Learning to Reason with Insight for Informal Theorem Proving

Anonymous ACL submission

Abstract

Informal theorem proving better matches large language models' natural-language strengths than fully formal proof search, but current models still struggle to identify the core techniques that unlock difficult proofs. We view this missing ability as a lack of mathematical *insight*: the capacity to recognize the pivotal construction, theorem, transformation, or proof organization before writing the full solution. We propose `DeepInsight`, an insight-oriented framework for informal theorem proving. It contains `DeepInsightTheorem`, a hierarchical dataset that annotates proofs with core techniques and proof sketches; a progressive multi-stage SFT curriculum that teaches proof writing, planning, and insight identification; and `InsightPO`, a GRPO-style reinforcement learning algorithm that assigns structured rewards over this hierarchy. Experiments on challenging theorem-proving benchmarks show that explicitly training models to recognize and apply core techniques substantially improves mathematical reasoning, outperforming standard SFT, structured prompting baselines, and strong open-source theorem-proving models.

1 Introduction

Automated theorem proving (ATP) is a long-standing benchmark for machine reasoning. While most recent LLM-based ATP systems rely on formal proof engines such as Lean, Coq, and Isabelle (Zheng et al., 2022; Liu et al., 2023a; Tsoukalas et al., 2024) or specialized languages (Welleck et al., 2022), informal theorem proving asks models to produce human-readable mathematical proofs in natural language and LaTeX. This setting is closer to how mathematicians communicate proofs and is well aligned with LLMs' strengths in language understanding, long-context generation, and mathematical exposition.

However, informal theorem proving remains underexplored (Welleck et al., 2022; Zhang et al., 2025b). Existing work mainly builds evaluation or data frameworks, leaving the proof-generation mechanism and its key bottlenecks less studied. In particular, simply fine-tuning on question-proof pairs often encourages models to imitate surface proof patterns, while difficult proof problems require discovering the hidden idea that determines the solution path. Additional related work is discussed in Appendix A.

We argue that a major bottleneck is the lack of *insight*: the ability to recognize the pivotal ideas or technical tools that make a proof possible. We call these tools *core techniques*, such as a key construction, lemma, theorem, or transformation, rather than routine logical steps. Human solvers often first form such a high-level view and then fill in the proof details; without this step, a model may generate fluent but directionless derivations. Our central idea is therefore to teach models this skill through a two-phase process: acquiring core techniques by reviewing existing proofs, and applying them by first identifying the relevant techniques for a new problem before constructing the final proof.

To realize this idea, we propose `DeepInsight`. First, we construct `DeepInsightTheorem`, extending DeepTheorem (Zhang et al., 2025b) with explicit core-technique and proof-sketch annotations. Second, we use a progressive multi-stage SFT curriculum that moves from proof writing to proof planning and then to technique identification, mirroring a learning process from solution imitation to insight-driven problem solving. Third, we introduce `InsightPO`, a GRPO-style reinforcement learning algorithm that uses the hierarchy to provide rewards for technique identification, planning, and proof completion, rather than treating a whole response as a flat scalar outcome. Experiments on FIMO, PutnamBench, and HMMT show consistent gains over standard SFT, structured-prompting baselines, and strong open-source theorem-proving models. Because both reinforcement learning and

evaluation use LLM-based verification, we also introduce a stratified human-audit protocol to calibrate verifier scores and inspect high-disagreement cases. Our contributions are:

- We identify core-technique recognition as a key bottleneck in informal theorem proving and formalize it through the notion of mathematical insight.
- We construct `DeepInsightTheorem`, a hierarchical dataset with core-technique, proof-sketch, and final-proof annotations.
- We design a progressive multi-stage SFT curriculum and `InsightPO`, a hierarchy-aware GRPO-style RL algorithm with structured verifier rewards.
- We validate the full framework across challenging benchmarks, structured prompting baselines, and a human-audited evaluation protocol.

2 Preliminary

We consider the context of informal mathematical proof generation. Let $\mathcal{V}$ be the vocabulary, which is a finite discrete set. The text sequence space $\mathcal{S}$ is defined as the union of the product spaces of any finite power $\mathcal{S} \triangleq \bigcup_{l=0}^{\infty} \mathcal{V}^l$.

We can then denote the space of theorems as a subspace $\mathcal{Q}$ of $\mathcal{S}$ which contains all well-defined math problems for theorem-proving. For any $q \in \mathcal{Q}$, we denote the space of proofs with respect to q as $\mathcal{P}(q) \subset \mathcal{S}$ that contains all valid proofs of q.

The LLM M_θ is an auto-regressive network which will predict the next token's distribution over the vocabulary when it is working. To train a generation policy π_θ to be a theorem-prover, the training datasets are collected from many resources of math corpus and commonly formed as theorem-proof pairs (q, p). For Supervised Fine-Tuning (SFT) on such a dataset, the objective is to solve $\max_\theta \mathbb{E}_{q\sim\mathcal{D}, p\sim\mathcal{D}_{P(q)}} \pi_\theta(p|q)$, where $\mathcal{D}$ is the distribution of the math problems in the training set and $\mathcal{D}_{P(q)}$ is the distribution over the valid proofs of each question in the training set.

2.1 Notion of Insight

For a mathematical proof question, *insight* is a cognitive act of identifying a set of pivotal ideas (e.g., "apply the Pigeonhole Principle", "utilize a specific invariant xxx") through the given conditions to capture the essence of the solution. We refer to such pivotal ideas as *core techniques* in this paper. Insight is not the proof itself, but instead refers to a preliminary, high-level perception and foresight of the core techniques required to construct the proof. Generally, we group these techniques into three main categories that cover the most common and essential technical patterns in math proofs:

- **Construction**: Introducing auxiliary objects, e.g., "Define the sequence $x_{n+1} = Ux_n$ for $n \geq 0$. This Picard iteration constructs the sequence $x_0, x_1 = Ux_0, x_2 = U^2x_0, \ldots$".
- **Theorem Call**: Invoking a known lemma, theorem, or any existing result, e.g. "By using Cantor's Theorem."
- **Mathematical Transformation**: Performing a reformulation that recasts the problem in a new framework, e.g., "Define a topology τ on the set of integers $\mathbb{Z}$..., shifting the language and tools entirely from number theory to topology."

Formally, we denote the space of techniques as $\mathcal{T} \subset \mathcal{S}$ that contains all possible techniques in those three classes. Generally, we can then write each proof p of q as an ordered sequence

$$p = (r_1, t_1, r_2, t_2, \cdots, r_k), \tag{1}$$

where $t_i \in \mathcal{T}$, r_i denote an elementary reasoning statement between successive techniques, and k is the total number of techniques in the proof.

3 Motivating Example: Insightfulness Evaluation of Off-the-Shelf LLMs

Expert mathematicians can often quickly form a big-picture view of a proof when facing a novel problem, drawing on insights cultivated through experience. In contrast, general-purpose LLMs often fail to reliably form such high-level ideas at the outset of proof generation, and are not as adept as experts at identifying the pivotal technique early.

We evaluated several powerful commercial LLMs with a non-thinking mode[1], including Gemini 2.5 Flash (Gemini Team, 2025) and DeepSeek-R1 (DeepSeek-AI, 2025), by prompting them to provide insights for math competition problems randomly sampled from Putnam and FIMO datasets (See Appendix H for detailed prompts). We then use o3-mini to review and evaluate the insights generated by these models from the dimensions of depth and completeness (Appendix C.1). Here, depth indicates the insight is specific and

[1] Long chain-of-thought reasoning often make the model first go over the whole proof and then summarize the core techniques, which is against the thinking pattern of insight. Therefore, it is more appropriate to disable the thinking mode.

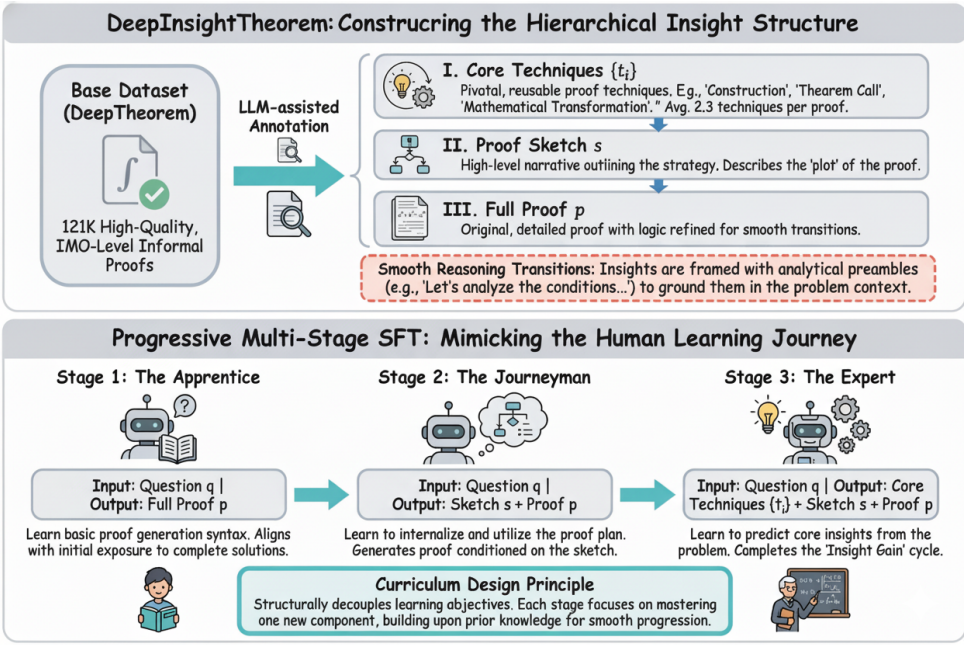

Figure 1: The top figure illustrates the construction of `DeepInsightTheorem`, and the bottom figure depicts our progressive multi-stage SFT procedure.

technically accurate, rather than a generic, high-level statement or a vague direction, while completeness measures whether all core techniques required for the solution are identified. Our results show that, in terms of depth, both models largely failed to provide specific and accurate insights, instead producing shallow guesses or vague hints at some directions. In terms of completeness, they always see just one or two low-hanging ideas while missing more central techniques.

