

大規模言語モデルの推論過程誤り分析に基づく推論能力の向上

吉田琉夏 西田光甫 西田京介

NTT 株式会社 人間情報研究所

{runa.yoshida, kosuke.nishida, kyosuke.nishida}@ntt.com

概要

大規模言語モデル (LLM) の推論能力は実用性に関わる重要な要素であり, LLM の推論過程にどのような誤りが生じるのかを明らかにすることは, 推論能力を改善するうえで重要である. 本研究では, LLM が数学データセットにおいて推論過程で示す誤りの傾向を明らかにする. 具体的には, LLM が正解に至らなかった問題を対象として, 出力された推論過程に含まれる誤りの性質を手で分析し, 21 種類の誤りクラスを定義し分類した. さらに, 得られた誤り分類のうち 9 種に特に注意を向けるプロンプトを設計し, それが推論性能の向上に有効であることを示した. また, 本研究で対象とした誤りは同規模の LLM で共通することが示唆された.

1 はじめに

大規模言語モデル (LLM) は多様なタスクにおいて高い性能を示しており, 実応用が進んでいる. 特に, LLM の推論能力は, 科学計算, ソフトウェア検証, 教育支援などの実応用に直結する重要な要素である. 推論能力の一例として, 複数の条件をすべて満たす結論を導く能力や, 複雑な問題をいくつかの簡単な問題に分け, 順番に解いていく能力が挙げられる. 数学的推論は厳密な論理の一貫性と計算過程の正確性を要求する課題であるため, LLM の推論能力の評価には, 数学領域のベンチマークが用いられてきた [1, 2]. しかし, 既存の多くのモデルは, 数学領域において, 問題構造や推論の複雑さが増すにつれて性能が大きく低下し, 複数段階の推論を要する問題では, 出力中の推論過程に誤りを生じやすいことが報告されている [1, 2, 3].

LLM の出力を改善する手法として, 推論過程で自身の出力を見直し, 修正させる枠組みである自己修正が注目されている [4, 5, 6, 7]. 一方で, LLM の自己修正能力の低さは, 推論誤りの修正能力の不足というよりも, 推論誤りを発見する能力の欠如に起

因することが指摘されている [8]. しかし, 既存研究の多くは, 各サンプルの最終解が正しいか否かに基づいて議論されており, LLM がなぜ誤答に至ったのかという観点から推論過程における誤り (以下, 推論誤りと呼ぶ) を詳細に分析する試みは十分に行われていない.

そこで本研究は, LLM の数学データセットにおける推論誤りの傾向を明らかにすることで, LLM の推論能力を改善することを目標とする. 具体的には, MATH ベンチマーク [2] のテストセット 5,000 問のうち, Llama-3.3-70B-Instruct モデル¹⁾ [9] が正解に至らなかった 583 問に対して推論誤りを収集し, 人手により 21 種類の誤りクラスを定義し, 推論誤りを分類した. 分類の結果, 推論方略に関する誤り, 問題文の条件や設定の文脈理解に関する誤り, 計算・代数操作の誤りが頻発することが明らかとなった. また, 推論誤りを LLM 自らに発見させる代わりに, 得られた頻出の推論誤り分類を明示的に考慮したプロンプトを設計し, その有効性を検証する. 実験の結果, 推論誤り分類を考慮して設計したプロンプトは数学データセットにおける性能を向上させることを確認した. さらに, 本研究での推論誤りは同規模の LLM に共通する課題であると示唆された.

2 関連研究

推論誤り分類 先行研究 [10] は, 8B 以下の言語モデルが数学データセットにおいて示す推論誤りを 8 クラスに分類し, 各クラスに対応した学習データが数学的推論性能が向上することを示した. しかし, このクラスは Geometric Errors といった問題の分野に依存するクラスや Logical and Reasoning Errors といった包括的な粒度にとどまっており, 具体的な推論誤りを十分に分析できていない. 推論誤りに関する他の先行研究 [11, 12, 13, 14, 15] でも LLM の推論誤りパターンが指摘されているが, 包括的な粒度のクラスにとどまっている.

