

# 蒸留による日英推論型大規模言語モデル構築戦略の探索

水木栄<sup>1,2,3</sup> 藤井一喜<sup>1</sup> 川村政貴<sup>1</sup> Nguyen Tien Dung<sup>1</sup> 片山結太<sup>1</sup> 齋藤幸史郎<sup>1</sup> 一瀬達矢<sup>1</sup> 宮本空<sup>1</sup>  
 松下直矢<sup>1</sup> 大井聖也<sup>1</sup> Youmi Ma<sup>1</sup> 太田晋<sup>1</sup> 大葉大輔<sup>1</sup> 高村大也<sup>2</sup> 横田理央<sup>1</sup> 岡崎直観<sup>1,2,4</sup>  
<sup>1</sup> 東京科学大学 <sup>2</sup> 産業技術総合研究所 <sup>3</sup> 株式会社ホットリンク <sup>4</sup> NII LLMC

## 概要

大規模言語モデル (LLM) に日英入力で深い推論を発現させるための蒸留レシピを体系的に検証する。推論型教師 LLM から、言語・学習ドメイン・教師 LLM・推論過程の有無等を制御して〈指示・推論過程・応答〉データを合成し、非推論型 LLM に対して教師あり微調整を行うことで、各要因の寄与を切り分けた。評価の結果、(i) 英語データの学習のみで日本語入力にも深い推論が発現し、(ii) 教師 LLM は性能と推論安定性に影響を与え、(iii) 汎用対話でも推論過程付き学習が有効であることを確認した。一方、推論過程なし日本語対話データの混在は日本語での推論発現を抑制し得ることが分かった。

## 1 はじめに

大規模言語モデル (LLM) は、問題の定式化・解法の探索・回答の検証を反復的に行う「深い推論」[1, 2] を伴う応答を生成することで、数学・科学・コード生成といった STEM 領域の高難度課題において飛躍的な性能向上を遂げた [3]。一方で、推論能力の獲得は英語のデータと評価を主軸に進んできたため英語以外の事例は乏しく [4, 5, 6]、日英両方で同等の推論挙動を効果的に発現させる学習指針は体系的に整理されていない。

推論能力を獲得する方法論は、(a) 推論型 LLM 教師の出力を教師あり学習 (SFT) で蒸留する [7, 8]、(b) 教師あり学習に加えて強化学習で最適化する [9, 10, 11]、(c) 強化学習のみにより獲得する [3, 12]、という系統に整理できる。本研究は、先行研究において高い有効性および再現性が報告されている (a) の推論蒸留 SFT に焦点を当てる。この枠組みは、教師 LLM や SFT データ合成プロセスを操作することで「何が推論発現や対話、言語理解の性能を左右するか」の要因分解が容易で、入力言語を含めた複数の実験因子の比較検証を行いやすい。したがって本稿の中心的な問いは、「日英入力で深い

推論を発現させ、STEM・汎用対話・言語理解を高度に解く推論型 LLM を獲得するために、推論蒸留 SFT をどのように設計すべきか」となる。

この問いに答えるため、本研究では推論過程を明示的に事前学習していない日本語継続事前学習モデルを出発点とし、推論型教師モデル (例: gpt-oss-120b [13]) に指示文を入力して〈指示・推論過程・応答〉の三つ組データを合成し、SFT を行う。その際に、入力言語、学習ドメイン、教師モデルの種別などを実験因子として統制し、SFT 実施後のモデルを高難度課題を含む日英ベンチマークで評価する。また推論型モデルは、推論過程で同じ文を無限に反復して応答に至らない「推論失敗」[14] がしばしば生じる。そこで、推論の安定性を</think>未閉鎖率 (推論失敗率) として測定し、推論能力と応答破綻の影響を切り分けて分析する。

実験の結果、STEM ドメインでは英語入力の推論過程を用いた蒸留のみでも日本語入力に対して深い推論挙動が発現することが確認された一方で、教師 LLM の種別や推論過程の長さがタスク性能と推論安定性に影響することが明らかになった。対話ドメインでは、汎用対話や指示追従にも推論過程付きの蒸留が有効であることが分かった。推論過程なしの日本語対話データを混合して学習させると、日本語入力での推論発現がタスクを問わず抑制されるというアンチパターンが存在し、既存の非推論型 SFT 資産を推論型に転用する際の注意点を見出した。

