細分化と要約を通じた数学証明の詳細度制御

服部 清志　松崎 拓也
東京理科大学大学院
1424521@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

概要

数学証明の推論の詳しさは、同じ命題に対する証明であっても書き手によって大きく異なり、これはしばしば読み手の理解を阻害する要因となる。本研究では、読み手の理解力に合わせた多様な粒度の自然言語証明を生成するため、3 種類に分類された「行間」に注目した証明の推論細分化手法、および細分化された証明の推論を指定された詳細度に応じて要約することによる行間調整手法を提案する。また、実解析の命題およびその自然言語証明から構成されるデータセットに対して本手法を適用し、自動評価の結果を分析した。その結果、提案手法による推論の細分化は、有意義かつ元の命題に忠実な推論を補完できることが確認された。一方で、要約による行間調整は、人間が記述する程度の詳しさをもつ証明に対しては適切に機能するものの、大きく縮約する場合には課題が残ることが示された。

1　はじめに

自然言語証明では、細かい推論過程や計算過程は省略されることが多い。しかし、何を「細かい」と見做すかは書き手が持つ知識や書き手が想定する読み手のレベルなどの様々な要因に依存して決まるため、同じ命題であっても証明ごとに詳細度は異なりうる。そのため、読み手と書き手との間で「細かい」と認識している操作に差があると、その違いは、説明が不足している場合には論理的な飛躍、すなわち「行間」として現れ、逆に説明が過剰な場合には冗長な操作として現れる。これらはいずれも読み手の理解を阻害する原因となる。

近年、大規模言語モデル (LLM) の研究及び発展に伴い、LLM が自然言語で数学の推論過程を生成する能力を十分に持つことが示されている [1, 2]。この能力を応用し、自然言語証明の詳細度を自由に調整することができれば、数学的推論データセットの生成や、教育分野への応用、自然言語証明を機械検証可能な形式証明へと変換する自動形式化分野への応用など、多方面での応用が期待できる。

しかし、LLM はハルシネーションを引き起こすリスクを抱えており、特に数学証明のような複雑な推論を要する場面では、誤った推論や意味のない推論を生成する傾向が強い [3]。さらに、推論の詳細度制御には証明全体の構造や文脈を踏まえた判断が必要であるため、推論を適切に補完・削減しつつ論理的に正しく再構成するというタスクは、現在の LLM にとっても極めて複雑かつ困難であろう。

これを解決するため、本研究では、「行間」を 3 種類に分類し、それらに対し段階的に推論を補っていくことによる、ハルシネーションのリスクを抑制した自然言語証明の推論詳細化手法をまず提案する。さらに、推論の詳しさの調整をマニュアルに基づいて立案を行うモデルと、実行を行うモデルの二つに分けて構成し、これら 2 モデルを用いて細分化された証明における行間を調整する手法を提案する。また、実解析の命題およびその自然言語証明から構成されるデータセットに対して本手法を適用し、その有効性を実験的に検証する。

2　関連研究

Welleck ら [4] は、与えられた命題に関連する定理や定義を検索モデルによって取得し、その定理名から内容を再構築したのち、それらをリファレンスとして推論ステップを生成する手法を提案している。この研究ではあくまで正しい推論の鎖を生成することが目的であるのに対し、本研究は生成された推論の詳細度を自由に制御することを目標としている、という点が異なる。

Chen ら [5] はアテンション機構に 5 段階の階層を導入し、それぞれの階層で異なる数学的構造に注目させることで数学証明が持つ「階層的な性質」を LLM に反映し、形式証明の生成能力を向上させる手法を提案した。本研究も数学証明の持つ階層的な構造には注目するが、アテンションを通じて間接的

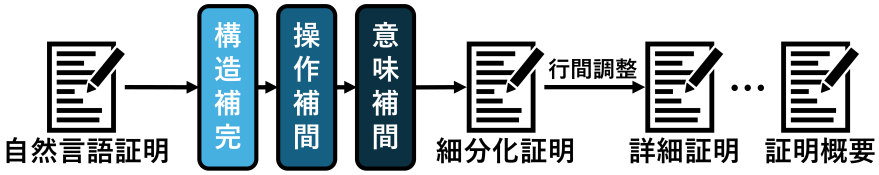

図 1　提案手法の概要図

に階層構造を捉えるのではなく、LLM を用いて証明内の階層構造を直接解析する。また、本研究は自然言語証明のみを対象としていること、数学的な階層を推論の精度向上ではなく詳細度の調整のために用いる点で異なる。

