaded.

*Work done during internship at Ant Group

†Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21301–21313
November 4-9, 2025

els struggle with competitions such as the USAMO (Petrov et al., 2025). Despite these setbacks, the demand for a reliable "mathematical copilot" is growing ever more urgent (Frieder et al., 2024).

Retrieval-augmented generation (RAG) is a natural direction for developing a mathematical copilot, as it has proven effective in other mathematical reasoning tasks (Yang et al., 2025a). Several key challenges must be carefully addressed when designing such a RAG system: (1) What to retrieve? Prior work has primarily focused on retrieving premise theorems (Welleck et al., 2021), but are there other types of information that could be beneficial for theorem proving? (2) How to retrieve? While it is straightforward to build a standard dense retrieval system by following approaches from other tasks (Lewis et al., 2020), an open question remains: can domain-specific knowledge improve retrieval effectiveness in the context of a mathematical copilot?

In this paper, we thoroughly investigate the role of RAG in natural language theorem proving, with the goal of building a practical mathematical copilot. We find that retrieval-augmented language models act as strong *mimetic* theorem provers: they can follow proof techniques found in unstructured contexts and effectively apply them to construct correct proofs for new theorems. An example is shown in Table 1. Without RAG, DeepSeek-R1 fails to generate a valid proof. However, when provided with a context containing a relevant proof technique, the LLM is able to complete the proof, even when the retrieved and target theorems are not exactly the same. This finding suggests that, in addition to theorems themselves, proofs can also be valuable retrieval targets, as they often contain techniques that generalize to new problems. To leverage this insight, we propose a simple yet effective system called Dual RAG that retrieves both relevant theorems and proof techniques, designed specifically for natural language theorem proving. Unlike standard RAG approaches, Dual RAG employs a dual augmentation mechanism that rewrites both the target theorem and the context, ensuring they are better aligned in the embedding space. This dual augmentation process enhances retrieval quality by bridging the gap between semantically dissimilar theorems that rely on similar proof strategies. The key idea behind our approach is grounded in domain knowledge: theorems that rely on similar proof strategies may not be semantically similar, but they tend to present similar challenges.

To assess the performance of Dual RAG, we introduce a new dataset, Exercise100, curated from graduate-level mathematics textbooks. While comparable in difficulty to the GHOSTS dataset (Frieder et al., 2023), our dataset includes a manually annotated gold context for each exercise. This allows for direct evaluation of retrieval quality and its downstream impact on theorem proving. Dual RAG achieves notable improvements in retrieval accuracy, with gains of up to 34.19%. Expert evaluations further validate that these improvements in retrieval lead to corresponding gains in proof generation quality. When integrated with the arXiv API for academic search, Dual RAG is capable of proving research-level theorems in theoretical machine learning, highlighting its potential as a foundation for a practical mathematical copilot. Our main contributions are as follows:

- We systematically study RAG for natural language theorem proving, with the goal of building a practical mathematical copilot. We show that language models equipped with both retrieved theorems and proofs, can act as mimetic theorem provers, effectively generalizing proof techniques from retrieved examples to novel problems.
- With the empirical insight, we propose Dual RAG, a simple yet effective RAG framework guided by domain knowledge. It identifies reasoning challenges using LLMs, and augments both target theorems and context to improve retrieval quality.
- We introduce a new dataset, Exercise100, from graduate-level math textbooks, with manually annotated gold contexts. This enables retrieval-level evaluation, where Dual RAG shows significant improvements in both retrieval and generation performance.
- Integrated with the arXiv API, Dual RAG can prove research-level theorems in theoretical machine learning, highlighting its potential for real-world mathematical assistance.

2 Related Works

LLMs for theorem proving. Automated theorem proving, a cornerstone of mathematical education and research, has long been a significant challenge in computer science (Robinson and Voronkov, 2001; Kovács and Voronkov, 2013). Recent advances in LLMs have introduced new possibilities

for mathematical reasoning, revitalizing research in this domain (Li et al., 2024). On benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), which only assess the final answers, LLMs have demonstrated remarkable performance. Substantial progress has also been made in theorem proving using formal languages like Lean, with numerous studies exploring LLM-based approaches (Zheng et al., 2021; Yang et al., 2023; Wang et al., 2024a; Xin et al., 2025; Hu et al., 2025). However, natural language theorem proving remains relatively understudied, primarily due to the challenges in evaluation (Yang et al., 2025b). Even state-of-the-art LLMs struggle with this task: GPT-4 failed to pass a graduate-level mathematics exam (Frieder et al., 2023), and recent models like o3-mini exhibit significant difficulties on advanced benchmarks such as the USAMO (Petrov et al., 2025). At present, the demand for *mathematical copilots* (i.e., agents capable of assisting in real-world education and research) has become increasingly urgent (Frieder et al., 2024). Our work, which reveals the mimetic capabilities of retrieval-augmented language models, represents a meaningful step toward this goal.

Premise selection. A related but distinct task is premise selection, which focuses on retrieving relevant lemmas from a large, structured library of proven theorems. Unlike our setting where retrieval operates over raw and unstructured mathematical contexts, premise selection addresses a more simplified and idealized retrieval scenario. As a result, it has been extensively studied in the literature (Kucik and Korovin, 2018; Piotrowski and Urban, 2020; Welleck et al., 2021, 2022). For a comprehensive overview of advances in this field, we refer readers to recent surveys such as (Li et al., 2024).