## 2 方法

推論蒸留 SFT の設計指針を得るため、蒸留にかかわる条件を実験因子として変更しながら SFT および評価を実施することで、各因子がタスク性能および推論安定性に及ぼす影響を検証する。

### 2.1 因子制御による SFT データの合成

**入出力形式** 各事例を〈指示, 推論過程, 応答〉の三つ組とし、推論過程は <think>...</think> でタ

グ付けする。

**合成方法** 学習ドメインの影響を検証するため、数学、科学、コード生成の設問からなる STEM ドメインおよび、プログラミング・創作・質問応答など多様な対話からなる Chat ドメインの指示文をシードにデータ合成を行う。STEM ドメインの指示は Nemotron-Post-Training-Dataset-v1 [15, 10] (NPT)<sup>1)</sup>から、Chat ドメインの指示は LMSYS-Chat-1M [16] (LMSYS) から取得した。言語の影響を考察するため、元の英語指示文を機械翻訳した日本語版指示文を用意した。英語 STEM については、NPT に元から存在する DeepSeek-R1-0528 [3] (R1) の推論過程・応答のほか、gpt-oss-120b [13] (OSS) を用いて合成した。日本語 STEM および日英 Chat は、OSS を用いて合成した。なお、gpt-oss は日本語による推論が不可能であるため、推論過程の言語は英語のみとした。また推論過程長の影響を検証するために、英語 STEM については OSS の思考の深さ<sup>2)</sup>を medium (OSS) および high (OSS-h) の 2 通りに設定して合成した。

**合成結果** 各実験におけるデータセット構成および、合成した SFT データの諸元を表 1 に示す。推論過程+応答の長さはドメインおよび教師 LLM によって大きく異なり、Chat よりも STEM, OSS よりも R1 および OSS-h が顕著に長くなった。データを観察したところ、数学やコード生成では、R1 および OSS-h は別解の検討や回答の検算を繰り返すことで長い推論過程が生成されることを確認した (付録 1)。指示の言語による違いは比較的小さかった。

## 2.2 教師あり微調整/SFT

本研究では、非推論型の日本語 LLM に対して事後学習で推論能力を獲得させる状況を想定する。そこで SFT 学習対象の LLM は、推論過程を含めないレシピ (付録 B) を用いて Qwen3-8B-Base [17] を継続事前学習して、日本語 QA および機械翻訳を主に強化したモデル (以下、Base モデルと表記) を用いる。Qwen3-8B-Base および Base モデルは、推論過程を含まない応答を生成することを確認した。

教師あり微調整 (SFT) は、データセット構成の異なる 10 通りの条件で実施する (表 1; SFT の詳細は付録 C)。学習事例数はドメインごとに一定とし、日英を併用する構成では各言語に等分する。

1) code, math, stem スプリットから 2:2:1 の割合で取得。

2) reasoning\_effort パラメータで指定する。

表 1: 実験条件一覧 (I/O: 入出力言語, Trace: 推論過程言語, 長さ: 平均トークン数, PPL: 学習前モデルで測定した推論過程+応答のパープレキシティ)

ID	ドメイン	教師	I/O	Trace	事例数	長さ	PPL
1	STEM	OSS	En	En	500k	2.3k	2.74
2	STEM	OSS	Ja	En	500k	2.1k	2.40
3	STEM	R1	En	En	1,500k	9.3k	2.00
4	STEM	R1	En	En	500k	9.3k	2.00
5	STEM Chat	R1 OSS	En Ja	En En	500k 100k	9.3k 1.4k	2.00 2.48
6	STEM Chat	R1 OSS	En En	En En	500k 100k	9.3k 0.8k	2.00 4.29
7	STEM Chat	R1 OSS	En Ja	En —	500k 100k	9.3k 1.2k	2.00 —
8	Chat Chat	OSS OSS	En Ja	En En	50k 50k	0.8k 1.4k	4.29 2.48
9	STEM	OSS-h	En	En	500k	7.7k	2.00
10	STEM STEM Chat Chat	R1 OSS OSS OSS	En Ja En Ja	En En En En	250k 250k 50k 50k	9.3k 2.1k 0.8k 1.4k	2.00 2.40 4.29 2.48

