ドメイン特化 LLM の推論能力向上を目的とした合成指示データセットの構築と金融ドメインにおける評価

大河内 悠磨 [1]　Fabio Milentiansen Sim[2]　岡田 智靖 [1]
[1] 株式会社野村総合研究所　[2]NRI インドネシア
{y2-okochi,t3-okada}@nri.co.jp　fabio.sim@nri.co.id

概要

特定ドメインへの LLM 適応において，専門知識と推論能力の両立は喫緊の課題である．本研究では，ドメイン語彙を起点としたデータ合成により，任意の分野で高品質な合成指示データを構築する汎用手法を提案する．実証として本手法を金融分野に適用し，総計約 95 億トークンの大規模な思考過程（Chain-of-Thought）付き指示データセットを構築した．評価の結果，金融ベンチマークにおいて公式モデル比で性能向上を確認し，手法の有効性を実証した．併せて，思考過程長が性能に与える影響とその限界についても報告する．

1　はじめに

大規模言語モデル（LLM）の社会実装が進む中，金融や法務といった高度な専門性が求められるドメインへの適応が活発に研究されている．従来，これらの分野への適応には継続事前学習（Continued Pre-Training）が多く用いられ，専門知識の習得において大きな成果を上げてきた [1]．これに加え，近年では単なる知識の有無だけでなく，複雑な事象の分析や将来予測（例：不正検知や業績予測 [2]）といった，高度な論理的推論を要する実務タスクへの対応が重要視されている．

近年，回答の前に思考過程（Chain-of-Thought）を生成することで性能を高める推論モデルが注目されている [3]．この能力獲得には思考過程を含む Supervised Fine-tuning（SFT）が有効だが，汎用的なデータセット [4] は存在するものの，特定ドメイン（特に日本語環境）に特化し，かつ推論過程を含んだ大規模データの構築手法は確立されていない．

そこで本研究では，特定ドメインにおける推論能力獲得を目的としたデータ構築パイプラインを提案し，以下の貢献を行う．

- ドメイン固有のトピックワードを起点に，Reasoning LLM を用いて思考過程付き指示データを合成・選別する手法を確立した．本手法は金融に限らず多様な分野へ適用可能である．
- 提案手法を日本の金融分野に適用し，総計約 95 億トークンのデータセットを構築した．これを用いた学習により，金融ベンチマークにおいて公式モデルを上回る性能向上を達成した．
- ドメイン特化タスクにおける思考過程長の分析を行い，推論トークン数が性能に与える影響とその境界に関する知見を得た．

2　関連研究

2.1　ドメイン適応と指示データセット

LLM を特定ドメインに適応させるアプローチとして，ドメイン特化データセットを用いた指示チューニングが盛んに研究されている．金融分野においては，Ke ら [5] や Lee ら [6] が大規模な指示データセットを提案しているが，これらは英語や中国語が中心である．日本語においては，Tanabe ら [7] が金融特化の指示データセットを提案しているが，推論過程（CoT）を含まない形式が主であった．本研究は，既存研究で不足していた「ドメイン知識」と「推論過程」の両立を，合成データ生成技術を用いることで解決するアプローチを採る．特に，トピックワード起点の生成手法を採用することで，任意のドメインに対して拡張可能なフレームワークを提示する点に特徴がある．

2.2　思考過程長に関する検証

複雑なタスクに対しては，解答を出力する前に思考過程を生成する手法が有効であり，思考過程を出力する reasoning model も多数提案されている [3, 8, 9]．数学ベンチマークを対象とした先行研究で

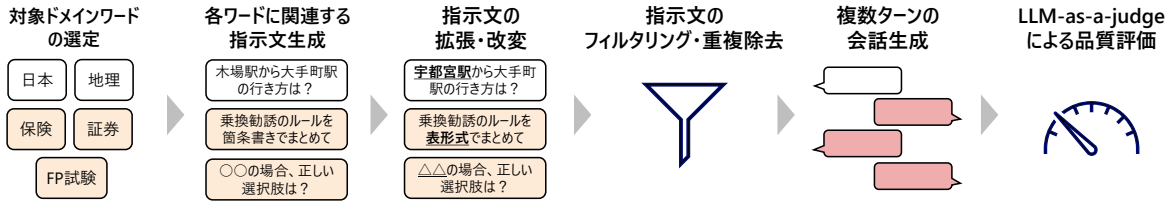

図 1　提案するデータセット構築パイプライン

は，思考過程長の増加による性能向上が報告されている [10] 一方で，一定の長さを超えると性能が低下する可能性も指摘されている [11]．これらの検証は主に数学タスクに限られており，専門知識や制度理解を要するドメイン特化タスクにおける検証は十分に行われていない．本研究では，金融ベンチマークを用いて思考過程長と性能の関係を検証する．