1) <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

表 1 MATH ベンチマークのテストセットの問題数

	Lev. 1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	合計
代数	135	201	261	283	307	1,187
計数と確率	39	101	100	111	123	474
幾何	38	82	102	125	132	479
中級代数	52	128	195	248	280	903
数論	30	92	122	142	154	540
代数入門	86	177	224	191	193	871
微積分入門	57	113	127	114	135	546
合計	437	894	1,131	1,214	1,324	5,000

プロンプトによる推論能力の向上 現在、多様な Chain-of-Thought (CoT) プロンプト手法が研究されており、それらが LLM の解決能力を向上させることが示されている [3, 16]. 例えば、質問の末尾に「Let's think step by step」と付加すると、LLM が段階的な推論過程を生成し、誤答が減少することが報告されている [17].

3 推論誤りの分析

本節では、分析対象とする問題の取得手法、推論誤りの分類、および分類結果について紹介する。

3.1 分析対象である問題の取得

分析対象とする問題の収集元 MATH ベンチマークのテストセットを採用した。MATH ベンチマークは、中等教育から大学初等レベルの広範な数学問題を含み、難易度 1 から 5 までのレベルで構成される数学データセットである。また、代数、計数と確率、幾何、中級代数、数論、代数入門、微積分入門の 7 分野で構成される。各問題には、推論過程および最終解がアノテーションされており、Llama-3-70B モデル²⁾ [9] においても正しい推論が困難な高難度問題が多く存在する [14] ため、分析対象として相応しいと考える。表 1 に MATH ベンチマークのテストセットの問題数を示す。

分析対象の決定方法 Llama-3.3-70B-Instruct モデルが正解に至らなかった問題を分析対象とする。8B モデルでは計算能力が推論能力の制約となることが指摘されているため [14], 70B モデルを採用した。各問題に対して系列長を 1,024 に固定した CoT 推論を行い、zero-shot 設定において greedy デコードにより解答を生成した。系列長の制約により最終解が出力されなかった問題は、分析対象から除外した。

2) <https://huggingface.co/meta-llama/Meta-Llama-3-70B>

表 2 分析対象の問題数

	Lev. 1	Lev. 2	Lev. 3	Lev. 4	Lev. 5	合計
代数	0	3	5	8	25	41
計数と確率	1	10	9	16	37	73
幾何	8	16	14	36	50	124
中級代数	1	9	23	31	68	132
数論	0	7	15	22	40	84
代数入門	2	7	10	17	40	76
微積分入門	2	7	10	12	22	53
合計	14	59	86	142	282	583

正誤判定 上記の分析対象の決定における正誤判定は、モデルが出力した最終解とデータセットでアノテーションされた最終解との Exact Match (EM) に基づいた。しかし、EM による評価では、両者が数学的に同値であるにもかかわらず、誤りと判定される場合がある。たとえば、 $-\sqrt{2} + 1$ と $1 - \sqrt{2}$ は同値であるが、EM による評価では誤りと判断される。そこで、本研究では gpt-oss-120b³⁾ [18] を用いて両者が同値であるかを判定し、同値と判定されたものを人手による確認のうえ分析対象から除外した。EM による評価で正解に至らなかったと判断された問題は全問題数の 29.6% 存在した。表 2 に分析対象の問題数を示す。