## 2.3 評価

**ベンチマーク** 評価対象のタスクおよびベンチマークは、大学院レベルの科学 GPQA [18], 競技数学 AIME [19], MATH-500 [20, 21], 競技プログラミング LiveCodeBench [22], 一般教養 MMLU-Pro [23] および、これらの日本語版として GPQA-Ja [24], MATH-100-Ja [25], JHumanEval [26], MMLU-ProX-Ja [27] を用いて、高難度の STEM や教養を日英で出題した場合の性能を評価する。さらに、日本固有知識 JamC-QA [28], 常識推論 HellaSwag [29], 機械翻訳 WMT20 [30], 対話 MT-Bench [31], 指示追従 M-IFEval-Ja [32] を用いて包括的に評価する (表 5)。

**詳細** ゼロショットの指示応答形式で出題する。正誤判定は、回答文字列を応答から抽出して正解と照合する。ただし</think>未閉鎖時は推論過程から途中回答を抽出する。実装は swallow-evaluation-instruct [33] を用いる。応答文生成は、推論型モデル評価の慣習 [9, 17] に倣って、温度 0.6, Top-p 0.95 の確率的デコーディングを主に用いる。コンテキスト長は 32,768 トークンとする。ベンチマーク評価と並行して</think> が未閉鎖の設問の割合を推論失敗率として定義し、推論安定性の定量指標とする。

表 2: 実験 ID ごとのベンチマーク結果. 略記は右の通り; MT-Bench: MTB, M-IFEval-Ja: MIF<sub>Ja</sub>, JamC-QA: JamC, WMT20 En-Ja: En/Ja, WMT20 Ja-En: Ja/En, MMLU-Pro-X-Ja: MPX<sub>Ja</sub>, GPQA: GPQ, MATH-100-Ja: M100<sub>Ja</sub>, JHumanEval: JHE, HellaSwag: HS, MMLU-Pro: MP, MATH-500: M500, LiveCodeBench: LCB. †B は学習前モデル (Base モデル) の結果.

ID	↑日本語-ベンチマークスコア									↑英語-ベンチマークスコア						
	MTB <sub>Ja</sub>	MIF <sub>Ja</sub>	JamC	En/Ja	Ja/En	MPX <sub>Ja</sub>	GPQ <sub>Ja</sub>	M100 <sub>Ja</sub>	JHE	MTB	HS	MP	GPQ	M500	AIME	LCB
B <sup>†</sup>	0.666	0.563	0.433	0.236	0.224	0.542	0.375	0.758	0.707	0.625	0.723	0.564	0.404	0.776	0.150	0.013
1	0.795	0.429	0.423	0.200	0.191	0.638	0.426	0.859	0.801	0.793	0.794	0.655	0.472	0.882	0.379	0.296
2	0.825	0.389	0.403	0.205	0.196	0.630	0.458	0.862	0.796	0.767	0.801	0.650	0.486	0.858	0.378	0.260
3	0.636	0.350	0.368	0.160	0.038	0.413	0.266	0.579	0.721	0.692	0.760	0.660	0.527	0.948	0.632	0.458
4	0.685	0.341	0.422	0.166	0.099	0.492	0.317	0.645	0.677	0.697	0.728	0.644	0.504	0.940	0.505	0.362
5	0.840	0.544	0.420	0.214	0.209	0.604	0.391	0.812	0.759	0.737	0.701	0.655	0.511	0.938	0.516	0.366
6	0.812	0.504	0.438	0.209	0.202	0.611	0.402	0.818	0.756	0.797	0.779	0.576	0.442	0.930	0.508	0.347
7	0.760	0.425	0.345	0.211	0.204	0.530	0.373	0.686	0.693	0.736	0.675	0.650	0.499	0.938	0.525	0.377
8	0.804	0.500	0.425	0.218	0.203	0.600	0.388	0.816	0.756	0.816	0.800	0.636	0.435	0.782	0.152	0.220
9	0.641	0.367	0.412	0.198	0.191	0.638	0.453	0.872	0.820	0.562	0.803	0.683	0.515	0.884	0.434	0.204
10	0.818	0.504	0.412	0.208	0.199	0.632	0.449	0.859	0.787	0.817	0.784	0.631	0.469	0.902	0.471	0.252

表 3: 実験 ID ごとの推論失敗率. 略記は表 2 に従う.