3　提案手法

本節では、自然言語証明の行間調整手法について具体的に述べる。まず、3.1 節で「行間」の分類を示す。次に、3.2 節では自然言語証明の推論過程の細分化手法について述べ、3.3 節では細分化証明の行間調整手法について述べる。提案手法の全体像を図 1 に示す。

3.1　証明の行間の分類

本研究では、行間を「説明量の不足によって生じるズレ」と捉える。既存の数学証明の観察に基づき、行間を以下のように分類する。

- **暗黙的な構造の導入**：証明の流れから明らかに読み取れるが、明示的に述べられていない構造的な要素（場合分け、背理法の開始など）
- **計算過程**：書き手にとって自明、ないし冗長であると判断された数式の変形
- **定理・補題の適用**：書き手が暗黙的に適用しても理解可能であると判断した定理・補題の適用
- **前提条件・仮定の適用**：その命題における前提条件、もしくは証明中で仮定された条件の適用
- **対称性・一般性を用いた証明**：証明が一部の推論のみを示し、残りを一般性や対称性を用いて示す場合（同様にして示せる等）
- **自明性**：「これは明らか」といった表現に代表される、書き手が意図的に省略した推論過程

さらに、各種行間を自動的かつ段階的に補う処理の特性に伴い、これらを**構造的な行間**、**操作的な行間**、**意味的な行間**の 3 つに分類する。構造的な行間は証明の構造に関わる行間であり、暗黙的な構造の導入がこれに該当する。操作的な行間は計算や定理適用などの具体的な操作に関わる行間であり、計算過程、定理・補題の適用、前提条件・仮定の適用の 3 つが該当する。意味的な行間は証明の意味や内容に関わる行間であり、対称性・一般性を用いた証明及び自明性がこれに該当する。本研究では、これら 3 種類の行間の特徴を考慮しながら、推論過程を細分化する手法を開発した。

3.2　証明の細分化

ここでは、自然言語証明の推論過程を細分化する手法について述べる。前項で述べた行間の種類に基づき、以下の 3 つの段階を通じて証明の行間補完及び細分化を行う。

構造的な行間　構造的な行間を補完するため、入力となる自然言語証明に対して、以下の 5 つの構造のスパンを検出し、そのスパンに基づいて証明の階層構造を同定する。なお、本稿では、証明中の連続する文の列をスパンと呼ぶ。スパンはそれぞれ XML タグ（<tag></tag>）の形で表され、文が特定のスパンに属しているかどうかはその文がタグに挟まれているかどうかで表現される。

- **場合分け**：証明が特定の変数ないし条件に基づいて複数の場合に分かれている場合、その場合分けが行われているスパン全体及び各場合の証明それぞれのスパンを検出する。
- **背理法**：背理法を用いた証明において、仮定から矛盾に至るまでのスパンを検出する。
- **帰納法**：帰納法を用いた証明において、その帰納法が行われているスパン全体及び基底ケースと帰納ステップそれぞれのスパンを検出する。
- **十分性**：証明が十分性（A を示すには B を示せば十分である）を用いている場合、十分性を示しているスパン及びその十分条件を証明する部分のスパンを検出する。
- **部分命題**：証明の中で部分命題が導入されている場合、その部分命題の宣言から証明完了に至るまでのスパンを検出する。

これらのスパンを抽出するために、各構造ごとの文表現の特徴を明示したプロンプトを用い、LLM

```
<proof>
  <statement>We will show that ...</statement>
  <by_contradiction>
    <step>
      Assume, for the sake of contradiction, ...
    </step>
  </by_contradiction>
</proof>
```

図 2　証明の構造化例

によって各構造のスパンを文のインデックスの区間（$[a, b]$ の形）で抽出したのち、それを XML 形式にルールベースで変換することで構造を表現した。このツリーは <proof>タグを根とし、命題を表す<statement>タグ及び構造を表す各種タグによって表現される。具体例を図 2 に示す。

また、暗黙的な構造の導入については、XML 形式に変換した証明を LLM に入力し、本文内の情報から適切な構造の導入文を生成させることで補完を行った。

操作的な行間　操作的な行間の補完については、上で述べた構造的な行間の補完を行った後の、階層構造を明示した証明に対して、計算過程、定理・補題の適用、前提条件・仮定の適用に関する行間を補完する。具体的には、各文について、その文が数式の変形を含む場合は計算過程の補完を行い、その後、定理・補題の適用及び前提条件・仮定の適用に関する行間補完を行う。本研究では、LLM による行間補完の実行後、LLM による生成結果の精査と修正を一定回数繰り返すことで実現する。