実装 DeepSeek-MATH [19] が提供する評価スクリプト⁴⁾を使用した。プロンプト設定を含む評価手順は、同スクリプトに準拠した。

3.2 推論誤りクラスの定義および分類

分析対象の問題で発生した推論誤りについて推論誤りクラスを定義し、人手で分類した。まず、分析対象である正解に至らなかった問題における推論過程を精査し、推論誤りを「I. 問題文の条件や設定の文脈理解に関する誤り」、「II. 数学概念の知識に関する誤り」、「III. 計算・代数操作の誤り」、「IV. 推論方略に関する誤り」、「V. その他」に大きく分類した。これは、先行研究 [10, 15] で挙げられた推論誤りクラスのうち、Geometric Errors のような問題の分野に依存するクラスを除外し、各推論誤りクラスが複数の大きな分類に該当しないよう考慮して決定した。次に、大きな分類それぞれに該当する推論誤りを比較し、類似する推論誤りが同じクラスになるように推論誤りクラスを定義した。その結果、推論誤りクラスは 21 種類となった。定義した推論誤りク

3) <https://huggingface.co/openai/gpt-oss-120b>

4) <https://github.com/deepseek-ai/DeepSeek-Math/tree/main/evaluation>

表 3 定義した推論誤りクラス, および分類された問題数. 一部の問題は複数の推論誤りクラスに該当する.

	問題数
I. 問題文の条件や設定の文脈理解に関する誤り	計 185
I-1. 図表に含まれる情報の読み取り誤り	69
I-2. 問題文に明示された条件を考慮しない誤り	53
I-3. 問題文に明示されていない条件を導入する誤り	31
I-4. 条件の誤解釈	27
I-5. 範囲・量・比較条件の解釈誤り	5
II. 数学概念の知識に関する誤り	計 32
II-1. 定理・公式の誤記憶	19
II-2. 定義や性質の異なる概念を区別しない誤り	13
III. 計算・代数操作の誤り	計 135
III-1. 数値計算誤り	61
III-2. 代数操作の誤り	39
III-3. 数え上げ誤り	31
III-4. 数の大小・順序関係の誤り	4
IV. 推論方略に関する誤り	計 230
IV-1. 不適切な定理適用	77
IV-2. 推測した仮定・制約の誤り	57
IV-3. 条件充足の誤判定	38
IV-4. 方略の誤り	18
IV-5. 公式・定理適用誤り	14
IV-6. 必要条件の推測誤り	12
IV-7. 規則性や反復的構造の定式化誤り	9
IV-8. 不適切な処理	5
V. その他	計 52
V-1. 転記誤り	37
V-2. 非誤答	15

ラス, および該当する問題数を表 3 に示す. 各推論誤りクラスの分野別問題数については付録 A を参照されたい. 分類の結果, 「IV. 推論方略に関する誤り」, 「I. 問題文の条件や設定の文脈理解に関する誤り」, および「III. 計算・代数操作の誤り」が頻発することが明らかとなった.

特に説明を要する推論誤りクラスについて, 以下にその定義を示す.

I. 問題文の条件や設定の文脈理解に関する誤り

I-4. 条件の誤解釈 問われている内容を誤解し, 本来とは異なる目的に基づいて解答する誤り

III. 計算・代数操作の誤り

III-3. 数え上げ誤り 2, 3, 4 と出力しているにもかかわらず, 整数の個数を 2 個と数えるなど, 出力した集合や範囲内の要素数の数え上げ誤り

IV. 推論方略に関する誤り

IV-1. 不適切な定理適用 定理の前提を満たしていない, あるいは前提を誤って解釈している状

{problem}

Please reason step by step, and put your final answer within $\boxed{\}$. When answering questions and explaining your reasoning process, please pay particular attention to the following perspectives at each step:

- Have all conditions within the text been extracted? (I-3)
- Have you introduced any implicit assumptions not stated in the problem? (IV-1)

- Have the prerequisites for the theorems and formulas used been verified? (IV-1, 2, 5)

- Does it satisfy all the conditions and prerequisites of the problem statement? (I-2, IV-1, 3)

- Is it clear what each step does and why the operation is valid? (IV-4)

Do not output confirmation for items that have already been verified. (I-3, 4)

図 1 推論誤りクラスを考慮した提案プロンプト.

青字は DeepSeek-MATH の zero-shot 標準プロンプトを示す. 各指示で考慮した推論誤りクラスの番号を括弧内に示す. 推論誤りクラスの番号は表 3 に対応している.