ID	↓日本語-推論失敗率									↓英語-推論失敗率						
	MTB <sub>Ja</sub>	MIF <sub>Ja</sub>	JamC	En/Ja	Ja/En	MPX <sub>Ja</sub>	GPQ <sub>Ja</sub>	M100 <sub>Ja</sub>	JHE	MTB	HS	MP	GPQ	M500	AIME	LCB
1	0.414	0.256	0.020	0.040	0.007	0.017	0.029	0.015	0.012	0.228	0.000	0.015	0.038	0.014	0.103	0.059
2	0.104	0.140	0.032	0.050	0.005	0.018	0.036	0.013	0.013	0.198	0.000	0.015	0.030	0.024	0.091	0.052
3	0.320	0.180	0.376	0.060	0.007	0.610	0.815	0.513	0.212	0.131	0.001	0.027	0.112	0.030	0.291	0.260
4	0.496	0.163	0.236	0.040	0.001	0.412	0.672	0.379	0.271	0.126	0.000	0.030	0.097	0.032	0.342	0.382
5	0.043	0.110	0.061	0.004	0.003	0.058	0.163	0.036	0.066	0.151	0.000	0.028	0.096	0.026	0.351	0.386
6	0.118	0.110	0.078	0.007	0.005	0.059	0.194	0.040	0.091	0.070	0.001	0.069	0.289	0.034	0.355	0.331
7	0.103	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.245	0.000	0.028	0.095	0.030	0.355	0.365
8	0.069	0.169	0.033	0.003	0.002	0.040	0.094	0.022	0.007	0.045	0.000	0.039	0.106	0.042	0.218	0.075
9	0.583	0.645	0.297	0.128	0.041	0.155	0.420	0.087	0.116	0.551	0.023	0.154	0.510	0.122	0.551	0.742
10	0.041	0.081	0.014	0.004	0.002	0.013	0.029	0.026	0.029	0.066	0.000	0.019	0.069	0.030	0.346	0.105

### 3 実験結果

ベンチマークのスコアを表 2 に, 推論失敗率を表 3 に示す. 以下では, 入力言語, 学習ドメイン, 教師 LLM などの実験因子 (表 1) が異なる実験同士を比較することで, 因子の影響を考察する.

#### 3.1 STEM ドメインデータの効果

英語 STEM データを学習すると (実験 1 vs. Base), MP, GPQ, AIME, LCB が 7 から 28pt 改善する. またこれらの日本語版である MPX-Ja, GPQ-Ja, M100-Ja, JHE も 10pt 改善する. したがって英語で STEM データを学習することで, 英語は当然ながら, 日本語で出題しても深い推論挙動が発現することが示唆される. さらに, 日本語 STEM データに変更して学習した場合 (実験 2 vs. 実験 1) にも, 上述した英日 STEM タスクのスコアおよび推論失敗率は変わらな

い. したがって STEM ドメインにおける推論過程の学習は言語間転移が生じること, それゆえ日本語で学習する効果は限定的だと考えられる.

STEM ドメインの学習事例数を 3 倍にすると (実験 3 vs 実験 4), AIME が 13pt, LCB が 10pt 改善し, 数学・コード生成でのスケール則 [10, 7] が確認できた. 一方で, 日本語で出題する MPX-Ja, GPQ-Ja, M100-Ja は推論失敗率の悪化に伴ってスコアが低下した. したがって英語 STEM 学習を大規模化する場合は, 日本語入力時の推論安定性を損なう副作用に注意が必要である.

#### 3.2 Chat ドメインデータの効果

対話タスク MTB および指示追従タスク MIF-Ja では, 対象言語の Chat データを学習すると推論失敗率が低下し, それに伴ってスコアが改善する<sup>3)</sup>. たど

3) 推論失敗時は推論過程が採点対象になるためである.

例えば日本語 Chat データを学習すると (実験 5 vs. 実験 4) MTB-Ja が 15pt, MIF-Ja が 20pt 改善し, 英語 Chat データを学習すると (実験 6 vs. 実験 4) MTB が 10pt 改善する. 次に, 日本語 Chat データから推論過程を除いて指示・応答ペアのみを学習すると (実験 7 vs. 実験 5), MTB-Ja が 8pt, MIF-Ja が 12pt 低下する. したがって汎用対話や指示追従でも深い推論挙動が効果的だと示された. また英語 STEM データも学習しているにもかかわらず, 日本語 STEM タスク (MPX-Ja, GPQ-Ja, M100-Ja, JHE) の性能改善効果は失われ (実験 7 vs. Base), 推論失敗率はゼロになった. すなわち, 指示・応答のみの Chat データを混ぜて学習すると, ドメイン転移によって STEM の設問でも深い推論挙動が抑制されることが示唆された. したがって, 非推論型モデル向けの既存の SFT データを推論型に転用する場合は, 推論過程を付与して再構築する必要があると考えられる.