3　データセット構築手法

本研究で提案するデータセット構築パイプラインを図 1 に示す．本手法は，Nemotron-4 340B の構築に用いられた合成手法 [12] を拡張し，特定ドメインの知識と推論能力を効率的に学習させるために最適化したものである．本手法はドメインに依存しない汎用的なプロセスであるが，本実験ではその有効性を検証するため，ターゲットドメインとして金融分野を設定し，構築および評価を行った．なお，合成に用いる LLM は，特に断りのない限り OpenAI 社の gpt-oss-120b [13] を用いている．

また，本パイプラインの各工程における生成数，フィルタリング閾値などの詳細なパラメータ設定については，まとめて Appendix A に示す．

ドメイン・トピックワードの選定

本手法の第一段階は，対象ドメインを網羅するトピックワードの選定である．対象ドメインの指示文を増やすため，金融分野を対象とした業界名や金融商品，関連技術などのシードワードを選定した．また，ドメイン特化による汎用性能の低下（Catastrophic Forgetting）を抑制するため，対象ドメイン以外のシードワードも混合したうえで，各ワードに関連するサブトピックワードを生成させた．

ドメインに関連する指示文生成

次に，生成した各サブトピックワードに関連するユーザの発話（指示文）を LLM に生成させた．タスクの多様性を担保するため，オープンな質問，数学的推論（計算問題），作文，多肢選択式問題の 4 つのタイプを定義した．各タイプについて，多様性を確保するために十分な数の指示文を生成した．

指示文の拡張・改変

ユーザの発話のバリエーションを広げるために，前段で生成した指示文に対する拡張・改変を行った．具体的には，文脈の追加，異なる形式・スタイルへの変換，特定の事柄に対する詳細な指示文への改変，関連する事柄に対する指示文への改変を行った．どのタイプの改変を行うかは LLM に任せることとし，元の指示タイプに応じた数のバリエーションを生成した．

指示文のフィルタリング・重複除去

生成されたデータの品質を担保するため，Swallow Corpus v2 [14] の手法を参考に，n-gram と単語数に基づくフィルタリング，および MinHash と LSH を利用した Fuzzy deduplication による重複除去を行った．n-gram によるフィルタリングは先行研究と同じパラメータを，単語数のフィルタリングは MeCab [15] による単語分割を行ったうえで，極端に短い指示文を除去した．

複数ターンの思考過程付き会話生成

前段で得た高品質な指示文を用い，複数ターン対話を生成して対話的タスク思考能力の向上を図った．1 ターン目は指示文をそのまま入力し，以降は直前までのやり取りを与えてユーザ発話と思考過程を含む LLM 応答を交互に生成した．

LLM-as-a-judge による品質フィルタリング

最後に，生成したデータが対象ドメインの知識を正確に反映し，かつ論理的な推論を含んでいるかを判定するため，LLM-as-a-judge による品質フィルタリングを行った．判定には gpt-oss-120b を用い，複数の評価観点に基づくスコアの平均値によって低品質データを除外した．詳細な評価手法および閾値設定については Appendix B に示す．

公開データセットの活用

ドメイン特化能力に加え，基礎的な数学・指示追従能力を補強するために，公開指示データセットを活用した．数学能力向上には NuminaMath-CoT [16] と Nemotron-Post-Training-Dataset-v1 [4] を，指示追従能力向上には smol-constraints [17] を用いた．これらのデータセットの質問部分のみを抽出し，図 1 におけるフィルタリング・重複除去以降の処理を実施した．なお，これらは 1 ターンのみの対話生成とした．