Retrieval-augmented generation. Recent advances in RAG have demonstrated significant effectiveness in enhancing the response quality of LLMs while reducing hallucination issues (Wang et al., 2024b; Zhao et al., 2024). Substantial research efforts have been devoted to optimizing RAG frameworks, particularly through innovations in three key components, such as document chunking (LangChain, 2024; Chen et al., 2024), embedding techniques (Zhang et al., 2023; Chen et al., 2023), and reranking methods (Sun et al., 2023). RAG systems have also been applied to mathematical question-answering, to improve the middle-school education (Henkel et al., 2024).

3 Problem Statement

In this paper, we study the task of natural language theorem proving with *unstructured* context. Formally, given a target theorem[1] $\mathcal{T}$, the goal is to generate a valid proof by leveraging a collection of potentially relevant documents $\mathcal{D} = \{D_1, D_2, \cdots, D_N\}$, where each D_i may be a theorem-related textbook or research paper. To construct the proof, the system must first retrieve a set of chunks $\mathcal{C}_\mathcal{T} = \{C_1, C_2, ..., C_n\}$ from these documents, where each chunk $C_i \subseteq D_j$ for some $j \in [N]$ is automatically segmented by the RAG system. Our setting differs from the classic premise selection task (Kühlwein et al., 2012; Irving et al., 2016; Ferreira and Freitas, 2020; Welleck et al., 2021) in important ways. While typical premise selection works with clean, pre-separated mathematical statements (usually from structured theorem libraries), we aim to handle raw documents where mathematical content appears alongside explanations, examples, and discussion. This better reflects how mathematicians actually work - they usually develop proofs by working with complete papers or books, instead of isolated formal statements.

4 Methodology

To optimize retrieval performance in natural language theorem proving with unstructured data, we introduce Dual RAG, a simple yet effective RAG framework. Our method employs dual data augmentation to align target theorems with their relevant context in the embedding space. We meticulously develop a complete retrieval pipeline for our task, from chunking to reranking. We systematically introduce our framework in this section.

Chunking. Segmenting unstructured documents effectively is a cornerstone for retrieval (Chen et al., 2024; Wang et al., 2024b), especially in theoretical texts where maintaining logical flow is essential. In our task, the documents include theorems, their proofs, along with examples and remarks. These examples show how to apply the known theorems, while the remarks provide intuitive explanations, both of which are extremely helpful when developing new proofs. To ensure no critical information is lost, we utilize an LLM-based chunking method. Specifically, we instruct the LLM to divide the

[1]Throughout this paper, we refer to the theorem or conjecture that needs to be proved as the "target theorem" or "query," and the theorems contained in the documents as "known theorems."

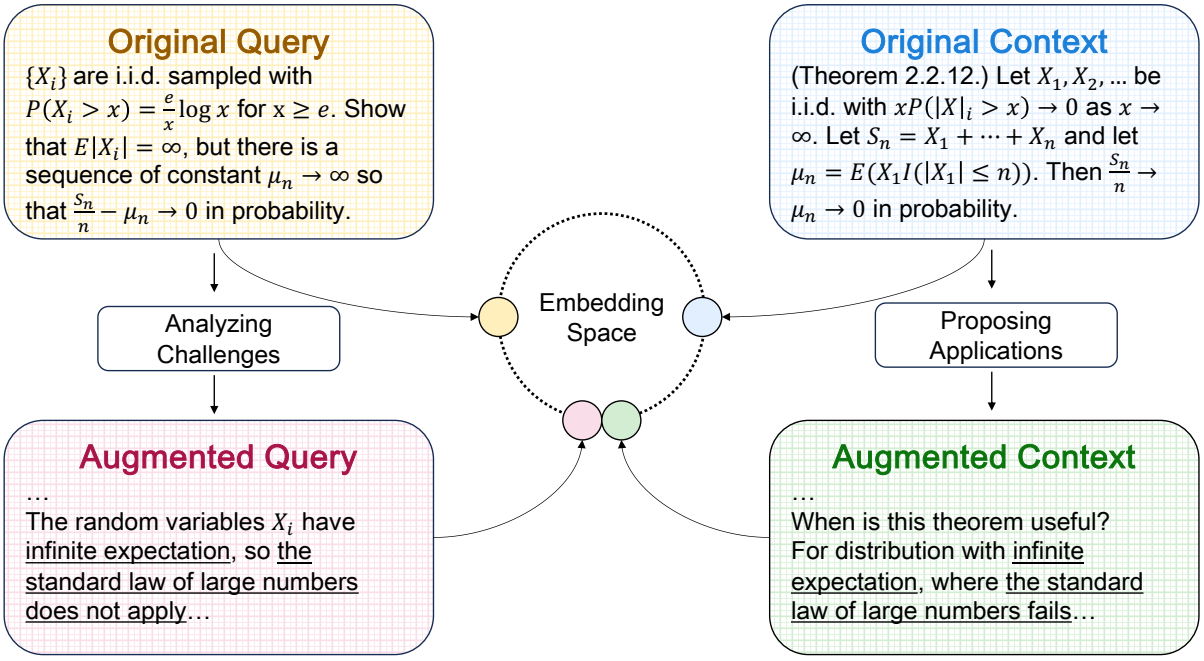

Figure 1: An example of dual augmentation. The original query (i.e., the target theorem) initially lies far from the original context in the embedding space. Dual augmentation reduces this distance by jointly analyzing the proof challenges of the query and identifying applications of known theorems in the context. Underlined text indicates semantically similar phrases between the augmented query and augmented context.

documents $\mathcal{D}$ in a manner that preserves the logical structure by grouping each known theorem, its proof, and associated examples/remarks into contiguous segments $\mathcal{C}$. This approach keeps logically connected content together, making it easier to retrieve and use for further analysis.