最後に, Chat データのみを学習すると, GPQ, M500, AIME は改善しなかった (実験 8 vs. Base). Chat データの出典である LMSYS は多様なドメインを包含している [16] が, それでも高難度の数学・科学を解くには不十分であることが示された.

### 3.3 常識推論・質問応答・翻訳への影響

常識推論 HS では, STEM データのみの学習 (実験 1 vs. Base) および Chat データのみの学習 (実験 8 vs. Base) の両方において約 8pt 改善した. したがって, これらのドメインで学習した深い推論は常識推論にも有効だと示唆される. 一方で, 日本知識 QA および機械翻訳 (JamC, En/Ja, Ja/En) は全実験で Base モデルを下回った. したがって STEM・Chat ドメインで深い推論を学習しても, 百科事典的知識に基づく質問応答や機械翻訳には効果がないといえる.

### 3.4 教師 LLM の影響

教師 LLM を R1 から OSS に変更して英語 STEM データを学習すると (実験 1 vs. 実験 4), MPX-Ja, GPQ-Ja, M100-Ja, GPQ, AIME, LCB の推論が顕著に安定化する一方で, 性能面は AIME および LCB の悪化と MTB の改善が混在した. さらに教師 LLM を OSS-h に変更すると (実験 9 vs. 実験 4), GPQ, AIME, LCB の推論は R1 の場合よりも不安定化した. したがって, 教師 LLM の違いはタスク性能および推論安定性に影響するが, R1 対 OSS というモデル種別の違いのみで決まるわけではなく, 推論過程の長さ

表 4: 英語ベンチマークの推論失敗率に対する実験因子の影響.  $\pm$  は影響の正負を,  $\{*, **, ***\}$  は, 偏回帰変数ゼロを帰無仮説とする検定の有意水準であり, 順に 10%, 5%, 2.5% に対応する.  $\dagger$  は教師 LLM が OSS(-h) か否か,  $\ddagger$  は Chat ドメインの学習有無.

	HS	MP	GPQ	M500	AIME	LCB	MTB
失敗率平均	0.003	0.044	0.150	0.039	0.294	0.266	0.174
回帰 $R^2$	0.44	0.37	0.35	0.99	0.77	0.78	0.33
教師 $\dagger$	***	***		***			**
PPL	***	***		***			
長さ	***	***	+	***	**	**	
Chat 学習 $\ddagger$							

なども影響する複雑な関係だと示唆される.

この複雑な関係を解きほぐすため, 表 1 の実験因子を説明変数に, 英語タスクの推論失敗率を目的変数にした線形回帰のモデル選択を行うことにより, 一連の実験結果を最もよく説明する実験因子の組み合わせを特定した (表 4, 詳細は付録 D). その結果, 推論過程+応答が長いほど推論失敗率が高くなる傾向が見出された. 一方で, STEM データに用いた教師 LLM が OSS 系列か否か, およびパープレキシティといった他の因子は, もともと推論失敗率が低い HS, MP, M500 のみと関連しており絶対的な影響は小さい. したがって実用的かつ STEM 重視のモデル構築では, 推論過程の長さに注意して SFT データを合成することが重要だと考えられる.

## 4 おわりに

本稿では, 推論型モデルの蒸留によって日英入力で深い推論が効果的に発現する LLM を構築する戦略を, 入力言語, 学習ドメイン (STEM および Chat), 教師 LLM, 推論過程の有無, 学習データ規模の観点から検証した. その結果, STEM ドメインは深い推論が言語間転移するため, 日本語での学習効果は限定的であると示唆された. また, 教師 LLM の選択は性能と推論安定性に影響し, 特に長い推論過程が安定性を損なうことを示した. Chat ドメインについては, 汎用対話・指示追従でも深い推論が有効であること, 学習した言語での推論が安定すること, 推論過程なしで学習すると非対話タスクでも推論発現を抑制するため要注意であることを示した.