4 Methodology

4.1 Challenge of Identifying Core Techniques

The main body of a proof is largely composed of some basic, easy-to-learn mathematical reasoning statements. For a theorem-proving LLM denoted by π_{TP} generating the proof for a problem, the uncertainty remains low until a core technique needs to be recognized. This likely arises for two main reasons. First, core techniques are intrinsically harder to identify. Second, since basic reasoning steps constitute a large fraction of the proof, LLM training may overly emphasize these basic reasoning details rather than core technique recognition.

As illustrated in Figure 2, in the proof generated by Qwen2.5-14B-Instruct (Qwen Team, 2025) for the well-known Baire Category Theorem in General Topology, tokens with spiking entropy consistently signal the presence of core techniques:

Construction: "`Constructing a Sequence of ...`" to construct a sequence of nested open sets.

Theorem call: "`Using completeness...`" to apply the completeness of the ground space.

Mathematical transformation: "`Consider the closures...`" to transfer the issue of open sets to their closures.

Note that the above three techniques are defined in Section 2.1. The rest of the proof is based on basic elementary reasoning statements that connect these core techniques, making the overall argument complete and smooth.

As discussed above, the model π_{TP} struggles to decide which core technique to apply, as evidenced by the token-level entropy spikes at positions where a core technique is introduced. To further formalize this, we assume

$$\pi_{\text{TP}}(t_i|u_i, q) < \delta \tag{2}$$

for some small constant $\delta > 0$ and any $t_i \in \mathcal{T}$, where $u_i := (r_1, t_1, \cdots, r_i)$ denotes the prefix before t_i and i indexes the i-th technique in a proof of form (1). This reflects the difficulty or the high uncertainty of identifying correct techniques. Consequently, we have the probability of directly predicting a core technique t_i from the question q bounded by δ: $\pi_{\text{TP}}(t_i|q) = \sum_{u_i} \pi_{\text{TP}}(t_i|u_i, q)\pi_{\text{TP}}(u_i|q) < \delta$. Similarly, by recursively applying (2), for the probability of generating a valid full proof p for q via π_{TP}, we have $\pi_{\text{TP}}(p|q) \leq \delta^k$, where k is the total number of techniques appearing in the proof. This indicates that the ability to complete a proof

also dense in \(X\). ### Step-by-Step Proof: 1. **Definition of Density**: A set \(A \subseteq X\) is dense in \(X\) if for every non-empty open set \(V \subseteq X\), the intersection \(A \cap V \neq \emptyset\). 2. **Given**: Each \(U_n\) is dense and open in \(X\). This means for any non-empty open set \(V \subseteq X\), \(U_n \cap V \neq \emptyset\) for all \(n \in \mathbb{N}\). 3. **Goal**: Show that \(\bigcap_{n=1}^\infty U_n\) is dense in \(X\). That is, for any non-empty open set \(V \subseteq X\), \(\left(\bigcap_{n=1}^\infty U_n \right) \cap V \neq \emptyset\). 4. **Constructing a Sequence of Nested Open Sets**: - Start with an arbitrary non-empty open set \(V \subseteq X\). - Since \(U_1\) is dense and open, there exists a non-empty open set \(V_1 \subseteq V \cap U_1\). - Since \(U_2\) is dense and open, there exists a non-empty open set \(V_2 \subseteq V_1 \cap U_2\). - Continue this process to construct a sequence of nested non-empty open sets \(\{V_n\}_{n=1}^\infty\) such that \(V_{n+1} \subseteq V_n \cap U_{n+1}\). 5. **Properties of the Sequence**: - Each \(V_n\) is non-empty and open. - \(V_{n+1} \subseteq V_n\) implies that \(\{V_n\}\) is a decreasing sequence of non-empty open sets. 6. **Using Completeness**: - By the completeness of \(X\), the intersection of any nested sequence of non-empty closed sets with diameters tending to zero is non-empty. - Consider the closures \(\overline{V_n}\) of these sets. Since \(V_n\) are open and nested, \(\overline{V_n}\) form a nested sequence of non-empty closed sets. - The diameter of \(V_n\) can be made arbitrarily small by choosing appropriate \(V_n\). Hence, \(\bigcap_{n=1}^\infty \overline{V_n} \neq \emptyset\). 7. **Intersection Point**: - Let \(x \in \bigcap_{n=1}^\infty \overline{V_n}\). Since each \(V_n \subseteq U_n\), it follows that \(x \in \overline{V_n} \subseteq \overline{U_n} = U_n\) (because \(U_n\) is open). - Therefore, \(x \in \bigcap_{n=1}^\infty U_n\). 8. **Conclusion**: - We have

Figure 2: The generated proof of the Baire Category Theorem generated by Qwen2.5-14B-Instruct. From light grey to deep warm color, the token entropy increases. Most tokens are of low entropy, composing basic mathematical reasoning. Tokens with a spiking entropy indicate core techniques.

is restricted by the uncertainty. A detailed derivation is in Appendix B. We can interpret the value $\pi_{\text{TP}}(t_i|q)$ as the *insightfulness* of π_{TP} in identifying technique $t_i \in \mathcal{T}$. A higher value indicates stronger insight in recognizing t_i as the appropriate technique to apply.

Therefore, the successful realization of core techniques is crucial for generating valid proofs in theorem proving. Core techniques also serve as more appropriate memory anchors for key pattern recognition in complex mathematical reasoning. This prompts a necessity of enhanced learning to bridge the gap between the question and the core techniques required to solve it.

4.2 Insight Acquisition and Application

It is critical to figure out the core techniques for the proof of a specific question if we want an LLM to grasp the essence of such a proof. After identifying the relevant techniques, the remaining challenge is to integrate these core techniques into a coherent solution, which relies primarily on the fundamental mathematical reasoning capability.

This reflects a shift in the underlying reasoning pattern for an LLM. In human learning, when presented with a problem and its corresponding proof, beginners typically need to carefully study and review the entire solution to gradually identify and understand the core techniques involved. Subsequently, when faced with similar problems later, they will first form an initial idea about the core techniques and then use the intuition to construct the proof.

Our framework comprises the following two complementary cognitive processes, mirroring how humans consolidate and review learning materials:

- **Acquisition phase**: extract core techniques by reviewing proofs in a training corpus;
- **Application phase**: develop insightful reasoning by first predicting the core techniques and then generating proofs by applying these techniques.

We consider the following generation process for a proof p via a policy π_θ parameterized by θ:

$$\pi_\theta(t, p|q) = \pi_\theta(t|q)\pi_\theta(p|t, q), \tag{3}$$

where $t = (t_1, \cdots, t_k)$. In (3), we first identify the required core techniques t via $\pi_\theta(t|q)$ before generating the proof p. The proof p is then generated conditioned on the identified techniques t via $\pi_\theta(p|t, q)$. Consequently, the performance of π_θ depends critically on how well the first factor $\pi_\theta(t \mid q)$ is learned. In the following sections, we construct a well-curated dataset and propose a novel multi-stage method to optimize θ such that both $\pi_\theta(t \mid q)$ and $\pi_\theta(p \mid t, q)$ are trained effectively, which corresponds to the acquisition phase. Subsequently, during the application phase, we apply (3) for proof generation.

Furthermore, we can consider a more refined generation process:

$$\pi_\theta(t, s, p|q) = \pi_\theta(t|q)\pi_\theta(s|t, q)\pi_\theta(p|s, t, q), \tag{4}$$

where s denotes a proof sketch generated conditioned on the identified core techniques t. In this scenario, the proof p is generated conditioned on both the core techniques t and the sketch s. This hierarchical decomposition will make the model first

commit to a high-level thinking structure before generating a full proof, reflecting a natural human problem-solving process. Furthermore, we follow the dual phases of insight acquisition and application for effectively generating mathematical proofs. Next, we show how to instantiate these processes in practice via more concrete procedures.

5 Training an Insightful LLM

5.1 Dataset Construction

We construct a hierarchical dataset to train an insightful LLM that can explicitly identify and subsequently generate the core techniques to guide the proof of a given problem. We design the dataset to provide a supervision signal that mirrors human review and consolidation. To ensure high quality and sufficient difficulty, we build our data based on the DeepTheorem (Zhang et al., 2025b) dataset, a recently introduced, large-scale resource for informal mathematical theorem proving.

Base Dataset. DeepTheorem (Zhang et al., 2025b) provides a robust foundation of 121K high-quality informal theorem-proof pairs at roughly IMO-level difficulty. It is constructed through a rigorous pipeline, including collection from diverse corpora and strict decontamination against major benchmarks (e.g., MATH (Hendrycks et al., 2021), AIME (AIME Dataset), miniF2F (Zheng et al., 2022)). Since DeepTheorem already covers many commonly used datasets for theorem-proving training and is built through a rigorous, high-quality construction pipeline, using it as the base dataset allows our subsequent hierarchical annotation to inherit these desirable properties, resulting in a curated dataset with strong overall quality. We refer readers to Zhang et al. (2025b) for further details.

DeepInsightTheorem. Since LLMs struggle to extract core techniques on their own (Section 4.1), we augment model training with a richer supervision signal by explicitly providing core techniques. Building upon DeepTheorem's theorem-proof data pairs (q, p), we perform additional annotations to transform each proof into a hierarchical representation as follows:

$$\big(q,\ \underbrace{(t_1, \ldots, t_m)}_{\text{core techs}},\ \underbrace{s}_{\text{sketch}},\ \underbrace{p}_{\text{proof}}\big).$$

This process involves a meticulous, LLM-assisted analysis to extract deep information from proof p, following a dedicated prompt to design each component as follows:

- **Core Techniques t_i.** Instead of only listing the core techniques, we first include a guiding statement as "`Let's analyze the conditions...`", then those core techniques are introduced with heuristic language, for example: "`The condition ... tells us to construct an auxiliary function...` " or "`... suggests using xxx Theorem.`". The core techniques are themselves finally summarized at the end of this component according to those three main classes.
- **Proof Sketch s.** We insert an intermediate component between the core techniques and the full proof. This sketch outlines a proof strategy based on the identified core techniques, bridging high-level insights and low-level derivations to encourage a smooth and coherent reasoning flow.
- **Proof p.** The original, detailed proof from the base dataset serves as the ground-truth instantiation.

We refer to the resulting dataset as **`DeepInsightTheorem`**, highlighting the insight-driven hierarchical structure derived from DeepTheorem. See the top of Figure 1 for an overview, and Figure 6 for a concrete example.