Dual augmentation. Dense retrieval, which leverages text embeddings to identify semantically proximate documents, has emerged as the mainstream approach in RAG systems (Lewis et al., 2020; Zhang et al., 2023; Chen et al., 2024). However, in our task, the target theorem $\mathcal{T}$ and relevant contextual information $\mathcal{C}_\mathcal{T}$ may not be close within the embedding space. For instance, when attempting to retrieve useful knowledge to prove a target theorem related to "Gaussian distribution," dense retrieval might inadvertently prioritize semantically proximate but irrelevant properties of Gaussian distributions, while the semantically distant yet critical knowledge (e.g., the law of large numbers) is missing due to its broader conceptual scope.

To address this issue, we propose a dual augmentation strategy that operates on both the target theorem and the documents. Specifically, for the target theorem, we leverage the LLM to analyze its underlying challenges and generate preliminary proof sketches that outline potential solution pathways. For the existing context, we employ the LLM to propose possible applications of known theorems and extract key techniques used in their proof. The proposed applications identified in the augmented context frequently align with the challenges posed by the target theorem, thereby reducing the semantic distance in the embedding space. Figure 1 illustrates this dual augmentation mechanism through a concrete example.

After augmenting the context, we feed the chunks into an embedding model and store the resulting representations in a vector database. Target theorem augmentation is carried out during inference, before the similarity search. We acknowledge the existence of some excellent RAG frameworks that employ similar query rewriting strategies, such as Rewrite-Retrieve-Read (Ma et al., 2023) and HyDE (Gao et al., 2023). However, directly adopting these methods without domain-specific adaptations is not feasible, as our task of theorem proving substantially diverges from conventional question-answering tasks. We provide empirical validation of this claim in §5.

Reranking. Reranking is also a powerful technique for enhancing retrieval performance (Wang et al., 2024b). Similar to the previous stages, we perform reranking in a zero-shot manner using

LLMs. Specifically, for the retrieved chunks containing known theorems and their corresponding proofs or explanations, we prompt an LLM to rank these chunks based on their relevance and usefulness to the target theorem $\mathcal{T}$. The ranked chunks are then used as context for powerful generative models (e.g., large reasoning models like OpenAI-o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025)) to assist in proving the target theorem.

5 Experiments

In this section, we conduct extensive experiments to study the task of natural language theorem proving with unstructured context. Specifically, we aim to answer the following research questions:

- **RQ1:** How does Dual RAG enhance retrieval performance on the Exercise100 dataset compared to existing RAG baselines?
- **RQ2:** Can improvements in retrieval quality translate into better proof generation for state-of-the-art reasoning models in natural language theorem proving?
- **RQ3:** Can the methods and insights developed in this work extend beyond textbook-level problems, e.g., support LLMs in addressing research-level theorem proving?
- **RQ4:** Do the retrieval improvements differ between exercises that rely solely on theorems and those that require proof techniques?
- **RQ5:** How do different components of Dual RAG contribute to its overall effectiveness?

5.1 Experimental setup

Datasets. To comprehensively evaluate both the retrieval and generation capabilities of our RAG framework, we constructed a new dataset, termed Exercise100, collected from graduate-level mathematics textbooks. Exercise100 comprises 100 theorem-proving problems evenly sampled from four standard textbooks: *Probability: Theory and Examples* (Durrett, 2010), *High-dimensional probability: An introduction with applications in data science* (Vershynin, 2018), *Introduction to Real Analysis* (Trench, 2009), and *Basic Topology* (M.A., 2004). For each problem, we manually annotated the most relevant contextual passages from the source textbooks as ground truth $\mathcal{G}_\mathcal{T}$ for retrieval evaluation. We will make this dataset publicly available to the greatest extent possible, adhering to the release policies of prior work (Frieder et al., 2023).

Retrieval setup. We evaluate the retrieval performance of Dual RAG and baseline methods on the Exercise100 dataset. Specifically, for each target theorem, we retrieve K chunks $\mathcal{C}_\mathcal{T}$ using different RAG frameworks and compute:

$$\text{Coverage@}K(\mathcal{T}) = \frac{|\mathcal{G}_\mathcal{T} \cap (\cup \mathcal{C}_\mathcal{T})|}{|\mathcal{G}_\mathcal{T}|}, \quad (1)$$

which measures the proportion of ground truth context covered by the retrieved chunks. We average the scores across all target theorems to obtain the dataset-level $\text{Coverage@}K$ metric. A higher coverage ratio indicates superior retrieval performance, which consequently enhances theorem proving capability of LLMs by providing more relevant contextual information.

For RAG baselines, we use Vanilla RAG (Lewis et al., 2020), Semantic Chunking (LangChain, 2024), Rewrite-Retrieve-Read (Ma et al., 2023), and HyDE (Gao et al., 2023). Since our work investigates a new task setting, we carefully developed our own implementations of the baseline approaches. The complete implementation code will be made publicly available to ensure reproducibility. For our framework, Dual RAG, we employ DeepSeek-V3 (DeepSeek-AI, 2024) as the base model due to its balance between performance and computational cost. The prompts used for each module are provided in Appendix A.