意味的な行間　意味的な行間の補完については、証明全体の構造を考慮しながら実行する必要がある。そのため、上記の操作的な行間を補完した後の証明に対して、自明性及び対称性・一般性を用いた証明が行われているかどうかを LLM を用いて判定し、該当する場合はそれぞれ補完を行う。このタスクは構造把握と証明全体の推論の流れの理解が必要であり、ハルシネーションが起きやすいため、補完の候補を生成する Generator と、候補に対して監査を行いその補完を実行するかどうかを決定する Auditor の 2 つのプロンプトを用いて実現した。具体例を図 3 に示す。図の例では、まず Generator が構造化された証明の中のある位置を xpath の形で指定し、その位置の文を置き換えることを提案している。それを受け取った Auditor は、証明の内容を元に提案内容を監査し、それを受理している。

細分化証明の生成　上記の行間補完を行ったのち、XML 形式の証明構造を接続詞などを補いながら証明文に戻し、その結果を最終的な細分化証明として出力する。

Generator :

```
{
  "action_type":"rewrite_step",
  "target_xpath":"./step[@id='0']",
  "text":"Since $a \leq b$ holds, ...."
}
```

Audit:

```
{
  "verdict": "accept",
  "reason": "Generated proof is enough."
}
```

図 3　意味的な行間の補完例

3.3　証明の行間調整

ここでは、3.2 節で細分化した証明に対して行間調整を行う方法について述べる。本研究では、推論過程をどの程度詳細に述べるかを表す「粒度」として以下の 4 段階を設定した。

- **詳細**：証明の構造だけでなく、すべての推論過程の背後にある定理や演算操作も含めて詳細に述べる。
- **標準**：証明の構造を反映しつつ、機械的 / 反復的な言い回しを除いて内容を述べる。
- **簡潔**：推論内容の骨子と、それを示すための最小限の事実のみを述べる。
- **概要**：証明の要点のみをごく簡潔に述べる。

本研究では粒度ごとに作成された要約のマニュアルに沿って、どのような操作を削って要約を行うかを考える立案モデルと、それに従って要約を行う実行モデルの二つを組み合わせた Planner-Executor アーキテクチャを採用して実装を行った。

4　実験

証明の細分化結果及び行間の調整結果の 2 つについて評価実験を行った。

4.1　実験設定

評価データには、DEMI-MathAnalysis [6, 7, 8] のベンチマークに含まれる実解析の命題 39 個とその証明を使用した。

証明の細分化及び行間調整モデルの学習データの作成には GPT-4.1-mini 及び GPT-5-mini を使用した。

表 1 証明の細分化による ROSCOE スコアの比較評価。スコアは [平均 (標準偏差)] の形で示している。

評価カテゴリ	元の証明	細分化後の証明
忠実性	0.859 (0.031)	**0.879 (0.032)**
論理的な一貫性	**0.140 (0.238)**	0.016 (0.063)
推論の有意義性	0.883 (0.026)	**0.937 (0.017)**
品質	**0.874 (0.033)**	0.842 (0.036)
自然さ	0.207 (0.038)	**0.211 (0.037)**

表 2 粒度に対する遵守率

粒度	適正	曖昧	詳細
概要	15.4%	0.0%	84.6%
簡潔	87.2%	0.0%	12.8%
標準	100.0%	0.0%	0.0%
詳細	97.4%	2.6%	0.0%

証明の細分化結果における推論の一貫性および意味的整合性を評価するために、ROSCOE [9] を使用した。ROSCOE は、入力された推論が命題に忠実であり論理的な飛躍を含まないか、また文章として自然であるかを評価し、NLI モデル等を用いて各観点ごとに 0〜1 に正規化されたスコアを算出する。本研究では命題に対する忠実性、論理的な一貫性、推論の有意義性（推論が証明の完了に向けて進んでいるかどうか）、品質、自然さの 5 項目で評価を行った。

行間の行間調整結果については指示遵守性の観点から評価するため、LLM-as-a-judge を使用し、細分化証明に含まれる構成要素を基準として，要約が指定された粒度に対して十分な内容を被覆しているか、指定された粒度に対して生成された証明の長さが適切かどうかを評価した。なお、評価モデルには GPT-5-mini を使用した。

4.2 実験結果

4.2.1 証明の細分化

細分化した証明に対する評価結果を表 1 に示す。まず、元の命題に対する忠実性は、元の証明と比較して向上した。この結果は、本手法が元の問題設定からの逸脱やハルシネーションを抑制しつつ、問題に忠実な推論を生成できていることを示唆している。また、証明を細分化することで推論の有意義性が大きく向上した。このことから、本手法による証明の細分化は冗長な推論を追加するものではなく、有意義な推論を付加できていると考えられる。一方で、論理的な一貫性および自然さの評価については、元の証明および細分化後の証明のいずれにおいても低いスコアとなった。特に、論理的な一貫性に関するスコアは、細分化によって大きく低下した。これは、LaTeX 形式の数式を含む詳細な推論テキストを ROSCOE が適切に認識できず、評価が正確に行われなかった可能性を示唆している。