Chain-of-Thought in Insight. A key design consideration in the construction of `DeepInsightTheorem` is to avoid presenting the core techniques as an isolated list. Instead, we frame insight generation with a short, analytical preamble that serves as a localized micro chain-of-thought for analyzing the question. This encourages a more accurate insight prediction than directly outputting a list of techniques.

Acquisition-Application Transition. There exists a subtle but crucial logical shift between acquisition and application phases, even though the two processes are basically similar. The former is backward-looking: "`Analyzing this proof, we see that xxx is the core tech.`", while the latter is forward-looking: "`Given these problem conditions, xxx might be a core tech.`" These two processes by our design are integrated into our LLM-assisted data annotations. Therefore, we need to carefully tune the prompt for the LLM to ensure the smoothness of the transformation. The specific prompt shows how the two processes are integrated, as for each problem, this prompt on one hand encourages the LLM to analyze the corresponding proof through summarization and

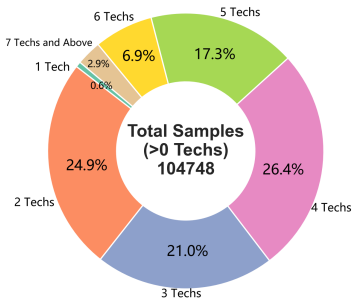

Figure 3: Distribution of core technique counts

extract the core techniques as required, and on the other hand shifts perspective to play the role of an expert encountering the problem for the first time, integrating these core techniques into the thought process of analyzing the problem. The prompts used are detailed in Appendix H.

Key Statistics. Our curated dataset contains approximately 100K problems. On average, each proof involves 3.6 core techniques, with more complex problems frequently combining four or more, as illustrated in Figure 3. A distribution of core technique counts is shown in Figure 10.

5.2 Progressive Multi-Stage SFT

The next step concerns applying the core techniques summarized above. Thanks to the hierarchical structure of `DeepInsightTheorem`, the model can already be trained to produce hierarchical outputs, generating a core technique analysis first (Section 5.1). Then the decompositions in (3) naturally suggest a two-phase training paradigm: the acquisition phase trains the LLM on the original theorem-proof pairs (q, p), while the application phase continues training on the insight-induced data.

We empirically demonstrate the necessity of this two-phase design by comparing training solely on `DeepInsightTheorem` against the baseline. Results for Qwen2.5-7B and Llama3-8B are reported in Table 1. The comparison indicates that directly training on `DeepInsightTheorem` does not necessarily allow the LLM to benefit from the insight structure. One of the reasons is that a higher value of (3) assumes a training on theorem-proof data to enlarge $\pi_\theta(p|q)$ and $\pi_\theta(r_i|r^{\leq i-1}, t, q)$ respectively. Without first training on theorem-proof pairs, the model may lack the basic mathematical reasoning and proof-writing competence required to exploit technique-level guidance. From a learning perspective, this can also be viewed as a mismatch with the natural progression of human learning: teaching

Methods	FIMO	Putnam	HMMT
Qwen2.5-7B	15.73	37.01	12.59
Llama3-8B	12.50	36.69	9.98

Table 1: Evaluation results for models solely trained on `DeepInsightTheorem`

a novice the expert's high-level reasoning pattern before solidifying foundational proof-writing skills can create a substantial comprehension gap.

Inspired this analysis, we adopt a multi-stage training strategy for fine-tuning LLMs. Specifically, we propose a three-stage curriculum that progresses from direct proof generation to sketch-conditioned generation and finally technique-guided reasoning, which we refer to as *Apprentice*, *Journeyman*, and *Expert*. Below we outline the main stages of our progressive multi-stage SFT, each conducted on the same dataset but using different components of `DeepInsightTheorem`:

- **Stage 1 (Apprentice):** (q, p)**.** We first train the model on plain theorem-proof pairs to acquire basic proof-generation ability. This stage corresponds to a novice learner studying complete solutions.
- **Stage 2 (Journeyman):** (q, s, p)**.** We then train the model on theorem-sketch-proof triples, encouraging it to condition proof generation on a high-level sketch. This introduces an intermediate, coarse-to-fine abstraction and reflects a more advanced learner who can follow and internalize the step-wise logical structure rather than merely reproducing surface forms.
- **Stage 3 (Expert):** $(q, \{t_i\}, s, p)$**.** Finally, we train on the full hierarchical structure, so that the model learns to predict core techniques from the problem and leverage them to derive the sketch and the final proof. This stage explicitly learns a direct mapping from the question to its core techniques, resembling an expert's technique-driven reasoning paradigm.

This multi-stage training strategy is illustrated in Figure 1 bottom. Overall, the progressive multi-stage design structurally decouples the learning objectives, yielding a clear and focused target for each training stage. Each stage's training primarily emphasizes the mapping from the problem to the newly introduced component (e.g., p, then s, then $\{t_i\}$). Since each mapping is not yet adequately learned by the model, continued training does not

cause overfitting.

5.3 Insight Policy Optimization

After the progressive multi-stage SFT, the model has learned to generate proofs in the hierarchical format of `DeepInsightTheorem`. We further propose `InsightPO`, a GRPO-style reinforcement learning algorithm tailored to this framework. The key idea is to align reinforcement learning with our structured data format: instead of assigning only an outcome reward to the entire proof, the verifier evaluates whether the generated insights, proof sketch, and final proof form a coherent hierarchy, providing a hierarchy-aware reward that reflects both the quality of the identified insight and the validity of the resulting proof. This structured reward is then used in group-relative policy optimization, encouraging the model to improve not only final proof writing but also the intermediate insight and planning behavior that supports difficult theorem proving.

6 Experiments

In this section, we conduct experiments on two open-source model families: Qwen2.5 (Qwen Team, 2025) and Llama3 (Llama Team, 2024). For each family, we start from their respective base models.

6.1 Experiment Details

Standard SFT. For each base model, we establish a baseline by performing standard SFT on the question-proof pairs for 3 epochs.

Progressive Multi-Stage SFT. We implement two variants of our proposed progressive training strategy as follows:

- **Full Three-Stage Training**: This variant follows the complete curriculum described in Section 5.2. For a given base model, we first fine-tune it on the theorem-proof pairs for 3 epochs, then on the theorem-sketch-proof for another 3 epochs, and finally on the full hierarchical data for a final 3 epochs.
- **Two-Stage Training**: This is a streamlined variant where we skip the intermediate theorem-sketch-proof stage. The motivation for this design is that the core part of our hierarchical structure is the insight and the conceptual jump from generating a full proof to generating a proof sketch might be relatively small in terms of reasoning difficulty. Therefore, removing the proof sketch part may not affect training and could also help save some token cost.

`InsightPO` Post-Training. Building on top of the three-stage SFT models, we conduct `InsightPO` post-training as described in Section 5.3. We use `DeepInsightTheorem` as the training set and DeepSeek-R1 as the verifier. The verifier evaluates insight quality, logical validity, completeness, and clarity with weights of 30%, 30%, 25%, and 15%, respectively. For each dimension, the verifier outputs a score from $\{0, 0.1, 0.2, \ldots, 1.0\}$; rather than using only the argmax score token, we extract the logits over the 11 score values, apply softmax normalization, and use the probability-weighted expected score $\sum_{i=0}^{10} p_i(i/10)$ as the dimension score before computing the weighted total. We train for 400 steps with batch size 128, group size 16, and a maximum rollout length of 8192 tokens. Following prior work on group-relative policy optimization (DeepSeek-AI, 2025), we do not apply KL regularization. The full verifier prompt is provided in Appendix H, and the test score dynamics during `InsightPO` training are shown in Figure 8 in Appendix C.

Evaluation Protocol. We evaluate generated proofs with an LLM-as-Judge protocol detailed in Appendix D. The automatic evaluation scores final proofs along logical validity, completeness, and clarity, and aggregates judgments from DeepSeek-R1 and o3-mini to reduce single-judge bias. Since our training and evaluation both rely on LLM-based verification, we further conduct a stratified human audit as a calibration and sanity-check mechanism. Human reviewers inspect sampled outputs across model families, benchmarks, and score ranges, checking whether the proof realizes the intended insight and whether any essential derivation is missing.

Moreover, we compare our method with two groups of baselines. The first group contains several open-source theorem-proving or math-reasoning models, many of which have been further enhanced via RL-based post-training; we compare them with our fine-tuned Qwen2.5-7B model. The second group contains three structural mathematical reasoning baselines, including SELF-DISCOVER, Least-to-Most Prompting, and Plan-and-Solve Prompting. These baselines are all trained on prompt-proof pairs (q, p).

Models	Methods	FIMO	Putnam	HMMT	Avg.
Qwen2.5-1.5B	Base	12.63	26.66	12.79	17.36
	Two-stage	13.09	30.06	13.42	18.86
	Three-stage	14.35	30.21	15.30	19.95
	DeepInsight	**17.60**	**35.50**	**17.40**	**23.50**
Qwen2.5-3B	Base	12.92	32.94	17.04	24.30
	Two-stage	15.05	35.00	18.88	28.31
	Three-stage	14.99	35.94	20.21	28.71
	DeepInsight	**21.00**	**47.60**	**24.55**	**31.05**
Qwen2.5-7B	Base	16.89	37.76	15.52	23.39
	Two-stage	18.39	45.20	17.76	27.12
	Three-stage	19.31	43.35	18.49	27.05
	DeepInsight	**27.80**	**55.50**	**24.60**	**35.97**
Llama3.2-1B	Base	6.72	16.20	4.25	9.06
	Two-stage	6.91	20.78	6.76	11.48
	Three-stage	8.49	21.55	8.02	12.69
	DeepInsight	**11.20**	**25.30**	**11.00**	**15.83**
Llama3.2-3B	Base	10.46	24.43	8.84	14.58
	Two-stage	11.98	24.69	11.09	15.92
	Three-stage	12.54	27.51	10.77	16.94
	DeepInsight	**17.65**	**36.50**	**14.10**	**22.75**
Llama3-8B	Base	15.61	38.65	13.67	22.64
	Two-stage	16.70	42.76	16.51	25.32
	Three-stage	15.98	41.61	16.64	24.74
	DeepInsight	**22.10**	**50.20**	**20.30**	**30.87**

Table 2: Average evaluation scores of models trained with our methods and the baseline by DeepSeek-R1 and o3-mini.