Generation setup. To evaluate the impact of retrieval on theorem proving and investigate whether enhanced retrieval performance translates to improved generation capabilities, we conduct a comparative analysis of three approaches: Raw LLMs, Vanilla RAG, and our framework Dual RAG. The experiments employ state-of-the-art reasoning models: DeepSeek-R1 (DeepSeek-AI, 2025) and OpenAI-o1 (OpenAI, 2024) as the backbone generative models. Detailed prompts for generation are provided in Appendix A. After collecting proofs for each target theorem, we compile them into a PDF format for manual evaluation. Our grading team for the Exercise100 dataset consists of four experts, all of whom are at least graduate-level students in mathematics-related fields and are familiar with the exercises they assess. We employ a 0/0.5/1 scoring scale, where 1 represents a perfect proof, 0.5 indicates a correct approach with minor errors

	PT		HDP		RA		TP	
# Chunks	4	8	4	8	4	8	4	8
Baselines								
Vanilla RAG	50.09	50.54	65.38	81.38	33.81	59.63	32.23	60.45
Semantic Chunking	15.80	21.55	46.67	55.20	33.62	60.95	23.47	25.09
Rewrite-Retrieve-Read	35.93	52.32	54.44	82.49	28.46	41.79	36.00	56.23
HyDE	51.21	56.65	64.00	83.38	33.02	59.02	44.23	68.23
Ours								
Dual RAG	**68.00**	**80.67**	**84.00**	**92.00**	**68.00**	**76.00**	**64.23**	**80.00**
Δ_r	17.91	30.13	18.62	10.62	34.19	16.37	32.00	19.55

Table 2: Retrieval performance of various RAG systems on Exercise100. We report the $\mathrm{Coverage@}K$ (in percentage) metric with $K \in \{4, 8\}$. Δ_r denotes the improvement of our proposed Dual RAG compared to Vanilla RAG. PT, HDP, RA, and TP refer to the four book sources, respectively. The best results are **bolded**.

	PT	HDP	RA	TP
DeepSeek-R1	62	90	80	64
+Vanilla RAG	70	94	86	82
+Dual RAG	**82**	**96**	**90**	**88**
Δ_g	12	2	4	6
OpenAI-o1	54	92	60	68
+Vanilla RAG	64	92	68	78
+Dual RAG	**76**	**98**	**76**	**86**
Δ_g	12	6	8	8

Table 3: Generation performance of two advanced reasoning models on Exercise100. We report the results of raw LLMs, LLMs with Vanilla RAG, and LLMs with Dual RAG. The scores are presented in percentages. Δ_g denotes the improvement of our proposed Dual RAG compared to Vanilla RAG. The best results are **bolded**.

or oversights, and 0 denotes an incorrect proof. For each textbook, we normalize the total scores to a scale of 100 to enhance readability.

5.2 Retrieval performance (RQ1)

The retrieval performance of Dual RAG and baseline methods on the Exercise100 dataset is detailed in Table 2. These results unequivocally demonstrate that Dual RAG consistently and substantially outperforms all evaluated RAG baselines across all four textbook sources. The magnitude of this improvement is highlighted by Δ_r, which indicates that Dual RAG surpasses Vanilla RAG by a margin ranging from 10.62% to as high as 34.19% in $\mathrm{Coverage@}K$. This robust enhancement in retrieval effectiveness underscores the advanced capabilities of our proposed framework. As we will demonstrate in §5.3, this marked improvement in retrieval quality subsequently translates into superior proof generation performance.

Notably, Dual RAG also exhibits a significant performance advantage over previous query rewriting techniques such as Rewrite-Retrieve-Read (Ma et al., 2023) and HyDE (Gao et al., 2023). While these methods also aim to improve retrieval by refining the query, our experiments reveal their limitations in the complex domain of natural language theorem proving. As shown in Table 2, these query rewriting approaches often yield inconsistent benefits over Vanilla RAG and, in several instances, even result in degraded performance (e.g., for exercises in the RA textbook with $K = 4$, where HyDE shows a 0.79% drop and Rewrite-Retrieve-Read shows a 5.35% drop compared to Vanilla RAG). This result confirms the importance of domain knowledge in retrieval performance.

5.3 Generation performance (RQ2)

The generation performance of two reasoning models, with and without RAG systems, is presented in Table 3. We observe that even these most advanced models struggle to excel in graduate-level theorem-proving tasks. However, incorporating retrieval mechanisms consistently enhances their performance. Notably, improvements in retrieval quality directly correlate with better proof generation. We observe that the generation gap (Δ_g) is smaller than the retrieval gap (Δ_r), primarily due to two types of exercises: (1) exercises requiring complex reasoning or deep search processes that exceed the capabilities of current LLMs, and (2)

Target theorem: For an oblivious adversarial crosslearning contextual bandit with a strongly observable feedback graph G and stochastic context distribution ν, a proposed algorithm for the problem achieves a regret bound of order $\widetilde{O}(\sqrt{\alpha T})$.

Retrieved documents: [Arxiv 2401.01857] We perform all our analysis conditioned on the following two events occurring with high probability, ..., conditioned on these two concentration events holding, ..., we then split $\mathrm{Reg}(u)$ into four terms, ..., terms $\mathbf{bias}_1$ and $\mathbf{bias}_2$, ..., term $\mathbf{ftrl}$...

[Arxiv 2012.05756v3] Lemma 4. Let G_t be a directed or undirected graph with vertex set $V := \{1, \ldots, K\}$. Let $\alpha(G_t)$ be the independence number of G_t and π be a distribution over V. Then,

$$\sum_{i \in V} \frac{\pi(i)}{c + \pi(i) + \sum_{j : j \xrightarrow{t} i} \pi(j)} \leq 2\alpha(G_t) \log\left(1 + \frac{\lceil K^2/c \rceil + K}{\alpha(G_t)}\right) + 2,$$

where c is a positive constant.

Generated proof: The proof follows the structure outlined in the analysis overview of Arxiv 2401.01857 (Section 3.2 and Appendix C), ..., 2. establishing high probability events, ..., 3. showing implications of event G, ..., 4. decomposing regret, ..., $\mathbf{FTRL}$ term, $\mathbf{Bias}$ terms...