表 3 指定した粒度と LLM-as-a-judge が入力に対して割り当てた粒度との対応表

指定＼監査	概要	簡潔	標準	詳細
概要	15.4%	33.3%	41.0%	10.3%
簡潔	0.0%	87.2%	12.8%	0.0%
標準	0.0%	0.0%	100.0%	0.0%
詳細	0.0%	0.0%	2.6%	97.4%

4.2.2 証明の行間調整

まず、粒度に対する遵守率を表 2 に示す。結果から、簡潔、標準、詳細の 3 粒度では高い遵守率の証明を生成できている一方で、概要レベルまで縮約した場合は精度が大きく低下することが確認された。

また、生成された証明に対してレビューモデルが割り当てた粒度との対応関係を表 3 に示す。この結果から、粗い粒度を指定した場合には細かな推論を十分に削減できないケースが生じる一方で、細かい粒度を指定した場合には推論を過度に削減してしまうケースがあることが分かる。これらの結果は、LLM が人間の書く証明の粒度に近い標準レベルの証明については高品質な出力を生成できる傾向がある一方で、情報量を大きく削減する概要レベルの証明においては、要約マニュアルなどの制御手法を導入した場合であっても、中間ステップを省略せずに記述する方向へのバイアスが生じやすいことを示唆している。

5 おわりに

本研究では、自然言語証明の行間補完及び削減を通じた自然言語証明の行間調整手法を提案した。また、実験を通じて提案手法の有効性を示した。今後の課題として、行間補完のローカルモデルへの置換や、より大規模なデータセットを用いた学習、さらなる評価実験の実施が挙げられる。これらの課題に取り組むことで、自動形式化や教育支援システムへの応用が期待される。

謝辞

本研究の一部は、キオクシア株式会社の支援をうけて実施したものです。

参考文献

[1] Alexander Wei. Openai imo 2025 proofs. https://github.com/aw31/openai-imo-2025-proofs/, 2025. Accessed: 2026-01-07.

[2] Google DeepMind. Gemini deep think for international mathematical olympiad 2025. https://storage.googleapis.com/deepmind-media/gemini/IMO_2025.pdf, 2025. Accessed: 2026-01-07.

[3] Hamed Mahdavi, Alireza Hashemi, Majid Daliri, Pegah Mohammadipour, Alireza Farhadi, Samira Malek, Yekta Yazdanifard, Amir Khasahmadi, and Vasant G Honavar. Brains vs. bytes: Evaluating LLM proficiency in olympiad mathematics. In **Second Conference on Language Modeling**, 2025.

[4] Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. Naturalprover: Grounded mathematical proof generation with language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 4913–4927. Curran Associates, Inc., 2022.

[5] Jianlong Chen, Chao Li, Yang Yuan, and Andrew C Yao. Hierarchical attention generates better proofs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 17506–17520, Vienna, Austria, July 2025. Association for Computational Linguistics.

[6] B.P. Demidovich. **Problems in Mathematical Analysis. Edited by B. Demidovich. Translated From the Russian by G. Yankovsky**. Russian Monographs and Texts on Advanced Mathematics and Physics. Mir Publishers, 1964.

[7] M. Hata. **Problems and Solutions in Real Analysis**. Series on number theory and its applications. World Scientific, 2007.

[8] Ziye Chen and Hao Qi. Large language models for mathematical analysis, 2024.

[9] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. 2022.

A　生成例

以下は、「数列 a_n および b_n が、それぞれ $n \to \infty$ のとき α、β に収束すると仮定する。このとき、$\frac{a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0}{n}$ が $n \to \infty$ のとき $\alpha\beta$ に収束することを示せ。」という命題に対する細分化結果及び要約例である。

A.1　元の証明

Let M be an upper bound of the two convergent sequences $|a_n|$ and $|b_n|$. For any $\epsilon > 0$ we can take a positive integer N satisfying $|a_n - \alpha| < \epsilon$ and $|b_n - \beta| < \epsilon$ for all integers n greater than N. If n is greater than N^2, then

$$|a_k b_{n-k} - \alpha\beta| \leq |(a_k - \alpha)b_{n-k} + \alpha(b_{n-k} - \beta)| \leq (M + |\alpha|)\epsilon$$

for any integer k in the interval $\left[\sqrt{n}, n - \sqrt{n}\right]$. Therefore