最終的にサンプル数約 144 万件，総計約 95 億トークンの指示データセットを作成した．データセットの詳細情報は Appendix C に示す．

4 実験評価

構築したデータセットを用いて LLM の学習と評価を行い，対象ドメインにおける提案手法の有効性を検証する．

4.1 金融コーパスを用いた継続事前学習

LLM のドメイン適応においては，事前学習モデルに対しドメイン特化コーパスを用いた継続事前学習 (Continued Pre-Training) を行うことが一般的である [18]．本研究では，Common Crawl などの公開コーパスに金融ドメイン分類や品質フィルタリングを適用し，日本語金融コーパスを構築した．このコーパスを用いて，Qwen3-14B-Base [9]，gpt-oss-20b [13] に継続事前学習を行った．

4.2 SFT による指示チューニング

本研究で構築したデータセットの効果検証を行うため，採用したベースモデルおよび 4.1 節で構築した継続事前学習済みモデルに対する Supervised Fine-tuning (SFT) を行った．エポック数は 2 とし，学習は AWS の p5en.48xlarge (NVIDIA H200 Tensor Core GPU x 8) インスタンス 1 台で実施した．各モデルで実行に約 240 時間を要した．

4.3 金融ベンチマークによる評価

訓練したモデルの金融分野における能力を，japanese-lm-fin-harness [19] および pfmt-bench-fin-ja [20] を用いて評価した．japanese-lm-fin-harness は有価証券報告書の感情分析や，金融系資格試験問題を含むベンチマークであり，モデルの金融知識の習得度を検証する目的で用いた．

pfmt-bench-fin-ja は金融分野における作文，情報抽出，アイデア出しなどのタスクにおいて，ユーザの質疑に複数ターンで適切に回答できるかを比較するベンチマークであり，金融系対話能力の検証に用いた．

各ベンチマークは評価方法を公開リポジトリの実装から改変しており，詳細を Appendix D に示す．

各ベンチマークによる評価結果を表 1 に示す．SFT のみを実施した場合，Qwen3-14B-Base に SFT を適用したモデルは，公式指示モデルと比較して低いスコアを示した．これは，モデル出力にループが生じる，あるいは所定の回答形式を遵守できないといった問題により回答抽出が失敗したことが主因であると考えられる．一方，最終結果に至る学習過程においては，学習ステップ数の増加に伴いこれらのエラーが減少する傾向が観測されたため，学習を継続することで性能向上が期待される．

CPT モデルに本手法で構築したデータセットを用いた SFT を施した結果，両方のベンチマークの平均スコアが向上した．gpt-oss-20b は全サブタスクで公式指示モデルを上回り，Qwen3-14B も一部を除き同様の傾向を示した．これは，提案するデータ構築手法が，ドメイン知識の注入（CPT）と推論能力の獲得（SFT）を効果的に結びつけ，特定ドメインにおけるタスク性能を最大化できることを示唆している．

また，ベースモデルの思考過程の有無による性能差の比較も行った．結果，いずれのベースモデルでも思考過程ありの推論は，思考過程なしに比べ，japanese-lm-fin-harness で約 4.5〜5.7pt，pfmt-bench-fin-ja で約 0.4pt 高い性能を示し，金融分野ベンチマークにおいても reasoning model の有効性が確認された．

4.4 思考過程長に伴う性能向上の検証

4.3 節の評価結果から，思考過程の出力が金融タスクの性能向上に寄与することが確認された．思考過程の長さがモデル性能に与える影響を検証するために，思考過程の長さを一定の値に固定し，その固定値を段階的に変化させた．

先行研究 [10] に倣い，思考過程の長さを指定値に固定するため，以下の手順を用いた．

指定長より短く終了した場合 reasoning 終了トークン（例：`</think>`）を除去し，末尾に “`Wait,`” を追加して推論を継続させた．

指定長以上で終了した場合 指定長を超えるトーク

表 1　金融ベンチマークによる各モデルの評価結果 (CoT = 思考過程の有無)