Using Lemma 4 from (Arxiv: 2012.05756v3), we have $\sum_{a \in V} \frac{\pi(a)}{c + \pi(\mathrm{NodeIn}(a))} \leq 2\alpha(G) \log(1 + \lceil K^2/c \rceil / \alpha(G)) + 2$.

Table 4: An end-to-end example of using a retrieval-augmented language model to prove a research-level theorem in theoretical machine learning. The LLM generates a correct proof for a recently proposed open problem by retrieving and leveraging relevant results from the arXiv corpus. We use Gemini2.5-Pro for its superior long context ability. The useful proof technique in the retrieved arXiv paper is underlined. Critical steps of the mimetic proof are shaded.

exercises based on well-known contexts that LLMs already encode in their parameters. Despite this disparity, Dual RAG remains an effective approach for advancing state-of-the-art LLMs in natural language theorem-proving.

5.4 Research-level example (RQ3)

A natural question to explore is whether RAG systems can extend their capabilities to solving research-level theorems, which is a key aspiration for developing mathematical copilots. Our study provides promising evidence toward this goal, demonstrating that language models augmented with retrieval can emulate advanced theorem-proving techniques.

As a case study, we applied our system to a recently proposed open problem in theoretical machine learning. This problem, introduced by Han et al. (2020); Wen et al. (2024), remained unsolved until late 2024, despite its seemingly modest complexity. It originates from the domain of online ad display auctions and is also of significant theoretical interest in its own right. Specifically, the problem lies in the field of Multi-Armed Bandits, a core area that investigates the tradeoff between exploration and exploitation, and is closely tied to reinforcement learning. The goal was to characterize the tight bound for a novel variant of a bandit problem. The details of this task are deferred to the Appendix B due to page limit.

We approached the task by feeding the problem statement and algorithm description to an LLM, prompting it to generate search queries. These were used with Dual RAG, which interfaces with the arXiv API to retrieve relevant LaTeX sources. Using this retrieved context, the LLM produced a complete proof that closely mirrors the structure of related works. Experts in theoretical machine learning have verified the essential correctness of the proof, and confirm that it only need minor adjustments to be fully correct.

Table 4 presents an overview of the generated proof, along with the documents retrieved from arXiv that the model relied on during reasoning. The complete proof is included in our supplementary data. This case illustrates the key insight of our work: retrieval-augmented LLMs can go beyond simply quoting or applying relevant theorems, they

are also capable of picking up on and adapting the proof strategies used in related literature. In this example, the model not only incorporates a lemma from one retrieved paper, but also mirrors the structural techniques of another, including conditioning on high-probability events and decomposing regret terms. This suggests that LLMs, when equipped with the right context, can begin to internalize and reproduce the reasoning patterns found in advanced mathematical writing.

5.5 Divided exercises (RQ4)

	Theorem		**Technique**	
# Chunks	4	8	4	8
Vanilla RAG	46.25	46.85	62.25	62.25
Dual RAG	63.16	73.68	83.34	100.0
Δ_r	16.91	26.83	21.09	37.75

Table 5: Retrieval performance on exercises from the *Probability Theory* textbook, divided into Theorem and Technique subsets.

As discussed in §1, a key contribution of our work is highlighting the importance of retrieving proof techniques. After demonstrating the effectiveness of Dual RAG and its retrieval improvements on Exercise100, we further investigate where these improvements originate. To this end, we divide the exercises in the textbook *Probability: Theory and Examples* (Durrett, 2010) into two subsets: *Theorem*, which require only retrieving theorems, and *Technique*, which rely on identifying and applying proof strategies.

Table 5 reports the retrieval performance of Vanilla RAG and Dual RAG on these subsets. We observe consistent improvements across both subsets with Dual RAG, but the gains are significantly higher for the Technique-based exercises. This suggests that LLM-augmented retrieval and domain knowledge is particularly beneficial in scenarios where reasoning and methodological insight are needed. By dually augment both target theorems and context, Dual RAG achieve superior retrieval performance on natrual language theorem proving task.

5.6 Ablation study (RQ5)

To better understand the contribution of each design component in Dual RAG, we conduct an ablation study on three key aspects: reranking, data augmentation, and chunking. Table 6 shows results on problems from two textbooks in the Exercise100 dataset: *Probability: Theory and Examples* and *Introduction to Real Analysis*. Removing the reranker leads to moderate performance drops, confirming its role in refining retrieval. Excluding data augmentation causes a larger degradation, highlighting its importance in bridging semantic gaps. We also find chunking to be essential, as it underpins other modules. Overall, all components are critical to the effectiveness of Dual RAG.

	PT		**RA**	
# Chunks	4	8	4	8
Dual RAG	**68.00**	**80.67**	**68.00**	**76.00**
w/o Rerank	65.23	72.00	64.00	72.00
w/o Aug.	53.23	60.21	54.64	68.99
w/o Chunk.	34.10	42.10	58.42	66.89

Table 6: Ablation study results for Dual RAG on problems from two textbooks in the Exercise100 dataset. We show the impact of removing individual components: reranking (w/o Rerank), data augmentation (w/o Aug.), and chunking (w/o Chunk.).