6.2 Main Results

Our framework achieves better performance. The main results are presented in Table 2 and Table 6. Table 2 shows the average scores from both DeepSeek-R1 and o3-mini, while Table 6 presents the scores from DeepSeek-R1. These results show that after training for insightful hierarchical reasoning, for both variants of our proposed progressive training strategy, the models of different sizes demonstrate superior performance to base models on all three benchmarks. This verifies that, via insightful reasoning, the model's reasoning capability improves.

In addition, Figure 7 in Appendix C reports the best scores for each benchmark, which are pushed higher when incorporating insight, especially for smaller backbones such as the 1.5B model. This implies that insight-guided thinking can help break the reasoning ability ceiling for small models.

InsightPO further boosts performance across all models. As shown in the "DeepInsight" rows in Table 6 and Table 2, InsightPO post-training on top of the three-stage SFT models yields consistent and substantial improvements on all benchmarks. For the Qwen2.5-7B model, InsightPO achieves an average score of 34.32, representing a relative improvement of 32.9% over the three-stage SFT baseline.

Our framework outperforms SOTA models. The comparison can be seen in Table 5 in Appendix C.3. These results demonstrate that, after only SFT, our model already achieves performance that is competitive with those of SOTA models of the same size and extensive post-training, and notably outperforms them after InsightPO training. For the details of evaluation, please see Appendix D.

Comparisons with structured prompting baselines. The results further show that generic decomposition or planning improves over plain SFT but remains consistently below DeepInsight. This suggests that the main gain does not come only from asking the model to produce a plan at inference time, rather, it comes from explicitly learning theorem-specific insight hierarchies during training and then refining them with InsightPO . Detailed settings and results are provided in Appendix C.2.

7 Conclusion

In this paper, we propose DeepInsight, a complete training framework consisting of: (1) DeepInsightTheorem, a hierarchical dataset with explicitly extracted core techniques; (2) a progressive curriculum-style SFT strategy that trains LLMs for insightful reasoning; and (3) InsightPO , a GRPO-style reinforcement learning algorithm with a probability-weighted LLM verifier and hierarchy-aware rewards. Comprehensive evaluations show consistent and substantial improvements over strong baselines on challenging benchmarks.

Limitations

While our insight-aware hierarchy improves infor636 mal theorem proving, several limitations remain. First, our $(q, \{t_i\}, s, p)$ annotations are produced with LLM-assisted decomposition, which may introduce noise; future work will strengthen this pipeline via self-consistency, cross-model agreement, and selective human auditing. Second, our evaluation relies on LLM-as-Judge, which can be prompt-sensitive; we will incorporate hybrid automatic checks (e.g., structure/consistency and domain-specific symbolic validation when applicable). Finally, our experiments cover a limited set of benchmarks and informal proofs cannot guarantee correctness; we plan broader out-of-distribution evaluations and tighter integration with verification-oriented feedback (e.g., checking key subclaims or partially formalized steps).

Ethical Statement

This work enhances LLM capabilities for informal theorem proving to support research and education, though it does not guarantee formal correctness. Because the model can generate convincing yet potentially erroneous proofs, we recommend expert oversight for high-stakes applications and appropriate safeguards against academic dishonesty. Additionally, as our evaluation relies on LLM judges, results should be interpreted with awareness of potential biases and limitations in detecting subtle mathematical flaws.

References

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*.

Kefan Dong, Arvind Mahankali, and Tengyu Ma. 2024. Formal theorem proving by rewarding llms to decompose proofs hierarchically. *Preprint*, arXiv:2411.01829.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google.

Fabian Gloeckle, Jannis Limperg, Gabriel Synnaeve, and Amaury Hayat. 2024. ABEL: Sample efficient online reinforcement learning for neural theorem proving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Runquan Gui, Zhihai Wang, Jie Wang, Chi Ma, Huiling Zhen, Mingxuan Yuan, Jianye Hao, Defu Lian, Enhong Chen, and Feng Wu. 2025. Hypertree planning: Enhancing llm reasoning via hierarchical thinking. *Preprint*, arXiv:2505.02322.

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*.

Jiewen Hu, Thomas Zhu, and Sean Welleck. 2025a. minictx: Neural theorem proving with (long-)contexts. *Preprint*, arXiv:2408.03350.

Jinwu Hu, Dongjin Yang, Langyu Bian, Zhiquan Wen, Yufeng Wang, Yaofo Chen, Bin Xiao, Yuanqing Li, and Mingkui Tan. 2025b. Beyond fast and slow: Cognitive-inspired elastic reasoning for large language models. *Preprint*, arXiv:2512.15089.

Haohan Lin, Zhiqing Sun, Sean Welleck, and Yiming Yang. 2025. Lean-star: Learning to interleave thinking and proving. *Preprint*, arXiv:2407.10040.

Chengwu Liu, Jianhao Shen, Huajian Xin, Zhengying Liu, Ye Yuan, Haiming Wang, Wei Ju, Chuanyang Zheng, Yichun Yin, Lin Li, Ming Zhang, , and Qun Liu. 2023a. FIMO: A challenge for- mal dataset for automated theorem proving. *CoRR*.

Chengwu Liu and 1 others. 2023b. Fimo: A challenge formal dataset for automated theorem proving. *arXiv preprint arXiv:2309.04295*.

Pan Lu, Jiayi Sheng, Luna Lyu, Jikai Jin, Tony Xia, Alex Gu, and James Zou. 2025. Solving inequality proofs with large language models. *Preprint*, arXiv:2506.07927.

Gabriel Poesia, David Broman, Nick Haber, and Noah D. Goodman. 2024. Learning formal mathematics from intrinsic motivation. *Preprint*, arXiv:2407.00695.

Haoran Sun and Shaoning Zeng. 2025. Hierarchical memory for high-efficiency long-term reasoning in llm agents. *Preprint*, arXiv:2507.22925.

AIME Dataset. *AIMO Validation AIME Dataset*. Url: https://huggingface.co/datasets/AI-MO/aimo-validation-aime.

Llama Team. 2024. The llama 3 herd of models. Technical report, AI @ Meta.

Qwen Team. 2025. Qwen2.5 technical report. Technical report, Qwen.

Harvard-MIT Mathematics Tournament. 2025. Hmmt dataset and resources.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Ami tayush Thakur, and Swarat Chaudhuri. 2024. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *Advances in Neural Information Processing Systems 38: Annual Conference*

on Neural Information Process- ing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. 2025. Hierarchical reasoning model. *Preprint*, arXiv:2506.21734.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *Preprint*, arXiv:2305.04091.

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

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

Ziyu Ye, Jiacheng Chen, Jonathan Light, Yifei Wang, Jiankai Sun, Guohao Li, Mac Schwager, Philip Torr, Yuxin Chen, Kaiyu Yang, Yisong Yue, and Ziniu Hu. 2024. Reasoning in reasoning: A hierarchical framework for (better and faster) neural theorem proving. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Ziyin Zhang, Jiahao Xu, Zhiwei He, Tian Liang, Qiuzhi Liu, Yansi Li, Linfeng Song, Zhenwen Liang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2025a. DEEPTHEOREM: Advancing llm reasoning for theorem proving through natural language and reinforcement learning.

Ziyin Zhang and 1 others. 2025b. Deeptheorem: Advancing llm reasoning for theorem proving through natural language and reinforcement learning. *arXiv preprint arXiv:2505.23754*.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations (ICLR)*. ArXiv:2109.00110.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-discover: Large language models self-compose reasoning structures. *Preprint*, arXiv:2402.03620.

A Related Work

Formal theorem proving. Formal automated theorem proving (ATP) relies on proof assistants such as Lean, Coq, and Isabelle (Zheng et al., 2022, Liu et al., 2023a, Tsoukalas et al., 2024), where correctness is enforced by a strict formal system. Recent work integrates large language models with formal proof environments to combine natural-language generation with machine-checkable verification (Gloeckle et al. (2024), Hu et al. (2025a), Lin et al. (2025), Poesia et al. (2024)). Representative benchmarks include miniF2F for cross-system olympiad-level evaluation (Zheng et al., 2022), as well as more challenging datasets that target IMO-/Putnam-style mathematics (Liu et al., 2023b; Tsoukalas et al., 2024). While formal approaches offer strong verifiability, they often face (i) a substantial gap between human mathematical exposition and formal languages, and (ii) a large proof search space that benefits from stronger decomposition, planning, and retrieval.

Informal theorem proving. Informal theorem proving generates proofs directly in natural language and standard mathematical notation (e.g., LaTeX), which aligns well with typical LLM pretraining. NaturalProofs constructs natural-language theorem–proof corpora (Welleck et al., 2022), and NaturalProver studies grounded proof generation in reference/retrieval settings (Welleck et al., 2022). Tencent's DeepTheorem further advances IMO-level informal proving with large-scale data and reinforcement-learning-style training recipes (Zhang et al., 2025b). Lu et al. (2025) focuses specifically on inequality proofs. Despite this progress, many systems remain largely end-to-end and do not explicitly identify or control the core techniques that drive a proof.

Hierarchical Reasoning. Hierarchical reasoning for LLMs decomposes complex problems into manageable sub-problems, enhancing reasoning accuracy and efficiency in complex scenarios. HyperTree Planning (Gui et al., 2025) constructs a structured, high-level outline before generating details. ReasonFlux (Yang et al., 2025) goes beyond raw text generation by scaling "thought templates". Plan-and-Solve Prompting (Wang et al., 2023) first asks the model to form an explicit plan before producing the final answer, while Least-to-Most Prompting (Zhou et al., 2023) decomposes a difficult problem into simpler subproblems and solves them sequentially. SELF-DISCOVER (Zhou et al., 2024) enables LLMs to compose reasoning structures from atomic modules. CoGer (Hu et al., 2025b) draws on Bloom's taxonomy to classify query complexity into four levels and trains an agent via reinforcement learning to dynamically select appropriate reasoning strategies. Another line of work lies in modifying model architectures to support hierarchical processing such as Wang et al. (2025), Sun and Zeng (2025).

In addition, several works focus on hierarchical reasoning for theorem proving. For example, Dong et al. (2024) proposes a reinforcement learning-based training algorithm that incentivizes LLMs to hierarchically decompose theorems into lemmas. Ye et al. (2024) decomposes complex theorem-proving tasks into small, achievable subgoals to abstract formal proof steps.