Model	CoT	japanese-lm-fin-harness						pfmt-bench-fin-ja		
		Avg.	chabsa	cma	cpa	fp2	ss1	Avg.	turn1	turn2
Qwen3-14B	なし	66.50	91.73	87.58	41.83	47.03	64.31	7.781	7.819	7.743
Qwen3-14B	あり	71.04	**91.96**	**93.26**	**49.37**	53.37	67.22	8.104	8.211	7.997
Qwen3-14B-Base + SFT (Ours)	あり	70.69	91.48	91.20	47.96	55.00	**67.81**	8.415	8.472	8.358
Qwen3-14B-Base + CPT + SFT (Ours)	あり	**71.78**	91.62	91.45	48.59	**60.00**	67.27	**8.455**	**8.514**	**8.395**
gpt-oss-20b	なし	61.17	91.09	81.33	33.45	41.82	58.17	7.480	7.425	7.534
gpt-oss-20b	あり	66.93	91.80	90.46	38.51	49.74	64.15	7.883	7.858	7.908
gpt-oss-20b + SFT (Ours)	あり	69.56	91.87	90.87	43.65	52.63	**68.80**	**8.432**	**8.439**	**8.424**
gpt-oss-20b + CPT + SFT (Ours)	あり	**72.50**	**91.89**	**94.24**	**45.51**	**62.71**	68.15	8.209	7.992	8.425

ンを削除し，reasoning 終了トークンと最終回答を促すテキスト（改行）“`Final Answer:`” を挿入したうえで，改めて最終回答を生成させた．

対象ベンチマークは japanese-lm-fin-harness とし，chabsa 以外の 4 サブタスクでは固定長を 128，256，512，1024，2048，4096，8192 トークンの 7 種類で設定した．chabsa は実行時間の都合上，128，256，512 トークンの 3 種類に限定して実験を行った．

各タスクにおける実行結果を図 2 に示す．chabsa タスクを除く各サブタスクでは，思考過程の固定長を 1024 トークンまで延ばすことで性能が向上した．chabsa についても，検証範囲である 512 トークンまでは同様の傾向が見られた．一方，固定長を 2048 トークン以上に設定しても性能の向上は確認されなかった．思考過程を分析した結果，その要因として以下の挙動が観測された．

思考の早期終了　“`Wait,`” を挿入しても，“`Wait, but the answer is D. So the final answer is D.`” のように直後に結論へ移行し，より深い推論をせずに思考過程を終了しようとする．

出力のループ　特に 4096 トークン以上では，結論の再確認のみを繰り返す出力が固定長に達するまで続き，実質的な推論が行われなくなる．

また，思考過程を強制的に終了させた場合，強制終了を行わない推論（図 2 中の★印）と比較して，すべてのタスクで数 pt の性能低下が確認された．ほぼ同程度の思考過程長であっても性能が低下することから，思考過程の終了方法自体が性能に影響を与える可能性が示唆される．例えば，Qwen3 での検証 [9] のように，“`Considering the limited time by the user, I have to give the solution based on the thinking directly now\n</think>\n\n`” といった自然な終了フレーズを用いる方法が考えられるが，今後は他の終了手法を含めた多様な対策の検討が必要である．

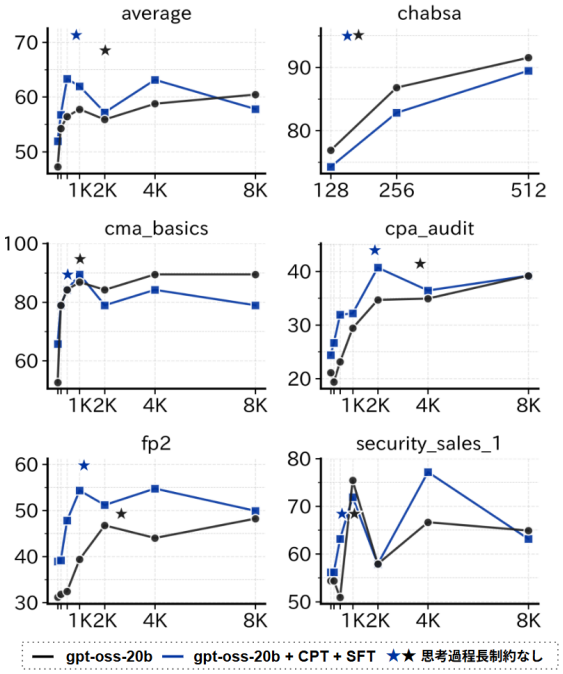

図 2　思考過程長に伴うタスク正解率の変化

5　おわりに

本研究では，特定ドメインにおける LLM の推論能力向上を目的とした，合成指示データセット構築の汎用的なフレームワークを提案した．提案手法の実証として，日本の金融分野を対象に総計約 95 億トークンの思考過程付きデータセットを構築し，評価を行った結果，公式指示モデルを上回る性能を達成した．これにより，トピックワードを起点とした合成データ生成が，ドメイン適応において極めて有効であることが示された．また，ドメイン特化タスクにおける思考過程長の分析から，1024 トークン付近までは性能が向上するものの，それ以降は頭打ちになる傾向を確認した．本手法は金融分野に限らず，専門性の高い他ドメインへの展開も容易であり，分野特化型 LLM 構築の基盤技術として貢献することが期待される．