6 Conclusion

In this paper, we study the task of natural language theorem proving with unstructured context. This setting is both practical and important for building a mathematical copilot. We observe that advanced LLMs exhibit strong mimetic capabilities in such scenarios. This insight suggests that effective retrieval should involve both related theorems and relevant proof techniques for each target theorem. Motivated by domain knowledge, we propose Dual RAG, a RAG system that dually augments the target theorem and its context with LLM-generated analyses. These analyses highlight the difficulties of the target theorem and suggest potential applications of the context, bringing semantically aligned query-context pairs closer and improving retrieval quality. This in turn leads to better proof generation on our graduate-level dataset, Exercise100. We further integrate Dual RAG with the arXiv API to address research-level problems in theoretical machine learning. Remarkably, we identify a case where an LLM, given the right context, successfully solves an unpublished research problem. While current models remain limited in their ability to tackle novel theorems requiring fundamentally new techniques, our work represents a meaningful step toward building a practical mathematical copilot.

Limitations

This paper investigates the effectiveness of retrieval-augmented language models for natural language theorem proving. While our method demonstrates strong mimetic capabilities on theorems that can be solved using existing proof techniques (both at the textbook and research level), it shows no significant improvement on problems that require fundamentally novel proof strategies. Although the theoretical machine learning example in our paper addresses an open question, its complexity is considered moderate by domain experts. This limitation is expected, as such problems can be viewed as extremely challenging out-of-distribution cases, where the necessary proof techniques are neither encoded in the LLM parameters nor present in the retrieved context. Addressing these types of problems is an important and open research direction, which we leave for future work.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2309.07597.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Rick Durrett. 2010. *Probability: Theory and Examples, 4th Edition*. Cambridge University Press.

Deborah Ferreira and André Freitas. 2020. Premise selection in natural language mathematical texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7365–7374. Association for Computational Linguistics.

Simon Frieder, Jonas Bayer, Katherine M. Collins, Julius Berner, Jacob Loader, András Juhász, Fabian Ruehle, Sean Welleck, Gabriel Poesia, Ryan-Rhys Griffiths, Adrian Weller, Anirudh Goyal, Thomas Lukasiewicz, and Timothy Gowers. 2024. Data for mathematical copilots: Better ways of presenting proofs for machine learning. *Preprint*, arXiv:2412.15184.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yanjun Han, Zhengyuan Zhou, and Tsachy Weissman. 2020. Optimal no-regret learning in repeated first-price auctions. *Oper. Res.*, 73:209–238.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Owen Henkel, Zachary Levonian, Chenglu Li, and Millie-Ellen Postle. 2024. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In *Proceedings of the 17th International Conference on Educational Data Mining, EDM 2024, Atlanta, Georgia, USA, July 14-17, 2024*. International Educational Data Mining Society.

Jiewen Hu, Thomas Zhu, and Sean Welleck. 2025. miniCTX: Neural theorem proving with (long-)contexts. In *The Thirteenth International Conference on Learning Representations*.

Geoffrey Irving, Christian Szegedy, Alexander A. Alemi, Niklas Eén, François Chollet, and Josef Urban. 2016. Deepmath - deep sequence models for premise selection. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2235–2243.

Laura Kovács and Andrei Voronkov. 2013. First-order theorem proving and vampire. In *Computer Aided Verification - 25th International Conference, CAV 2013, Saint Petersburg, Russia, July 13-19, 2013. Proceedings*, volume 8044 of *Lecture Notes in Computer Science*, pages 1–35. Springer.

Andrzej Stanisław Kucik and Konstantin Korovin. 2018. Premise selection with neural networks and

distributed representation of features. *Preprint*, arXiv:1807.10268.

Daniel Kühlwein, Twan van Laarhoven, Evgeni Tsivtsivadze, Josef Urban, and Tom Heskes. 2012. Overview and evaluation of premise selection techniques for large theory mathematics. In *Automated Reasoning - 6th International Joint Conference, IJCAR 2012, Manchester, UK, June 26-29, 2012. Proceedings*, volume 7364 of *Lecture Notes in Computer Science*, pages 378–392. Springer.

LangChain. 2024. How to split text based on semantic similarity.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. 2024. A survey on deep learning for theorem proving. In *First Conference on Language Modeling*.

Armstrong M.A. 2004. *Basic Topology*. Springer (India) Pvt. Limited.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

OpenAI. 2024. Learning to reason with LLMs.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. Proof or bluff? evaluating llms on 2025 usa math olympiad. *Preprint*, arXiv:2503.21934.

Bartosz Piotrowski and Josef Urban. 2020. Stateful premise selection by recurrent neural networks. In *LPAR 2020: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning, Alicante, Spain, May 22-27, 2020*, volume 73 of *EPiC Series in Computing*, pages 409–422. EasyChair.

John Alan Robinson and Andrei Voronkov, editors. 2001. *Handbook of Automated Reasoning (in 2 volumes)*. Elsevier and MIT Press.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

William Trench. 2009. *Introduction to Real Analysis*. Pearson College Div.

Roman Vershynin. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, Zhenguo Li, and Xiaodan Liang. 2024a. Lego-prover: Neural theorem proving with growing libraries. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024b. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736, Miami, Florida, USA. Association for Computational Linguistics.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hanna Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. Naturalproofs: Mathematical theorem proving in natural language. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. Naturalprover: Grounded mathematical proof generation with language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yuxiao Wen, Yanjun Han, and Zhengyuan Zhou. 2024. Stochastic contextual bandits with graph feedback: from independence number to mas number. In *Advances in Neural Information Processing Systems*, volume 37, pages 64499–64522. Curran Associates, Inc.

Huajian Xin, Z.Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Haowei Zhang, Qihao Zhu, Dejian Yang, Zhibin Gou, Z.F. Wu, Fuli Luo, and Chong Ruan. 2025. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. In *The Thirteenth International Conference on Learning Representations*.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J. Prenger, and Animashree Anandkumar. 2023. Leandojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*

2023, New Orleans, LA, USA, December 10 - 16, 2023.