The key idea in these works is to decompose a complex problem into smaller parts, which corresponds to the proof-sketch component (Section 5.1) in our work. Our hierarchical method instead identifies core techniques, operating at a higher level than a sketch. The work most similar to ours is ReasonFlux (Yang et al., 2025). However, it depends on constructing and retrieving from a template library for theorem proving, which does not endow the LLM itself with intrinsic hierarchical thinking capabilities. In our work, the LLM has been trained to develop insightful thinking, which essentially empowers the LLM with high-level thinking patterns.

B Proof Details Omitted in Section 4.1

Recall that $\pi_{\text{TP}}(p|q)$ is the probability of π_{TP} generating a valid proof p for question q. The general form of a proof is shown in (1). We let k be the number of core techniques in a proof p. Since we assume $\pi_{\text{TP}}(t_i|u_i, q) < \delta$ for all $i \leq k$, by Bayes' rule, we have

$$\begin{aligned}
&\pi_{\text{TP}}(p|q) \\
&= \prod_{i\leq k} [\pi_{\text{TP}}(r_i|t_{i-1}, u_{i-2}, q)\pi_{\text{TP}}(t_i|u_i, q)] \\
&\leq \prod_{i\leq k} \delta \cdot \pi_{\text{TP}}(r_i|t_{i-1}, u_{i-2}, q) \qquad (5) \\
&\leq \delta^k, \qquad (6)
\end{aligned}$$

where the prefix u_i is defined by $u_i := (r_1, t_1, \cdots, r_i)$. Note that (5) is obtained by the assumption $\pi_{\text{TP}}(t_i|u_i, q) < \delta$ for all $i \leq k$ and (6) is due to $\pi_{\text{TP}}(r_i|t_{i-1}, u_{i-2}, q) \leq 1$.

O3-mini's Evaluation-DeepSeek-R1

1. **mixed**
2. **mixed**
3. **incomplete**

Below is an explanation for each answer:

1. **Mixed:**
   The insight mentions several techniques—substitution (setting $y = 1$ or $x = 1$), exploring injectivity/surjectivity, and even hinting at Cauchy-type ideas as a route to establishing linearity. However, these ideas remain quite standard for this type of functional equation problem. They are useful, but the answer doesn't pinpoint any single, non-trivial transformation or method that clearly stands out as a breakthrough insight; it rather relies on a collection of standard tactics. Thus, while not entirely shallow, the ideas include both non-shallow observations and standard, almost routine deductions.

2. **Mixed:**
   The insight does provide some accurate general directions, which are key in many functional equation proofs. However, the description remains somewhat vague and does not detail enough concrete, non-trivial steps. This mix of accurate generalizations and lack of detailed construction leads to a "mixed" evaluation.

3. **Incomplete:**
   While the insight mentions common techniques, it does not capture all potential core ideas needed to solve the problem. There could be additional subtleties—such as a rigorous step-by-step approach to prove that f must be affine or specific manipulations that pivot the equation to reveal the structure $f(x) = x + 1$. Therefore, the core techniques mentioned are not fully comprehensive to cover all aspects of the solution; they provide an initial guide but remain incomplete.

Figure 4: o3-mini's evaluation response for the insight generated by DeepSeek-R1

O3-mini's Evaluation-Gemini2.5-Flash

1. **mixed**
2. **simple scratch**
3. **incomplete**

Below is the explanation for each:

1. **Mixed:**
   The insight mentions exploring properties like injectivity, surjectivity, particular substitutions (e.g., $x = 1$, $y = 1$), and fixed points, and notes that the structure of the equation suggests multiplicative features. While these are useful starting points, they represent standard techniques in functional equation problems. They do hint at a deeper structure but do not provide a concrete key step that addresses the heart of obtaining $f(x) = x + 1$. Thus, there exists a mix of non-trivial ideas combined with basic, common observations.

2. **Simple Scratch:**
   The insight lists several exploratory steps but does so in a rather vague manner without a precise roadmap tailored to this particular question. The ideas are on the right track but lack the detail needed for a detailed, step-by-step core technique application.

3. **Incomplete:**
   While the insight addresses several common techniques for approaching functional equations, it does not cover all the essential steps needed to rigorously derive $f(x) = x + 1$. Additional deeper reasoning, such as deducing specific values (like evaluating $f(1)$) and showing uniqueness or employing clever substitutions that eventually simplify the functional relationship, is necessary. Thus, the list of techniques provided is not comprehensive enough to fully solve the problem without filling in significant logical and algebraic deductions.

Figure 5: o3-mini's evaluation response for the insight generated by Gemini 2.5 Flash

Question:
Let X be a non-negative random variable and $g : [0, \infty) \to [0, \infty)$ be a monotone nondecreasing function such that $g(x) \to \infty$ as $x \to \infty$. If $g(X)$ is integrable, i.e., $E[g(X)] < \infty$, prove that $\lim_{x\to\infty} g(x)P(X > x) = 0$.

Response:
<tech>
Let's analyze the conditions in this question. The monotonicity and divergence of g naturally suggest linking the tail event $\{X > x\}$ to the growth of $g(X)$, forming a crucial inclusion that leverages g's behavior. The integrability of $g(X)$ prompts Markov's inequality to bound the tail probability by an expectation, exploiting the non-negativity and finite mean. The pointwise convergence to zero, combined with the integrable bound $g(X)$, calls for the Dominated Convergence Theorem to justify limit-interchange, which is indispensable since the expectation structure captures the asymptotic decay. Therefore, the potential techniques are summarized as:
<construction>: Event inclusion $\{X > x\} \subseteq \{g(X) \geq g(x)\}$ via monotonicity of g
<theorem call>: Markov's inequality and Dominated Convergence Theorem
<transformation>: no
</tech>

<sketch>
1. Use the monotonicity and divergence of g to establish $\{X > x\} \subseteq \{g(X) \geq g(x)\}$, implying $P(X > x) \leq P(g(X) \geq g(x))$.
2. Apply Markov's inequality to the nonnegative integrable $g(X)$, yielding $g(x)P(X > x) \leq E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}]$.
3. Observe pointwise convergence $g(X)\mathbf{1}_{\{g(X)\geq g(x)\}} \to 0$ due to $g(x) \to \infty$, and dominate by $g(X)$.
4. Invoke the Dominated Convergence Theorem to show $E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}] \to 0$.
5. Conclude $\lim_{x\to\infty} g(x)P(X > x) = 0$ via the squeeze theorem on $0 \leq g(x)P(X > x) \leq E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}]$.
</sketch>

<proof>
Since g is monotone nondecreasing, for any $x \geq 0$, the condition $X > x$ implies $g(X) \geq g(x)$. This gives the event inclusion:

$$\{X > x\} \subseteq \{g(X) \geq g(x)\},$$

which, by monotonicity of probability, yields:

$$P(X > x) \leq P(g(X) \geq g(x)).$$

As $g(X)$ is nonnegative and $E[g(X)] < \infty$, Markov's inequality for $a = g(x) > 0$ (for large x since $g(x) \to \infty$) states:

$$P(g(X) \geq g(x)) \leq \frac{E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}]}{g(x)}.$$

Multiplying by $g(x)$ and using the earlier inequality, we obtain:

$$g(x)P(X > x) \leq g(x)P(g(X) \geq g(x)) \leq E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}].$$

Fix $\omega \in \Omega$. As $x \to \infty$, $g(x) \to \infty$ so $g(x) > g(X(\omega))$ for large enough x, forcing $\mathbf{1}_{\{g(X(\omega))\geq g(x)\}} = 0$. Thus:

$$\lim_{x\to\infty} g(X(\omega))\mathbf{1}_{\{g(X(\omega))\geq g(x)\}} = 0, \quad \forall\omega.$$

Moreover, $0 \leq g(X)\mathbf{1}_{\{g(X)\geq g(x)\}} \leq g(X)$, and $E[g(X)] < \infty$ by hypothesis. By the Dominated Convergence Theorem:

$$\lim_{x\to\infty} E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}] = E\left[\lim_{x\to\infty} g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}\right] = E[0] = 0.$$

Finally, sandwiching with:

$$0 \leq g(x)P(X > x) \leq E[g(X)\mathbf{1}_{\{g(X)\geq g(x)\}}] \xrightarrow{x\to\infty} 0,$$

we conclude $\lim_{x\to\infty} g(x)P(X > x) = 0$ as required.
</proof>

Figure 6: A data example from DeepInsightTheorem

C Additional Experimental Results

C.1 Experiments in Section 3

We conduct an experiment on how insightful the recent advanced commercial LLMs can be. We choose Gemini 2.5 Flash and DeepSeek-R1 to generate insights of several problems. The problems are chosen from four categories including algebra, analysis, combinatorics and number theory, each contains 10 mathematical competition level questions from Putnam and FIMO dataset. The insight generation prompt is listed in Appendix H titled 'Insightfulness Evaluation'. Then we evaluate these generated insight by o3-mini through the prompt titled 'Insight Evaluation'. The results show that most of the insights are basically listing the general standard techniques according to the conditions in questions and can not grasp the essential approaches. See a randomly chosen evaluation result of the following question. The evaluations for both models are shown in Figure 4 and Figure 5.

Question: Let $\mathbb{R}^+$be the set of positive real numbers. Determine all functions $f : \mathbb{R}^+ \to \mathbb{R}^+$ such that, for all positive real numbers x and y,

$$f(x + f(xy)) + y = f(x)f(y) + 1,$$

the final answer is $f(x) = x + 1$.

Insights generated by DeepSeek-R1: The functional equation involves terms like $f(x + f(xy))$ and $f(x)f(y) + 1$, suggesting the use of **substitution techniques** such as setting $y = 1$ or $x = 1$ to simplify and obtain relations involving $f(1)$. A potential approach is to **explore injectivity/surjectivity** by fixing one variable and comparing expressions, then using **Cauchy-type ideas** after establishing f is linear or affine. The structure $f(x + f(xy)) - f(x)f(y) = 1 - y$ hints at eventually proving f is of the form $ax + b$ and then determining constants via substitution into the original equation.

Insights generated by Gemini 2.5 Flash: The presence of $f(x + f(xy))$ and $f(x)f(y)$ suggests looking for properties like injectivity, surjectivity, or particular values of f. Substituting specific values (e.g., $x = 1$, $y = 1$) or analyzing the fixed points of f might be crucial. The term $f(xy)$ combined with $f(x)f(y)$ often points to multiplicative properties or comparison with Cauchy's functional equation variants.