況において, 当該定理を適用する誤り

IV-2. 推測した仮定・制約の誤り 既存の条件や構造から導いた仮定・制約の誤り

IV-3. 条件充足の誤判定 条件, および条件が成り立つかどうかについて言及しているが, 当該条件の成立可否を正しく判断できていない誤り

IV-4. 方略の誤り 適切でない解法や手法を選択し, 不必要に複雑な, あるいは不合理な解答過程を採用する誤り

IV-5. 公式・定理適用誤り 公式の適用条件や定理の前提は満たしているが, 代入すべき値や対応関係を取り違える誤り

IV-8. 不適切な処理 後続の計算で誤差のない値を前提とするにもかかわらず, 中間段階で数値近似や丸め処理を導入するような, その後の計算や推論の前提と整合しない不適切な操作や処理を導入してしまう誤り

V. その他

V-1. 転記誤り 正しい結論に到達しているにもかかわらず, 最終解に誤った値を記述する誤り

V-2. 非誤答 データセットでアノテーションされた最終解の誤り, または単位の有無などモデルが出力した最終解が内容的に妥当であるもの

4 推論誤り分析による推論性能向上

推論誤りを分析することが LLM の推論能力向上に寄与することを示す.

4.1 提案手法

本研究では、LLM が数学問題を解く際に生じる推論誤りの分類結果を明示的に考慮したプロンプトを提案する。具体的には、推論誤りの分類結果として特に頻繁に観測された「I. 問題文の条件や設定に関する文脈理解の誤り」、および「IV. 推論方略に関する誤り」の2種類に着目し、推論の各段階において、問題文の条件整合性や推論の妥当性を逐次確認するようモデルに促すプロンプトを設計した。提案プロンプトを図1に示す。

4.2 実験設定

ベースライン DeepSeek-MATH の zero-shot 標準プロンプト (以下、標準と呼ぶ。図1の青字参照) を採用した。さらに、既存研究 [10] の推論誤りクラスである Logical and Reasoning Errors に基づき、標準の末尾に推論過程への注意を促す文言「Please pay particular attention when answering questions and explaining your reasoning process.」を付加したプロンプト (以下、注意喚起と呼ぶ) を採用した。

開発・評価データ分割 開発データとして、MATH データセットのテストセットから、各分野および各難易度レベルから 20 問ずつランダムに抽出した計 700 問を使用した。残りの 4,300 問を評価データとした。提案手法におけるプロンプトの文言は、開発データ上における Llama-3.3-70B-Instruct の性能を踏まえて決定した。

評価モデルと設定 Llama-3.3-70B-Instruct および Qwen3-32B⁵⁾ [20] の2種類を評価モデルとして使用した。Temperature は 0.6 とし、5 つの乱数シードによる結果を平均して評価した。系列長は 8,192 に設定し、問題の誤答判定には EM を用いた。

評価指標 推論誤りの種類を考慮したプロンプトと各ベースラインとの間に有意な差が存在するかを検証するため、マンホイットニーの U 検定 [21] による p 値、および Cohen's d [22] を算出した。

4.3 実験結果

提案プロンプトは推論性能をより向上させるか？

表4の通り、提案した推論誤り分類考慮プロンプトは、標準プロンプトと比較して一貫して性能を向上させ、推論能力改善への有効性を示した。また、注意喚起プロンプトと比較しても性能が向上している

表4 ベースラインと提案手法のプロンプトを評価した結果。誤答問題数は、EM による評価による誤答問題数の平均と標準偏差を示す。p 値と Cohen's d は、各ベースラインに対する提案手法の値を示す。