Ling Yang, Zhaochen Yu, Bin Cui, and Mengdi Wang. 2025a. Reasonflux: Hierarchical llm reasoning via scaling thought templates. *Preprint*, arXiv:2502.06772.

Shu-Xun Yang, Cunxiang Wang, Yidong Wang, Xiaotao Gu, Minlie Huang, and Jie Tang. 2025b. Stepmathagent: A step-wise agent for evaluating mathematical processes through tree-of-error. *Preprint*, arXiv:2503.10105.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *Preprint*, arXiv:2310.07554.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *Preprint*, arXiv:2409.14924.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

A Prompt

Prompt for raw LLMs generation. We use the following prompt for raw LLMs generation (i.e., without RAG):

Raw LLMs Prompt

```
You are a helpful and highly
knowledgeable assistant specialized in
mathematical proofs. Your task is to
provide a rigorous proof for the given
problem. Ensure that the proof is
logically sound, clearly explained, and
formatted in correct LaTeX syntax. Break
 the proof into distinct steps and
summarize the conclusion at the end.
Avoid any logical fallacies or omissions
.

{Target theorem}
```

Prompt for retrieval-augmented generation. We use the following prompt for retrieval-augmented generation:

RAG Prompt

```
You are a helpful and highly
knowledgeable assistant specialized in
mathematical proofs. Your task is to
provide a rigorous proof for the given
problem. Ensure that the proof is
logically sound, clearly explained, and
formatted in correct LaTeX syntax. Break
 the proof into distinct steps and
summarize the conclusion at the end.
Avoid any logical fallacies or omissions
.

{Target theorem}

Here are relevant context documents that
 may help with the proof:

[Document 1]
{Document 1}
[End Document 1]
...
```

Prompt for LLM-based chunking. We use the following prompt for LLM-based chunking:

Chunking Prompt

```
Please carefully segment the provided
LaTeX source code while adhering to
these requirements:
1.Keep all theorem/proposition/lemma
environments and their corresponding
proofs intact within the same segment.
2.Create segments whenever 1-2 theorems/
propositions appear.
3.Preserve all original content exactly
- only insert '[SEP]' markers between
segments
4.Output only the processed text with
segmentation markers - no additional
commentary

{LaTex source code}
```

Prompt for query augmentation. We use the following prompt to augment target theorems:

Target Theorem Augmentation Prompt

```
You are a mathematics expert skilled in
analyzing and solving complex
mathematical proof problems. Your task
is to analyze the following proof
problem, identify its challenges, and
determine the potential theorems and
proof techniques that might be required.
 You do not need to finish the proof,
but you should provide a detailed
analysis of the problem and the
potential strategies to solve it.

{Target theorem}
```

Prompt for document augmentation. We use the following prompt to augment known theorems in documents:

Known Theorem Augmentation Prompt

```
You are a mathematics expert skilled in
understanding and analyzing mathematical
 theorems and their proofs. I will
provide you with some mathematical
theorems and their proofs. Please help
me with the following tasks:

Theorem Summary: Summarize the core idea
 of the theorem in concise language.
Potential Applications: Based on your
knowledge, suggest possible applications
 of this theorem (e.g., give example
exercise that can be proved with this
theorem).
Proof Technique Analysis: Identify the
key techniques or methods used in the
proof (e.g., induction, contradiction,
constructive proof, limit arguments, etc
.) and briefly explain how these
techniques are applied.
Please analyze the following theorem and
 its proof:

{Known theorems}
```

Prompt for chunk reranking. We use the following prompt to rerank the retrieved chunks:

Reranking Prompt

```
You are an intelligent assistant that
can rank mathematical theorems and
statements based on their usefulness in
proving a given theorem. Your goal is to
 identify the most relevant theorems or
statements that directly contribute to
the proof of the target theorem, while
avoiding those that are semantically
similar but irrelevant to the specific
proof.

I will provide you with {K} passages,
each indicated by number identifier [].
Your task is to rank these passages
based on their usefulness in proving the
 following theorem:
{Target theorem}.

Focus on selecting theorems or
statements that are directly applicable
to constructing the proof, rather than
those that are merely semantically
similar but do not contribute
meaningfully to the proof.
```

B Details of the theoretical machine learning example

We present the details of the theoretical machine learning example. Specifically, we want to use LLMs to prove the following theorem. For readers not familiar with theoretical machine learning, we provide a detailed problem statement in the following subsection. The full proof generated by Gemini-2.5-pro-preview is uploaded as data.

B.1 Problem Statement

We study a contextual K-armed bandit problem over T rounds with graphical feedback, where contexts belong to the set $[M]$. We consider the oblivious adversarial bandits. At the beginning of the problem, an oblivious adversary selects a sequence of losses $\ell_{t,c}(a) \in [0,1]$ for every round $t \in [T]$, every context $c \in [M]$, and every arm $a \in [K]$. Note that all of our results apply to stochastic bandits as well, since the oblivious adversarial bandits are strictly stronger than stochastic bandits.

We assume that there is a directed feedback graph G over the set of arms $[K]$ with edge set E. We use the following graph-theoretic notations. For each arm a, let $N_{\text{out}}(a) = \{v \in [K] : a \to v\}$ be the set of out-neighbors of a (including a itself), and let $\mathrm{N}_{\text{in}}(a) = \{v \in [K] : v \to a\}$ be the set of in-neighbors of a. The independence number $\alpha(G)$ is defined as the cardinal number of the maximal independence set of G. For any vector p of dimension $[K]$ and set $V \subset [K]$, we denote $p(V) \triangleq \sum_{v \in V} p(v)$. For example, the notation $p(\mathrm{N}_{\text{in}}(a))$ means $\sum_{a' \to a} p(a')$.