We see that most of the insights generated by both Gemini 2.5 Flash and DeepSeek-R1 for this question are merely generic analysis, without getting to the essence. The insights generated by Gemini are even less accurate, and can even be regarded as "simple scratch."

C.2 Additional Baseline Settings

We evaluate three additional baselines under the same theorem-proving setting as our main experiments. The full prompts for these baselines are provided in Appendix H, under "Plan-and-Solve Baseline Prompt", "Least-to-Most Prompt", and "SELF-DISCOVER Prompt". For Plan-and-Solve, the model is prompted to first write a concise proof plan and then generate the final proof conditioned on that plan. For Least-to-Most Prompting, the model is prompted to decompose the theorem into a sequence of easier subclaims, solve them in order, and combine the intermediate results into the final proof. For SELF-DISCOVER, the model is prompted to select and compose general reasoning modules into a task-specific reasoning structure before generating the proof.

For fairness, all three baselines start from the same standard SFT checkpoint trained only on prompt-proof pairs (q, p). They are applied only at test time on Qwen2.5-7B and Llama3-8B. All generations use the same benchmark split, decoding budget, proof extraction rule, LLM-as-judge protocol, and human-audit calibration as our main experiments. Table 3 reports the resulting comparison.

C.3 Parameter Settings in Section 6

Here we provide main parameters setting for our experiments. For both model series, we use the learning rate of 2e-5, and training batch size 256. The maximal sequence length is set to be 4096 for all experiments on the base dataset, and 8192 for all on `DeepInsightTheorem`. We also conducted additional experiments with the baseline's max sequence length set to 8192 for completion. See Table 4 for details.

D Evaluation

Since our task involves generating informal mathematical proofs by natural language, we adopt an LLM-as-Judge evaluation protocol. Following the established practice in DeepTheorem (Zhang et al., 2025b), our evaluation is conducted on a set of challenging benchmarks to test the model's reasoning capability. Specifically, we use theorem-proving problems drawn from FIMO (Liu et al.,

Models	Methods	FIMO	Putnam	HMMT	Avg.
Qwen2.5-7B	SFT(q, p) + Plan-and-Solve	22.40	48.70	20.90	30.67
	SFT(q, p) + Least-to-Most	23.65	47.90	17.35	29.63
	SFT(q, p) + SELF-DISCOVER	22.95	51.60	16.70	30.42
	`DeepInsight`	**27.80**	**55.50**	**24.60**	**35.97**
Llama3-8B	SFT(q, p) + Plan-and-Solve	17.35	41.80	12.60	23.92
	SFT(q, p) + Least-to-Most	16.80	44.20	13.75	24.92
	SFT(q, p) + SELF-DISCOVER	19.45	43.60	14.95	26.00
	`DeepInsight`	**22.10**	**50.20**	**20.30**	**30.87**

Table 3: Comparison with additional baselines on 7B/8B models. All baselines use the standard SFT checkpoint trained only on (q, p) and are applied at test time.

Methods	FIMO	Putnam	HMMT	Avg.
Qwen2.5-7B	14.98	36.05	15.24	22.09
Llama3-8B	13.29	38.84	11.99	21.37

Table 4: Evaluation of baseline with max length 8092

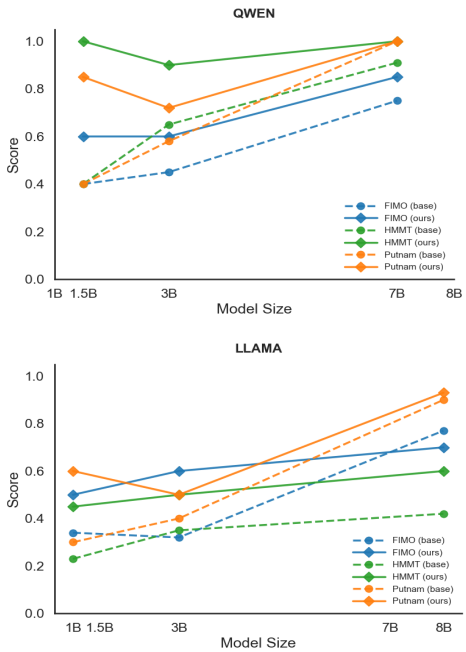

Figure 7: Max scores in evaluation across benchmarks.

2023a), PutnamBench (Tsoukalas et al., 2024), and a newly constructed theorem-proving subset of the Harvard-MIT Mathematics Tournament (HMMT) (Tournament, 2025).

To assess the quality of the generated proofs, we apply the evaluation framework in DeepTheorem but omit a separate "correctness" judgment, which is a fabricated verifiable answer constructed by DeepTheorem that is yet unrelated to the quality of proof itself and hence not applicable to our task. Our evaluation centers on the following three core dimensions of the proof text itself (Zhang et al., 2025b):

- **Logical Validity.** Check whether each step follows logically from the preceding step and indicate any logical errors.
- **Completeness.** Verify whether all necessary steps are included to fully prove the theorem.
- **Clarity.** Assess whether the proof is clear, unambiguous, and well-explained.

We use DeepSeek-R1 (DeepSeek-AI, 2025) as the judge model. For each generated proof, the judge is prompted to analyze and score it on a continuous scale from 0 to 1 for each of the three dimensions above. The final score for a proof is calculated as a weighted average of these three-dimensional scores. The specific weighting scheme and the full prompts used for this evaluation are provided in Appendix H.

Using only one LLM as a judge may introduce bias in evaluation. To address this, we also incorporate o3-mini as another LLM judge and combine the scores from both judges to evaluate the performance of our method. See Table 2 in Appendix C.3 for details.

Human Audit and Calibration. In addition to multi-judge aggregation, we use a lightweight

Models	FIMO	Putnam	HMMT	Avg.
Qwen2.5-Inst-7B	15.29	42.39	17.86	25.12
Qwen2.5-Math-Inst-7B	17.01	41.06	16.57	24.92
DS-Prover-v1.5-RL-7B	17.39	42.00	13.68	24.36
DS-Prover-v2-7B	16.25	41.50	15.43	24.39
Ours (2-stage)	16.33	43.34	15.59	25.09
Ours (3-stage)	18.03	41.67	17.78	25.83
Ours (`DeepInsight`)	**26.10**	**53.70**	**23.15**	**34.32**

Table 5: Evaluation score by Deepseek-r1 comparison against Open-Source Baselines

human-audit protocol to reduce systematic bias in LLM-as-Judge evaluation. We stratify generated proofs by model family, benchmark, and LLM-judge score range, and then sample a small subset for expert inspection. Human reviewers evaluate the same dimensions used by the verifier: insight quality, logical validity, completeness, and clarity. For insight quality, reviewers check whether the generated core techniques capture the key mathematical idea rather than merely restating surface conditions. For proof quality, reviewers check whether the proof sketch and final proof correctly realize the stated techniques and whether any essential derivation is missing. The audited subset is used to calibrate verifier behavior by comparing human scores with LLM scores, checking score deviations across dimensions, and identifying cases where the verifier systematically over-rewards verbose or superficially plausible proofs. As shown in Figure 9, the statistics are computed on a single stratified audit pool aggregated across FIMO, PutnamBench, and HMMT

Finally, since the model trained by our method would generate a formatted output containing three components, we extract only the final proof component for evaluation to ensure a fair comparison with baseline methods.

E Data Construction

As we discussed before, the construction of our data is based on the DeepTheorem (Zhang et al., 2025a) by annotating each piece of data with the assistance of DeepSeek-R1 to help generate the core techniques and corresponding proof sketches. After the generation process, we apply a filtering process to remove some data that failed to be annotated due to some reasons like temporary API calling failure or annotated with an undesired structure. Also we notice some repeated questions in the base dataset hence we also remove those redundant ones.

Finally, we collect 104,751 pieces of hierarchical data with rigorous structure as shown in Figure 6. The distribution of each class of techniques is summarized in Figure 10.

Maximizing Data Utility. A fundamental principle of `DeepInsightTheorem` is its self-contained nature. The creation of the hierarchical structure $(q, \{t_i\}, s, p)$ is achieved solely through the analysis of the information already embedded within the original proofs. By doing so, we increase the informational density of each training example, offering a powerful pathway to improve data efficiency when scaling high-quality mathematical proof data.

F Benchmark Statistics

Table 7 is a simple statistic on the number of questions in three benchmarks we use for evaluation. Note that in Zhang et al. (2025a), they manually expanded each benchmark with variants of each question. Here we do not need such a design.

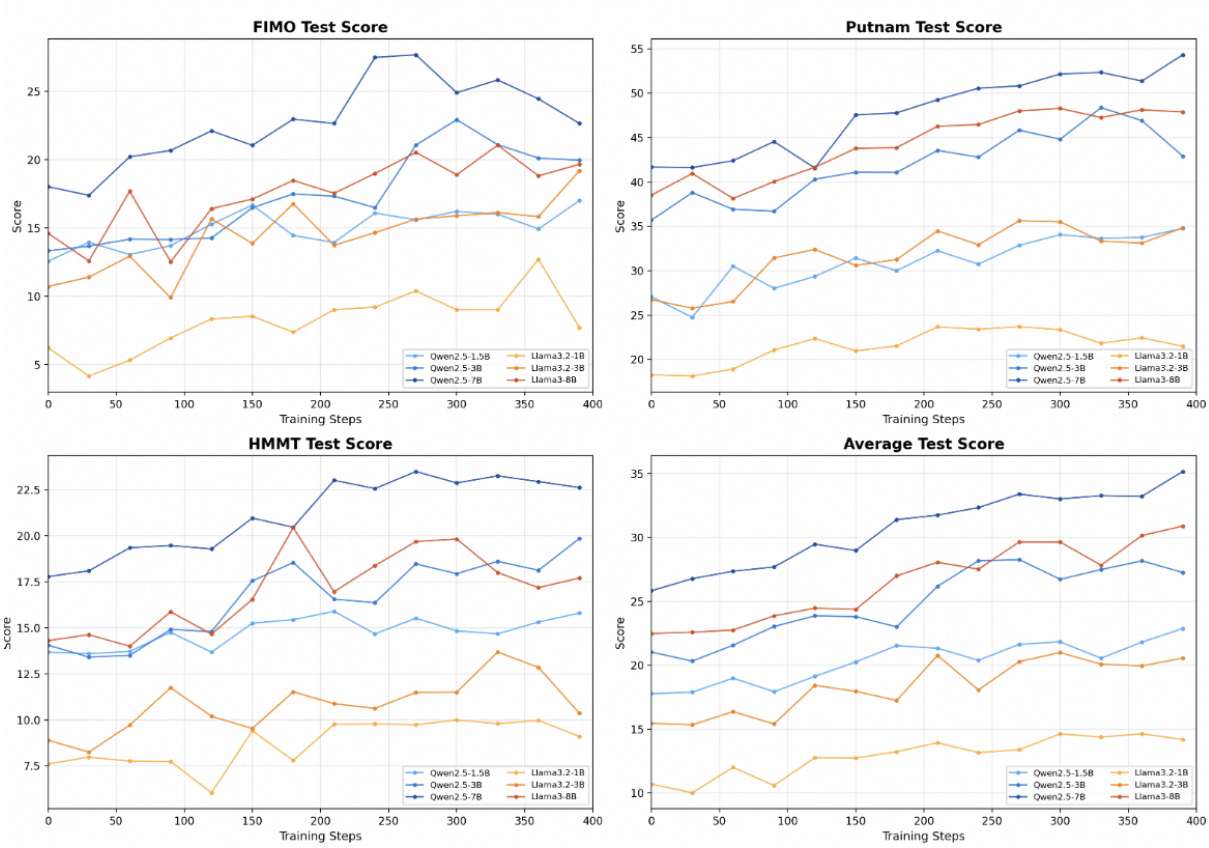

Figure 8: Test score curves during `InsightPO` training.