今後は, 事前学習や強化学習も含めた全体最適な推論型モデル構築戦略の探索および, 日本語で思考する推論型モデルの実現に取り組む.

## 謝辞

本研究は、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の助成事業（JPNP25006）、産総研政策予算プロジェクト「フィジカル領域の生成 AI 基盤モデルに関する研究開発」および、文部科学省の補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けた。また、産総研及び AIST Solutions が提供する ABCI 3.0「ABCI 3.0 開発加速利用」支援ならびに東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した。

## 参考文献

- [1] Philip Lippmann and Jie Yang. Style over substance: Distilled language models reason via stylistic replication. In **Second Conference on Language Modeling**, 2025.
- [2] Kanishk Gandhi, Ayush K Chakravarthy, Anikait Singh, Nathan Lile, and Noah Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective STaRs. In **Second Conference on Language Modeling**, 2025.
- [3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shitong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, Vol. 645, No. 8081, pp. 633–638, 2025.
- [4] Minjun Kim, Hyeonseok Lim, Hangeul Yoo, Inho Won, Seungwoo Song, Minkyung Cho, Junhun Yuk, Changsu Choi, Dongjae Shin, Huije Lee, Hoyun Song, Alice Oh, and Kyungtae Lim. KORMo: Korean open reasoning model for everyone. arXiv:2510.09426, 2025.
- [5] Kunat Pipatanakul, Pittawat Taveekitworachai, Potsawee Manakul, and Kasima Tharnpipitchai. Adapting language-specific LLMs to a reasoning model in one day via model merging – an open recipe. arXiv:2502.09056, 2025.
- [6] Changjiang Gao, Zixian Huang, Jingyang Gong, Shujian Huang, Lei Li, and Fei Yuan. Llamax2: Your translation-enhanced model also performs well in reasoning. arXiv:2510.09189, 2025.
- [7] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raouf, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, and et al. OpenThoughts: Data recipes for reasoning models. arXiv:2506.04178, 2025.
- [8] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 20275–20321, 2025.
- [9] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and et al. OLMo 3. arXiv:2512.13961, 2025.
- [10] Llama-Nemotron Team. Llama-nemotron: Efficient reasoning models. arXiv:2505.00949, 2025.
- [11] Zhoujun Cheng, Richard Fan, Shibo Hao, Taylor W. Killian, Haonan Li, Suqi Sun, Hector Ren, Alexander Moreno, Daqian Zhang, Tianjun Zhang, and et al. K2-think: A parameter-efficient reasoning system. arXiv:2509.07604, 2025.
- [12] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. OctoThinker: Mid-training incentivizes reinforcement learning scaling. arXiv:2506.20512, 2025.
- [13] OpenAI. gpt-oss-120b & gpt-oss-20b model card. arXiv:2508.10925, 2025.
- [14] Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages. arXiv:2508.14828, 2025.
- [15] Dhruv Nathawani, Igor Gitman, Somshubra Majumdar, Evelina Bakhturina, Ameya Sunil Mahabaleshwar, Jian Zhang, and Jane Polak Scowcroft. Nemotron-Post-Training-Dataset-v1. <https://huggingface.co/datasets/nvidia/Nemotron-Post-Training-Dataset-v1>, 2025.
- [16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In **Twelfth International Conference on Learning Representations**, 2024.
- [17] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and et al. Qwen3 technical report. arXiv:2505.09388, 2025.
- [18] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In **First Conference on Language Modeling**, 2024.
- [19] Art of Problem Solving Wiki: American invitational mathematics examination (aime). <https://artofproblemsolving.com/wiki/>, 2026.
- [20] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical prob-