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=0}^{n} a_k b_{n-k} - \alpha\beta\right| \\
&\leq \frac{1}{n}\sum_{\sqrt{n}\leq k\leq n-\sqrt{n}} |a_k b_{n-k} - \alpha\beta| + 2\left(|\alpha\beta| + M^2\right)\frac{\lfloor\sqrt{n}\rfloor + 1}{n} \\
&\leq (M + |\alpha|)\epsilon + 2\left(|\alpha\beta| + M^2\right)\frac{\sqrt{n} + 1}{n}.
\end{aligned}$$

We can take n so large that the last expression is less than $(M + |\alpha| + 1)\epsilon$.

A.2　細分化証明

Suppose that a_n and b_n converge to α and β as $n \to \infty$ respectively. Show that the sequence

$$\frac{a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0}{n}$$

converges to $\alpha\beta$ as $n \to \infty$.

Let M be an upper bound of the two convergent sequences $|a_n|$ and $|b_n|$.

Since $a_n \to \alpha$ and $b_n \to \beta$, there exist positive integers N_1 and N_2 such that $|a_n - \alpha| < \epsilon$ for all $n > N_1$ and $|b_n - \beta| < \epsilon$ for all $n > N_2$, and hence taking $N = \max\{N_1, N_2\}$ gives $|a_n - \alpha| < \epsilon$ and $|b_n - \beta| < \epsilon$ for all integers $n > N$.

Take $N = \max\{N1, N2\}$ so that $|a_n - \alpha| < \varepsilon$ and $|b_n - \beta| < \varepsilon$ for all $n > N$. Let $k = \lfloor\sqrt{n}\rfloor$. If $n > N^2$ then $k \geq N$ and $n - k \geq N$, hence for any integer j with $k \leq j \leq n - k$ we have $|a_j - \alpha| < \varepsilon$ and $|b_{n-j} - \beta| < \varepsilon$. For such j,

$$\begin{aligned}
&|a_j b_{n-j} - \alpha\beta| \\
&= |(a_j - \alpha)b_{n-j} + \alpha(b_{n-j} - \beta)| \\
&\leq |a_j - \alpha||b_{n-j}| + |\alpha||b_{n-j} - \beta| \\
&\leq \varepsilon \cdot M + |\alpha| \cdot \varepsilon \\
&= (M + |\alpha|)\varepsilon,
\end{aligned}$$

using $|b_{n-j}| \leq M$.

We have: Therefore

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{k=0}^{n} a_k b_{n-k} - \alpha\beta\right| \\
&\leq \frac{1}{n}\sum_{\sqrt{n}\leq k\leq n-\sqrt{n}} |a_k b_{n-k} - \alpha\beta| + 2\left(|\alpha\beta| + M^2\right)\frac{\lfloor\sqrt{n}\rfloor + 1}{n} \\
&\leq (M + |\alpha|)\epsilon + 2\left(|\alpha\beta| + M^2\right)\frac{\sqrt{n} + 1}{n}.
\end{aligned}$$

Choosing $n > N^2$ so large that $2\left(|\alpha\beta| + M^2\right)\frac{\sqrt{n}+1}{n} < \epsilon$, the preceding inequality implies that the last expression is less than $(M + |\alpha| + 1)\epsilon$.

A.3　要約結果 （粒度: 概要）

We show that if $a_n \to \alpha$ and $b_n \to \beta$ as $n \to \infty$, then

$$\frac{1}{n}\sum_{k=0}^{n} a_k b_{n-k} \longrightarrow \alpha\beta \quad \text{as } n \to \infty.$$

Using a common bound M for $|a_n|$ and $|b_n|$, together with the convergences $a_n \to \alpha$ and $b_n \to \beta$, one obtains for sufficiently large n the estimate

$$\left|\frac{1}{n}\sum_{k=0}^{n} a_k b_{n-k} - \alpha\beta\right| \leq (M + |\alpha|)\varepsilon + 2\big(|\alpha\beta| + M^2\big)\frac{\sqrt{n} + 1}{n}.$$

Choosing n so large that

$$2\big(|\alpha\beta| + M^2\big)\frac{\sqrt{n} + 1}{n} < \varepsilon$$

gives the bound

$$\left|\frac{1}{n}\sum_{k=0}^{n} a_k b_{n-k} - \alpha\beta\right| < (M + |\alpha| + 1)\varepsilon.$$

Letting $\varepsilon \to 0$ yields

$$\frac{1}{n}\sum_{k=0}^{n} a_k b_{n-k} \longrightarrow \alpha\beta.$$