G Qualitative Case Study

We present a qualitative case study to demonstrate the improved evaluation results. It compares the output generated by the Qwen2.5-7B model trained on `DeepInsightTheorem` with the baseline output for a sample problem from HMMT. The chosen problem is "2025-02-combinatorics-04".

Evaluation of Our Method. We achieved a score of 0.9 for this problem with a minor completeness issue.
Logical Validity: "`The algorithm is correct and would work, no logical flaw in the structure.`"
Completeness: "`the algorithm is complete and the answer is stated, but the computation is omitted. It's partial.`"
Clarity: "`The steps are well-explained and clear.`"

Evaluation of the Baseline. The baseline produces a low-quality proof, receiving a score of 0.19.
Logical Validity: "`The proof is logically invalid due to an incorrect reduction to only right and up moves, a flawed bijection in Step 3 and the contradiction between Step 4 and Step 5.`"
Completeness: "`The proof is incomplete as it misses paths using diagonal moves, does not account for variable path lengths, and fails to properly handle the no-revisit constraint.`"
Clarity: "`The proof has a structured outline, but the marking system in Step 3 is ambiguous, and the numerical inconsistency creates confusion.`"

In summary, our method helps the model establish a solid logical ability and cultivate its insight to plan for the reasoning path.

H Prompt Templates

We present all prompts on the remaining pages.

I Statement on the Use of AI-Assisted Tools

We used AI-assisted tools to support the writing and editing of this manuscript. Specifically, these tools were employed to help refine grammar, improve readability, and enhance the clarity and consistency of the presentation. All technical content including the problem formulation, algorithm de-

Models	Methods	FIMO	Putnam	HMMT	Avg.
Qwen2.5-1.5B	Base	11.17	25.39	11.25	15.94
	Two-stage	11.81	28.72	11.80	17.44
	Three-stage	12.57	27.06	13.68	17.77
	`DeepInsight`	**16.10**	**33.50**	**15.90**	**21.83**
Qwen2.5-3B	Base	11.92	31.87	12.69	18.83
	Two-stage	12.85	34.86	11.98	19.90
	Three-stage	13.34	35.69	14.06	21.03
	`DeepInsight`	**19.50**	**45.60**	**18.40**	**27.83**
Qwen2.5-7B	Base	15.27	36.75	14.82	22.28
	Two-stage	16.33	43.34	15.59	25.09
	Three-stage	18.03	41.67	17.78	25.83
	`DeepInsight`	**26.10**	**53.70**	**23.15**	**34.32**
Llama3.2-1B	Base	5.04	14.61	4.48	8.04
	Two-stage	4.19	18.26	5.78	9.41
	Three-stage	6.25	18.26	7.60	10.70
	`DeepInsight`	**9.70**	**23.30**	**9.50**	**14.17**
Llama3.2-3B	Base	9.29	23.90	7.04	13.41
	Two-stage	11.08	24.73	9.27	15.03
	Three-stage	10.72	26.73	8.89	15.45
	`DeepInsight`	**16.15**	**34.80**	**12.25**	**21.07**
Llama3-8B	Base	13.48	37.68	12.25	21.14
	Two-stage	14.70	40.51	14.15	23.12
	Three-stage	14.62	38.51	14.30	22.48
	`DeepInsight`	**20.60**	**48.50**	**18.70**	**29.27**

Table 6: Evaluation of models trained with our methods and the baseline by DeepSeek-R1

Benchmark	Num.
FIMO	71
Putnam	166
HMMT	76

Table 7: The number of questions in each benchmark

sign, theoretical analysis, experimental setup, results, and conclusions, was developed and verified by the authors. The authors reviewed, validated, and edited all AI-assisted outputs to ensure correctness, originality, and compliance with the venue's policies.

Score range	Samples	LLM verifier	Human audit	Difference pattern
0.0–0.2	22	0.13	0.15	slight under-scoring
0.2–0.4	34	0.32	0.30	close agreement
0.4–0.6	41	0.51	0.46	gap on incomplete sketches
0.6–0.8	36	0.68	0.64	moderate over-scoring
0.8–1.0	27	0.84	0.79	gap on subtle errors

Figure 9: Human-audit calibration of the LLM verifier, aggregated over the stratified audit subset sampled from FIMO, PutnamBench, and HMMT. Average human scores broadly follow verifier scores across score ranges, while deviations are not monotonic: larger gaps appear both for mid-score proofs with incomplete sketches and for high-score proofs with subtle errors. The bar length qualitatively indicates the absolute score gap.

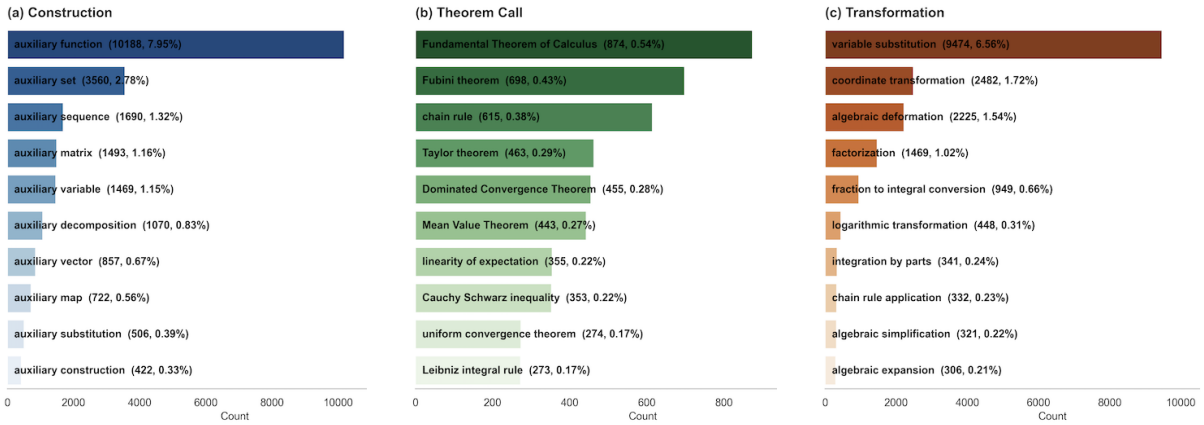

Figure 10: Top technique distribution for each class in `DeepInsightTheorem`

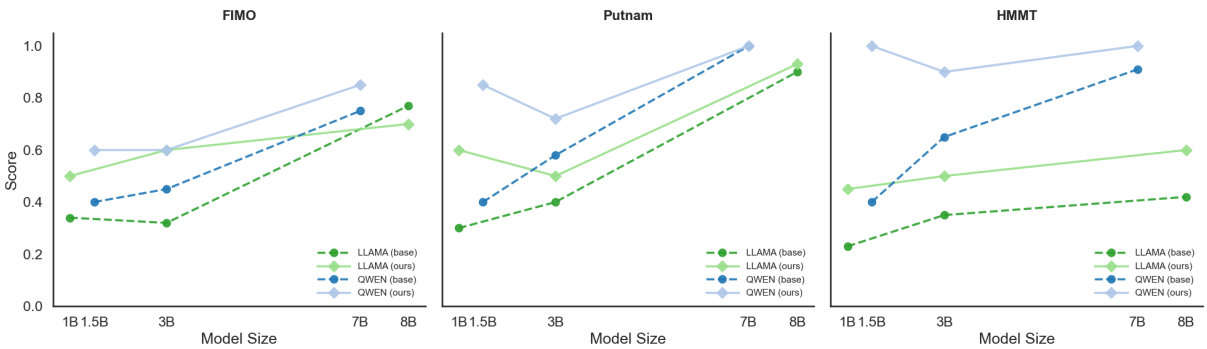

Figure 11: Max scores in evaluation across benchmarks.

Insight Generation

```
Now you are tested the insight this math question:
{question}
From now on you can not go on thinking and expanding the proof step by step.
Instead, try to infer the most potential core techniques that would be used
in this question by ONLY analyzing the conditions in the question. You can
only have a short and quick guess without detailed and long deduction.

The core techniques may refer to specific mathematical construction, use
of known theorem or existing results or smart and subtle mathematical
transformations, instead of those fundamental logic deduction.

Output: Latex formatted core precise techniques in your prediction for around
3 sentences.
```

Insightfulness Evaluation

The core techniques in math proof may refer to specific mathematical construction, use of known theorem or existed results or smart and subtle mathematical transformations for particular question, instead of those fundamental mathematical and logic deduction details. Now you have the following question:
{}.
And a given insight which contains the prediction of core techniques to this question:
{}
Now you need to review and evaluate this insight from the following aspects:

1. see if there exists an idea in this insight that is highly key to the solution, which in general is likely to be found or realized after several thinking steps, or the other steps in the solution are far easier after dealing with this core idea. Note that such idea can not vaguely hints without providing concrete methods and identifying the core non-trivial step. And hence determine whether the insight is shallow quick guess or deep identification.
2. see if the core ideas described in the insight is accurate enough. Check if the idea gives the precise techniques key and adapted particularly to the question. It does not need to be containing details but accurately describe mentioned techniques. And hence determine whether the idea is a simple scratch or an accurate expression.
3. see if the core techniques mentioned in this insight are all core techniques for the question. check whether the question can be solved by filling basic mathematical and logical deductions ONLY under the core techniques in the insight. And hence determine whether the insight is comprehensive or incomplete.