	誤答問題数↓	p 値↓	Cohen's d↑
Llama-3.3-70B-Instruct			
標準	1,117.8 (± 16.5)	0.008	2.94
注意喚起	1,111.6 (± 24.2)	0.008	1.85
提案手法	1,077.0 (± 10.7)	-	-
Qwen3-32B			
標準	677.0 (± 29.8)	0.032	2.30
注意喚起	614.4 (± 30.4)	0.056	2.30
提案手法	476.4 (± 79.3)	-	-

ことから、単に注意を促すのではなく、推論誤りの種類を踏まえて、誤りが生じやすい観点に明示的に注意を向けさせることが重要であることが示唆された。注意喚起プロンプトでは改善が見られず、提案した推論誤り分類考慮プロンプトによって改善された推論誤りの例を付録 B に示す。

本研究の推論誤りはモデル固有のものか？

表4の通り、Qwen3-32B においても性能向上が確認されたことから、本研究で分析した推論誤りは Llama3 系列に固有のものではなく、同規模の LLM に共通する課題である可能性が示唆される。

5 おわりに

本研究は、LLM が数学データセットにおける推論誤りを人手で分析し、分類を行った。

本研究の独自性 本研究では、既存研究に基づく大きい分類を定義したのち、さらに細分化することで、新たに 21 種類の推論誤りクラスを定義した。さらに、得られた推論誤り分類に基づいて注意を促すプロンプトが、既存研究の大きい分類に基づき推論過程への注意を促すプロンプトよりも、数学的推論性能の向上に寄与することを実験的に示した。

本研究の重要性 LLM の推論能力は Agentic AI をはじめとした LLM アプリケーションの根幹を担う能力である。本研究は、推論誤りの詳細な定性分析を通じて、現在の LLM が抱える推論能力の課題をより具体的に明らかにした。さらに、ベンチマークにおける EM 評価ではなく推論誤りの定性分析に基づくことで LLM の出力が改善できることを示した。これらより、本研究は、LLM の実応用における課題とその改善に向けた指針を明らかにした。

5) <https://huggingface.co/Qwen/Qwen3-32B>

参考文献

- [1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, Vol. abs/2110.14168, , 2021.
- [2] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- [3] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large 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. 24824–24837. Curran Associates, Inc., 2022.
- [4] Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. Iterative translation refinement with large language models. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, **Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)**, pp. 181–190, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT).
- [5] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 46534–46594. Curran Associates, Inc., 2023.
- [6] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.
- [7] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, Vol. 12, pp. 484–506, 2024.
- [8] Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 13894–13908, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [9] AI@Meta. Llama 3 model card. 2024.
- [10] Erxin Yu, Jing Li, Ming Liao, Qi Zhu, Boyang Xue, Minghui Xu, Baojun Wang, Lanqing Hong, Fei Mi, and Lifeng Shang. Self-error-instruct: Generalizing from errors for LLMs mathematical reasoning. 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. 8504–8519, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [12] Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In **The Eleventh International Conference on Learning Representations**, 2023.
- [13] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems 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. 3843–3857. Curran Associates, Inc., 2022.
- [14] Kathrin Seßler, Yao Rong, Emek Gözlüklü, and Enkelejda Kasneci. Benchmarking large language models for math reasoning tasks, 2024.
- [15] Johan Boye and Birger Moell. Large language models and mathematical reasoning failures. *arXiv preprint arXiv:2502.11574*, 2025.
- [16] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In **ICLR 2023**, 2023.
- [17] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 22199–22213. Curran Associates, Inc., 2022.
- [18] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.
- [19] Qihao Zhu Runxin Xu Junxiao Song Mingchuan Zhang Y.K. Li Y. Wu Daya Guo Zhihong Shao, Peiyi Wang. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.
- [20] Qwen Team. Qwen3 technical report, 2025.
- [21] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947.
- [22] J Cohen. **Statistical Power Analysis for the Behavioral Sciences (2nd ed.)**. Routledge, 1988.