- lem solving with the MATH dataset. In **Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks**, 2021.
- [21] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. arXiv:2305.20050, 2023.
  - [22] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In **The Thirteenth International Conference on Learning Representations**, 2025.
  - [23] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhramil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In **Advances in Neural Information Processing Systems**, Vol. 37, pp. 95266–95290, 2024.
  - [24] Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. BenchMAX: A comprehensive multilingual evaluation suite for large language models. In **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 16751–16774, 2025.
  - [25] Guijin Son, Jiwoo Hong, Hyunwoo Ko, and James Thorne. Linguistic generalizability of test-time scaling in mathematical reasoning. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 14333–14368, 2025.
  - [26] 佐藤美唯, 高野志歩, 梶浦照乃, 倉光君郎. LLM は日本語追加学習により言語間知識転移を起こすのか? 言語処理学会第 30 回年次大会 (NLP2024), 2024.
  - [27] Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjie Wang, Fan Gao, Jinghui Lu, Yang Jiang, Huitao Li, et al. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 1513–1532, 2025.
  - [28] 岡照晃, 柴田知秀, 吉田奈央. JamC-QA: 日本固有の知識を問う多肢選択式質問応答ベンチマークの構築. 言語処理学会第 31 回年次大会 (NLP2025), pp. 839–844, 2025.
  - [29] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4791–4800, 2019.
  - [30] Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, and Eric et al. Joanis. Findings of the 2020 conference on machine translation. In **Proceedings of the Fifth Conference on Machine Translation**, pp. 1–55, 2020.
  - [31] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and Eric P. Xing et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In **Annual Conference on Neural Information Processing Systems 2023**, pp. 46595–46623, 2023.
  - [32] Antoine Dussolle, Andrea Cardena Díaz, Shota Sato, and Peter Devine. M-IFEval: Multilingual instruction-following evaluation. In **Findings of the Association for Computational Linguistics: NAACL 2025**, pp. 6161–6176, 2025.
  - [33] Sakae Mizuki, Koshiro Saito, Masanari Ohi, Tatsuya Ichinose, Naoya Matsushita, Sora Miyamoto, Tien Dung Nguyen, and Sangwhan Moon. 大規模言語モデルの評価フレームワーク swallow-evaluation-instruct v202510. <https://github.com/swallow-llm/swallow-evaluation-instruct>, 2025.
  - [34] Koshiro Saito, Sakae Mizuki, Masanari Ohi, Taishi Nakamura, Taihei Shiotani, Koki Maeda, Youmi Ma, Kakeru Hattori, Kazuki Fujii, Takumi Okamoto, Shigeki Ishida, Hiroya Takamura, Rio Yokota, and Naoaki Okazaki. Why we build local large language models: An observational analysis from 35 japanese and multilingual llms. arXiv:2412.14471, 2025.
  - [35] 服部翔, 岡崎直観, 水木栄, 藤井一喜, 中村泰士, 大井聖也, 塩谷泰平, 齋藤幸史郎, Youmi Ma, 前田航希, 岡本拓己, 石田茂樹, 横田理央, 高村大也, 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), pp. 94–99, 2025.
  - [36] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsang Loem, Rio Yokota, and Sakae Mizuki. Building a large Japanese web corpus for large language models. In **First Conference on Language Modeling**, 2024.
  - [37] Xinyi Zhao, Masafumi Hamamoto, and Hiromasa Fujihara. Laboro-paracorus: A web-based japanese-english parallel corpus. <https://github.com/Laboroai/Laboro-ParaCorpus>, 2021.
  - [38] Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia. <https://huggingface.co/datasets/HuggingFaceTB/cosmopedia>, 2024.
  - [39] Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostafa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics**, pp. 2459–2475, 2025.
  - [40] Kazuki Fujii, Yukito Tajima, Sakae Mizuki, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Masanari Ohi, Masaki Kawamura, Taishi Nakamura, Takumi Okamoto, Shigeki Ishida, Kakeru Hattori, Youmi Ma, Hiroya Takamura, Rio Yokota, and Naoaki Okazaki. Rewriting pre-training data boosts llm performance in math and code. arXiv:2505.02881, 2025.
  - [41] Clifford M Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, Vol. 76, No. 2, pp. 297–307, 1989.