-output:
For 1, if all ideas in the insight are not shallow quick guess, and there are no flaws in the whole insight, then output 'deep identification'. If there are also some ideas that are shallow quick guess, or even there are some flaws in this insight, output 'mixed'. Note that mixed means there indeed exist non-shallow ideas, but also other standard observations. Otherwise, if there are all shallow quick guess or just spread out standard general techniques in the insight without an accurate orientation, then output 'shallow quick guess'.

For 2, if all ideas in the insight are satisfied, then output 'detailed expression'. If all are not satisfied, output 'simple scratch'. otherwise output 'mixed'.

For 3, if all ideas in the insight are enough core ideas for the question, then output 'comprehensive'. If not, output 'incomplete'.

format:

1. 'deep identification'/ 'shallow quick guess'/'mixed'
2. 'detailed expression'/'simple scratch'/'mixed'
3. 'comprehensive'/'incomplete'/'mixed'

Data Construction

Analyze the mathematical problem and solution below:
Problem:
{question}
Solution:
{response}

Perform these tasks:
1. First identify 1-3 core mathematical techniques used in the solution by considering if there are some specific constructions, theorem or existed results calling and smart mathematical transformations, where the smart transformation may not be known results and are subtle and hard to note.

Note that the core mathematical techniques are not those basic logic deductions. They need to be crucial to the solution.

Then Write the analysis for each technique on how a person can realize such technique when reading the question. The analysis should be from several aspects, including how the problem settings or assumptions suggest this technique, how it might be potentially useful to prove the result and why it might be crucial to the whole proof.

The analysis should be like an insightful and experienced math professor's thoughts when he is trying to solve the question. The whole analysis should start with 'Let's analyze the conditions in this question.'

The analysis contains 2-3 sentences for each technique. The analysis should avoid mentioning the solution. Make sure the logic of the analysis is coherent. After analysis, write the extracted techniques following the analysis through 'Therefore, the potential techniques are summarized as...', with three tags:

<construction>: identify the specific construction used in the solution if any;
<theorem call>: specify the theorem or any existed results used in the solution if any;
<transformation>: specify the smart mathematical transformations used in the solution if any.

If there are no such techniques, just specify 'no' after the tag, but you must write the tag <construction>, <theorem call> and <transformation> even there are no such techniques.

Do not write one technique under two tags both. The whole technique analysis, e.g., the analysis and technique extraction, should be with LaTeX inline math ($... $) where appropriate.

2. Create a proof sketch integrating these techniques analyzed from task 1, which serves as a high-level proof organization:
- Format: Numbered steps in LaTeX
- Each step: 1 sentence with key mathematical reasoning
- Include essential formulas in math mode
- Example: '\\begin{{enumerate}}\\item Assume P is countable: $P = \\{{x_1, x_2, \\. . . \\}}$\\end{{enumerate}}'

Data Construction (cont'd)

[t] 3. Based on the original given solution, improve the solution by elaborating each step in the proof sketch, making it well-organized.

4. Output:
  - mathematical techniques: string (LaTeX formatted)
  - proof sketch: string (LaTeX enumerated steps)
  - solution: string (LaTeX formatted)

The mathematical techniques are enclosed within <tech></tech>, the proof sketch is enclosed within <sketch></sketch> and the solution is enclosed within <proof></proof>, respectively, i.e., <tech> mathematical technique analysis here </tech> <sketch> proof sketch here </sketch> <proof> solution here </proof>

Proof Evaluation

```
The following question asks to prove a statement.
The question:
{question}
The test subject's solution:
{response}

Your task is to evaluate the proof's quality and assign a score from 0 to
1 based on three criteria: logical validity (40%), completeness (30%), and
clarity (30%).

Instructions:
1. Analyze the proof step by step.
2. For each criterion:
    - Logical Validity: Check if each step follows logically from the previous
      one. Flag any logical errors.
    - Completeness: Verify if all necessary cases and steps are included to
      prove the theorem.
    - Clarity: Assess if the proof is clear, unambiguous, and well-explained.
3. Assign a sub-score (0 to 1) for each criterion and compute the total score
   using the weights: (0.4 × validity) + (0.3 × completeness) + (0.3 ×
   clarity).
4. Provide a brief explanation (2-3 sentences) summarizing any errors or issues
   and justifying the score.

Output:
 - Your total score: float;
 - Your sub-scores and corresponding brief explanation:
    - Sub-score: float;
    - Brief explanation: string (LaTeX formatted)

The total score is enclosed in <score></score>, and the sub-scores with
corresponding explanation are enclosed in <exp></exp>, e.g.,

- Format:
'<score>
your final scores here. Just write a single number here.
</score>
<exp>
"validity": your sub-score for validity here
explanation: your explanation for validity score here
"completeness": your sub-score for completeness here
explanation: your explanation for completeness score here
"clarity": your sub-score for clarity here
explanation: your explanation for clarity score here
</exp>'
```

InsightPO Verifier Prompt

You are an expert mathematical proof evaluator specializing in insight-driven theorem proving. You will evaluate a model-generated response to a mathematical theorem-proving question.

Question:
{question}
Model's Response:
{response}

Evaluation Framework

The response should follow a hierarchical insight-driven reasoning format containing three components:

- <tech>...</tech>: Core technique analysis — identifying key mathematical techniques by analyzing the problem's conditions
- <sketch>...</sketch>: Proof sketch — a high-level proof plan organized around the identified techniques
- <proof>...</proof>: Full mathematical proof — the complete, rigorous proof

If the response does not contain this hierarchical structure, evaluate only the proof quality and assign 0 to Insight Quality.

Evaluate the response across four dimensions:

1. Insight Quality (Weight: 30%)
Evaluate the core technique analysis in the <tech> section. Core techniques refer to specific mathematical constructions, invocations of known theorems or existing results, or clever mathematical transformations — NOT basic logical deductions.

Check: Does the analysis begin by examining the conditions? Are the identified techniques specific, non-trivial, and genuinely crucial? Are techniques properly categorized under <construction>, <theorem call>, or <transformation>? Does the analysis demonstrate genuine mathematical intuition?

Scoring: 0.0–0.2: No meaningful techniques identified; 0.3–0.4: Only generic techniques; 0.5–0.6: Some correct techniques but missing key ones; 0.7–0.8: Most techniques correctly identified; 0.9–1.0: All essential techniques with deep analysis.

2. Logical Validity (Weight: 30%)
Evaluate mathematical rigor in the <proof> section. Check logical flow, justification of claims, absence of errors or circular reasoning.

Scoring: 0.0–0.2: Logically invalid; 0.3–0.4: Major logical errors; 0.5–0.6: Some gaps but sound direction; 0.7–0.8: Minor issues, fundamentally correct; 0.9–1.0: Every step rigorous.

3. Completeness (Weight: 25%)
Evaluate whether the proof fully establishes the theorem. Check coverage of all cases, edge conditions, and sufficiency of steps.

Scoring: 0.0–0.2: Absent or restates problem; 0.3–0.4: Significant portions missing; 0.5–0.6: Partial proof; 0.7–0.8: Nearly complete; 0.9–1.0: Fully complete.

InsightPO Verifier Prompt (cont'd)

Scoring Instructions

4. Clarity (Weight: 15%)
Evaluate presentation quality, notation consistency, and coherent flow of the hierarchical structure.

Scoring: 0.0–0.3: Incomprehensible; 0.4–0.6: Somewhat clear; 0.7–0.8: Clear with minor issues; 0.9–1.0: Exceptionally clear.

IMPORTANT: The final score MUST be exactly one of these 11 discrete values: {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. No intermediate values are permitted.

1. Analyze the response carefully, step by step.
2. Assign a sub-score for each dimension from {0, 0.1, 0.2, ..., 1.0}.
3. Compute the weighted total: raw $= 0.30\times$ insight_quality $+0.30\times$ logical_validity $+0.25\times$ completeness $+0.15\times$ clarity.
4. Round the raw score to the nearest value in {0, 0.1, ..., 1.0} to obtain the final_score.

Output Format

'<score>
{final_score — must be one of: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
</score>
<exp>
"insight_quality": {sub-score}
explanation: {1-2 sentence justification}
"logical_validity": {sub-score}
explanation: {1-2 sentence justification}
"completeness": {sub-score}
explanation: {1-2 sentence justification}
"clarity": {sub-score}
explanation: {1-2 sentence justification}
</exp>'

Plan-and-Solve Baseline Prompt

You are given a mathematical theorem-proving problem.
Problem:
{question}

First devise a concise proof plan. The plan should divide the theorem into necessary steps, identify useful intermediate claims, and specify how these claims will be connected. Do not assume access to any external proof sketch or annotated core techniques.

Then solve the problem rigorously by following your plan. The final response should contain only the complete mathematical proof after the plan has guided your reasoning.

Output the final proof enclosed by <proof></proof>.

Least-to-Most Prompt

You are given a mathematical theorem-proving problem.
Problem:
{question}

Decompose the problem into a sequence of simpler subproblems or subclaims ordered from easiest to hardest. Each later subproblem may use the conclusions of previous subproblems. The decomposition must be generated only from the problem statement, without using any annotated proof sketch or core technique labels.

Solve the subproblems one by one, and then assemble their conclusions into a complete proof of the original theorem. Make sure every intermediate claim used in the final proof is justified.

Output the final proof enclosed by <proof></proof>.

SELF-DISCOVER Prompt

You are given a mathematical theorem-proving problem.
Problem:
{question}

Select useful reasoning modules from the following general set: identify assumptions, infer hidden constraints, search for relevant lemmas, construct auxiliary objects, decompose into subclaims, prove intermediate claims, check edge cases, and assemble a final proof.

Compose the selected modules into a task-specific reasoning structure for this problem. The structure should guide the proof process but must not rely on any annotated core techniques, proof sketches, or training-time hierarchy labels.

Use the self-composed reasoning structure to generate a rigorous complete proof. Ensure that the final proof is logically valid, complete, and clearly written.

Output the final proof enclosed by <proof></proof>.