A 推論誤り分類の詳細

各推論誤りクラスに該当する問題数を表 5 に示す。

表 5 各推論誤りクラスに該当する問題数 (分野別)

	代数	計数と確率	幾何	中級代数	数論	代数入門	微積分入門	合計
I. 問題文の条件や設定の文脈理解に関する誤り	8	31	52	22	23	34	15	185
I-1. 図表に含まれる情報の読み取り誤り	3	6	32	2	3	19	4	69
I-2. 問題文に明示された条件を考慮しない誤り	4	11	7	8	7	9	7	53
I-3. 問題文に明示されていない条件を導入する誤り	0	3	6	7	8	3	4	31
I-4. 条件の誤解釈	0	10	7	3	4	3	0	27
I-5. 範囲・量・比較条件の解釈誤り	1	1	0	2	1	0	0	5
II. 数学概念の知識に関する誤り	1	6	10	9	1	2	3	32
II-1. 定理・公式の誤記憶	0	0	9	7	1	0	2	19
II-2. 定義や性質の異なる概念を区別しない誤り	1	6	1	2	0	2	1	13
III. 計算・代数操作の誤り	22	15	6	44	27	12	9	135
III-1. 数値計算誤り	14	6	4	13	20	1	3	61
III-2. 代数操作の誤り	2	0	1	27	2	1	6	39
III-3. 数え上げ誤り	5	9	1	2	5	9	0	31
III-4. 数の大小・順序関係の誤り	1	0	0	2	0	1	0	4
IV. 推論方略に関する誤り	8	27	66	55	21	22	31	230
IV-1. 不適切な定理適用	1	4	24	15	3	15	15	77
IV-2. 推測した仮定・制約の誤り	0	7	24	15	5	1	5	57
IV-3. 条件充足の誤判定	3	9	1	7	7	5	6	38
IV-4. 方略の誤り	0	1	4	8	3	1	1	18
IV-5. 公式・定理適用誤り	1	1	10	2	0	0	0	14
IV-6. 必要条件の推測誤り	2	3	1	3	1	0	2	12
IV-7. 規則性や反復的構造の定式化誤り	1	1	1	3	2	0	1	9
IV-8. 不適切な処理	0	1	1	2	0	0	1	5
V. その他	3	3	9	7	15	13	2	52
V-1. 転記誤り	2	3	4	7	10	9	2	37
V-2. 非誤答	1	0	5	0	5	4	0	15

B 推論誤り分類考慮プロンプトにより改善した例

注意喚起プロンプトでは改善せず、推論誤り分類考慮した提案プロンプトにより改善した例を図 2 に示す。

【問題】

Your friend has an egg collection comprising at least 200 eggs. ...{略}... What is the minimum number of eggs that your friend could have?

【データセットに含まれる最終解】 : 313

【標準プロンプトおよび注意喚起プロンプト】

Step 1: Understand the problem

The problem involves finding the minimum number of eggs a friend could have, given that when stored in dozen-egg containers, there is 1 egg left over, and when stored in baker-dozen-egg containers (holding 13 eggs), there is also 1 egg left over. ...{略}... Since $N-1$ must be a multiple of the LCM of 12 and 13, the smallest possible value for $N-1$ is 156. Therefore, the minimum number of eggs N is $156 + 1 = 157$. The final answer is: $\boxed{157}$

【提案手法のプロンプト】

Step 1: Understand the problem conditions

The problem states that the friend has at least 200 eggs and wants to store them in dozen-egg containers. ...{略}...

Step 10: Verify the minimum solution

Given that 157 is the first common solution but is less than 200, and knowing that the pattern repeats every 156 eggs, the next possible solution would indeed be $157 + 156 = 313$. This satisfies both modular conditions and is above the minimum of 200 eggs. It's essential to verify that no smaller solution exists within the given constraints. The final answer is: $\boxed{313}$

図 2 注意喚起プロンプトでは改善せず、提案プロンプトで改善した誤りの例。この例の推論誤りは「I-2. 問題文に明示された条件を考慮しない誤り」に分類される。この例の注意喚起プロンプトの出力は、標準プロンプトと同じであった。