謝辞

この成果は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の助成事業「ポスト5G 情報通信システム基盤強化研究開発事業／競争力ある生成 AI 基盤モデルの開発（GENIAC）」の結果得られたものである．

また，本研究の構想にあたり，ご助言をいただいたスタンフォード大学 Assistant Professor の橋本龍範氏に感謝いたします．

参考文献

[1] Shijie Wu, Ozan Irsoy, Steven Lu, et al. BloombergGPT: A large language model for finance. **arXiv preprint arXiv:2303.17564**, December 2023.

[2] Issa Sugiura, Takashi Ishida, Taro Makino, Chieko Tazuke, Takanori Nakagawa, Kosuke Nakago, and David Ha. EDINET-Bench: Evaluating LLMs on complex financial tasks using Japanese financial statements. **arXiv preprint arXiv:2506.08762**, June 2025.

[3] OpenAI Team. OpenAI o1 system card. **arXiv preprint arXiv:2412.16720**, December 2024.

[4] NVIDIA. Nemotron-Post-Training-Dataset-v1. https://huggingface.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v1, 2025.

[5] Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, et al. Demystifying domain-adaptive post-training for financial LLMs. **arXiv preprint arXiv:2501.04961**, October 2025.

[6] Sangmin Lee, Suzie Oh, Saeran Park, et al. FINALE: Finance domain instruction-tuning dataset with high-quality rationales via chain-of-thought prompting. In **Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing (FinNLP) and 1st Agent AI for Scenario Planning (AgentAI)**, pp. 89–106, Jeju, South Korea, August 2024.

[7] Kota Tanabe, Masahiro Suzuki, Hiroki Sakaji, et al. JaFIn: Japanese financial instruction dataset. **arXiv preprint arXiv:2404.09260**, July 2024.

[8] DeepSeek-AI. DeepSeek-V3 technical report. **arXiv preprint arXiv:2412.19437**, 2024.

[9] Qwen Team. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[10] Niklas Muennighoff, Zitong Yang, Weijia Shi, et al. s1: Simple test-time scaling. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)**, pp. 20286–20332, Suzhou, China, January 2025. Association for Computational Linguistics.

[11] Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, et al. Does thinking more always help? Mirage of test-time scaling in reasoning models. In **NeurIPS 2025**, October 2025.

[12] NVIDIA. Nemotron-4 340B technical report. **arXiv preprint arXiv:2406.11704**, August 2024.

[13] OpenAI. GPT-Oss-120B & GPT-Oss-20B model card. **arXiv preprint arXiv:2508.10925**, 2025.

[14] 服部翔, 岡崎直観, 水木栄, 藤井一喜. Swallow コーパス v2: 教育的な日本語ウェブコーパスの構築. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[15] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to Japanese morphological analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)**, pp. 230–237, July 2004.

[16] Jia Li, Edward Beeching, Lewis Tunstall, et al. NuminaMath. https://huggingface.co/datasets/AI-MO/NuminaMath-CoT, 2024.

[17] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, et al. SmolLM2: When Smol goes big – data-centric training of a small language model. **arXiv preprint arXiv:2502.02737**, 2025.

[18] Haizhou Shi, Zihao Xu, Hengyi Wang, et al. Continual learning of large language models: A comprehensive survey. **ACM Comput. Surv.**, Vol. 58, No. 5, November 2025.

[19] Masanori Hirano. Construction of a Japanese Financial Benchmark for Large Language Models. In **Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing (FinNLP) and 1st Agent AI for Scenario Planning (AgentAI)**, pp. 1–9. Association for Computational Linguistics, May 2024.

[20] 平野正徳, 今城健太郎. 金融分野に特化した複数ターン日本語生成ベンチマークの構築. **Jxiv**, 2024.