In each round t, we begin by sampling a context $c_t \sim \nu$ i.i.d. from an unknown distribution ν over $[M]$, and we reveal this context to the learner. Based on this context, the learner selects an arm $a_t \in [K]$ to play. The adversary then reveals the function $\ell_{t,c}(a)$ for all $a \in N_{\text{out}}(a_t)$, and the learner suffers loss $\ell_{t,c_t}(a_t)$. Notably, the learner observes the loss for every context $c \in [M]$ and every arm $a \in N_{\text{out}}(a_t)$.

We aim to design learning algorithms that minimize regret. Fix a policy $\pi : [M] \to [K]$. With a slight abuse of notation, we also denote $\pi_c = e_k \in \Delta([K])$ for each $c \in [M]$. The expected regret with respect to policy π is

$$\mathrm{Reg}(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,c_t}(a_t) - \ell_{t,c_t}(\pi_{c_t})\right]$$

We aim to upper bound this quantity (for an arbitrary policy π).

B.2 The Theorem

Theorem 1. *For an oblivious adversarial cross-learning contextual bandit with a strongly observable feedback graph G and stochastic context distribution ν, for $\iota = 2\log(8KT^2), L =$*

$\sqrt{\frac{\iota\alpha T}{\log(K)}} = \widetilde{\Theta}(\sqrt{\alpha T}), \gamma = \frac{16\iota}{L} = \widetilde{\Theta}(1/\sqrt{\alpha T})$, *and* $\eta = \frac{\gamma}{2(2L\gamma+\iota)} = \widetilde{\Theta}(1/\sqrt{\alpha T})$, *Algorithm 1 yields a regret bound of*

$$\mathrm{Reg}(\pi) = \widetilde{O}(\sqrt{\alpha T}).$$

Algorithm 1 The algorithm for the unknown distribution setting

Input: Parameters $\eta, \gamma > 0$ and $L < T$.
$\widehat{w}_2 \leftarrow 0$
$s_{1,c} \leftarrow \frac{1}{K}$ for each c
$s_{2,c} \leftarrow \frac{1}{K}$ for each c
for $t = 1, \ldots, L$ **do**
  Observe c_t
  Play $A_t \sim s_{1,c_t}$
  for $a \in [K]$ **do**
    $\widehat{w}_2(a) \leftarrow \widehat{w}_2(a) + \frac{s_{2,c_t}(\mathrm{N_{in}}(a))}{2L}$
for $e = 2, \ldots, T/L$ **do**
  $\widehat{w}_{e+1} \leftarrow 0$
  for $t = (e-1)L+1, t = (e-1)L+3, \ldots, eL-1$ **do**
    Set $p_{t,c} = \operatorname{argmin}_{p \in \Delta[K]} \left(\left\langle p, \sum_{s=1}^{t-1} \widehat{\ell}_s(c) \right\rangle - \eta^{-1} F(p) \right)$
    for $t' = t, t+1$ **do**
      Observe $c_{t'}$
      if $p_{t,c_{t'}}(a) \geq s_{e,c_{t'}}(a)/2$ *for all* $a \in [K]$ **then**
        Set $q_{t',c_{t'}} = p_{t,c_{t'}}$
      else
        Set $q_{t',c_{t'}} = s_{e,c_{t'}}$
      Play $A_{t'} \sim q_{t',c_{t'}}$
      Observe $\ell_{t',A_{t'}}$
    $t_f, t_\ell \leftarrow \mathsf{RandPerm}(t, t+1)$
    for $a \in [K]$ **do**
      $\widehat{w}_{e+1}(a) \leftarrow \widehat{w}_{e+1}(a) + \frac{s_{e+1,c_{t_f}}(\mathrm{N_{in}}(a))}{2(L/2)}$
      Sample $S_{t,a} \sim \mathcal{B}\left(\frac{s_{e,c_{t_\ell}}(\mathrm{N_{in}}(a)}{2q_{t,c_{t_\ell}}(\mathrm{N_{in}}(a))} \right)$
      $\widehat{\ell}_{t_\ell,c}(a) \leftarrow \frac{2\ell_{t_\ell,c}(a)}{\widehat{w}_e(\mathrm{N_{in}}(a)) + \frac{3}{2}\gamma} \mathbb{I}\left(A_{t_\ell} \to a, S_{t,a} = 1\right)$
  $s_{e+2} \leftarrow p_t$

C Potential risks

Our approach has the potential to assist researchers in tackling complex mathematical proofs, which could enhance productivity in mathematical and theoretical research. However, we caution against the uncritical use of language models for tasks such as automatic paper writing or theorem generation without human oversight. The mimetic capabilities demonstrated by our system should be viewed as a tool for exploration and support, rather than a replacement for rigorous human reasoning and scholarly standards.

D Artifacts information

Similar to the GHOST dataset (Frieder et al., 2023), some parts of the datasets contain information that may be protected under copyright, so we will release data without these information. The retrieved documents in the theoretical machine learning example are all arXiv papers, the license can be found in `https://info.arxiv.org/help/license/index.html`. The data contains no information that names or uniquely identifies individual people or offensive content.

E Human Subjects

We use humman annotators for the evaluation of both the textbook-level and research-level theorem proving. The annotators are recruited from the graduate school and supported by grants. All annotators agree to share data. The data collection protocol is approved by an ethics review board.