A　データセット構築パイプラインのパラメータ詳細

本研究のデータセット構築パイプラインにおける，各ステップの詳細な設定パラメータを表 2 に示す．

表 2　データセット構築パイプラインの設定パラメータ一覧

ステップ	パラメータ項目	設定値
1. トピックワード選定	金融ドメインシードワード数	135
	汎用ドメイン（混合用）シードワード数	20
2. ユーザ質問文生成	生成数 / サブトピック（多肢選択式）	8
	生成数 / サブトピック（その他）	10
3. 指示データの拡張	拡張バリエーション数 / 質問（多肢選択式）	3
	拡張バリエーション数 / 質問（その他）	5
4. フィルタリング	最小単語数（これ未満を除外）	10
5. 複数ターン生成	最大ターン数	3

B　LLM-as-a-judge による品質フィルタリングの詳細

生成した複数ターンの対話的指示データセットに対し，LLM-as-a-judge を用いた品質評価およびフィルタリングを実施した．判定モデルには gpt-oss-120b を用い，各サンプルにおける LLM の応答が適切であるかを，表 3 に示す 5 つの評価観点から評価した．各観点について 1–5 点の 5 段階評価を行い，評価者モデルには客観的かつ厳格な基準での評価を指示した．

表 3　品質評価の観点と評価基準

観点	評価内容	評価基準（5 段階）
正確性	事実の正確さ，誤情報の有無	5: 完全に正確 / 3: 一部誤り有 / 1: 重大な誤情報
関連性・指示追従	プロンプトへの適切な対応，指示の遵守	5: 完全に適切 / 3: 一部未対応 / 1: 的外れ
有用性・網羅性	回答の有用性，情報の網羅性	5: 極めて有用 / 3: 普通 / 1: 不十分
推論・深度	推論の質，分析の深さ，論理的一貫性	5: 優れた推論 / 3: 普通 / 1: 論理破綻
安全性・適切性	安全性，倫理性，適切な表現	5: 完全に安全 / 3: 軽微な問題 / 1: 危険

データセットに対する評価の結果，各観点において 5 点を獲得した割合は，正確性 81.7%，関連性・指示追従 97.3%，有用性・網羅性 97.6%，推論・深度 91.2%，安全性・適切性 99.95%であった．正確性が最も低く，事実誤認やハルシネーションの発生が確認された．本研究では，SFT 用データの品質を最大限に担保するため，全 5 観点において 5 点を獲得したサンプルのみを採用する厳格なフィルタリング基準を設定した．この基準により，839,398 サンプルから 632,636 サンプル（75.4%）が選択され，残りの約 24.6%が除外された．除外されたサンプルには，repetition（同一フレーズの繰り返し），指示に対する不適切な応答，推論過程の破綻などが含まれていた．

C　作成したデータセットの詳細

学習に用いたデータセットのサンプル数，トークン数，マルチターンの有無の情報を表 4 に示す．

表 4　学習に用いた各データセットの統計量

データセット名	カテゴリ	サンプル数	総トークン数	平均トークン数 (/sample)				Multi-turn
				user	assistant	reasoning	total	
nri-fin-reasoning	金融/汎用	632,636	6,352,341,377	191	8,996	855	10,041	Yes
NuminaMath-CoT-modified	数学	88,727	141,678,621	83	508	1,006	1,596	No
Nemotron-Post-Training-math-modified	数学	694,168	2,999,298,386	169	640	3,512	4,320	No
smol-constraints-modified	指示追従	24,411	12,803,102	71	199	254	524	No
Total	—	**1,439,942**	**9,506,121,486**	**171.7**	**4,295.4**	**2,134.6**	**6,601.7**	—

D　金融ベンチマークによる評価方法の改変

japanese-lm-fin-harness は，公開リポジトリの実装では選択肢に対応する token の log-likelihood を用いて判定するが，思考過程を出力する指示モデルでは評価が不安定になるため，プロンプトを Chat 形式に改変し，回答を `\boxed{}` に含めて回答させるように改変した．また，近年のモデルは深い推論を安定させるために特定の temperature や top-p などのサンプリングパラメータを指定する傾向があるため，貪欲法による決定的デコーディングを行わず，Pass@K 形式による評価を実施した．

pfmt-bench-fin-ja は，公開リポジトリの実装と異なり，LLM-as-a-judge 用のモデルとしてより性能の高い GPT-5 mini (version: 2025-08-07) を利用した．