## A 評価ベンチマークの詳細

表 5: 評価ベンチマーク一覧

言語	タスク	ベンチマーク名	評価尺度	設問数	生成パラメータ†
日	汎用対話	MT-Bench-Ja	LLM-as-a-Judge	80 問 ×2 ターン	N=5, Tは既定値
日	指示追従	M-IFEval-Ja	正解率/指示	172 問, 226 指示	T=0.6, P=0.95
日	日本知識 QA	JamC-QA	正解率/択一	2,309	T=0.6, P=0.95
日	英日翻訳	WMT20 En/Ja	BLEU	1,000	T=0.6, P=0.95
日	日英翻訳	WMT20 Ja/En	BLEU	993	T=0.6, P=0.95
日	一般教養	MMLU-Pro-X-Ja	正解率/択一	11,759	T=0.6, P=0.95
日	科学	GPQA-Ja	正解率/択一	448	T=0.6, P=0.95
日	数学	MATH-100-Ja	正解率	99	T=0.6, P=0.95, N=16
日	コード生成	JHumanEval	Pass@1	164	T=0.2, P=0.95, N=10
英	汎用対話	MT-Bench-En	LLM-as-a-Judge	80 問 ×2 ターン	N=5, Tは既定値
英	常識推論	HellaSwag	正解率/択一	10,042	T=0.6, P=0.95
英	一般教養	MMLU-Pro	正解率/択一	12,032	T=0.6, P=0.95
英	科学	GPQA (Diamond)	正解率/択一	198	T=0.6, P=0.95, N=16
英	数学	MATH-500	正解率	500	T=0.6, P=0.95
英	競技数学	AIME 2024-2025	正解率	60	T=0.6, P=0.95, N=16
英	競プロ	LiveCodeBench v5-6 の新設問	Pass@1	342	T=0.6, P=0.95, N=10

LLM-as-a-Judge は gpt-4o-2024-08-06 を用いた。非推論型モデルである Base モデルの場合のみ、T=0, N=1 に設定して評価した (MT-Bench, JHumanEval, LiveCodeBench を除く)。

## B 継続事前学習の詳細

本研究で SFT の学習対象とした LLM は、Qwen3-8B-Base [17] を日本語・英語・数学・ソースコード等のテキスト約 52B トークンを用いて継続事前学習したモデルである。学習に用いたコーパス (表 6) の主な内訳は、日本語ウェブテキストおよびウェブテキストをシードとする合成テキストが約半分を占める。また、数学テキストおよびソースコードがそれぞれ約 10%、本実験 5 で用いた英語 STEM データおよび日本語 Chat データから推論過程を除去した指示応答形式のデータ (表 6 ではそれぞれ NPT-R1-En-wo-Trace, LMSYS-OSS-Ja-wo-Trace と表記) が 15% 弱、および対訳コーパスが 1% 弱を占める。継続事前学習の主な効果は、日本固有知識の獲得 [34] (JamC-QA +7pt)、翻訳の改善 [34] (WMT20 En/Ja +6pt)、および一般教養・科学の強化 [35] (MMLU-Pro +13pt, GPQA +7pt) だった。ハイパーパラメータは、学習率を  $1.5 \times 10^{-5}$ 、コンテキスト長さを 32,768 とした。

表 6: 事前学習データの構成

言語	タイプ	テキスト名	割合 [%]
日	ウェブ	日本語 Wikipedia	1.1
日	ウェブ (合成)	Swallow コーパス v3 [36] + 日英 QA [35]	43.4
日	日英対訳	Laboro ParaCorpus [37]	0.3
日	日英対訳	kaken-trans-ja-en <sup>3)</sup>	0.3
日	Chat/指示応答	LMSYS-OSS-Ja-wo-Trace	0.1
英	ウェブ	英語 Wikipedia	2.1
英	ウェブ (合成)	Cosmopedia [38]	3.0
英	ウェブ	Nemotron-CC-HQ actual [39]	14.8
英	STEM/指示応答	NPT-R1-En-wo-Trace	13.7
英	STEM/QA	SwallowCode-v2 [40] <sup>4)</sup>	10.1
英	STEM/QA	SwallowMath-v2 [40] <sup>4)</sup>	11.3

## C 教師あり微調整/SFT の詳細

SFT のハイパーパラメータは、学習率を  $1.5 \times 10^{-5}$ 、エポック数を 1、バッチサイズを 64、コンテキスト長さを 32,768 とした。コンテキストの余剰はパディングした。

## D 実験因子・推論失敗率の統計分析

表 1 の実験因子を説明変数、英語タスクの推論失敗率を目的変数とする線形回帰に対してモデル選択を行った。これにより、実験 8 を除く一連の実験結果を最もよく説明する実験因子の組み合わせを特定した。具体的には、英語ベンチマーク  $b$  の推論失敗率  $Y_b$  を目的変数として、Chat データの学習の有無  $X_c \in \{0, 1\}$ , STEM

- 1) <https://huggingface.co/datasets/hprcr/kaken-trans-ja-en>
- 2) <https://huggingface.co/datasets/tokyotech-llm/swallow-code-v2>, <https://huggingface.co/datasets/tokyotech-llm/swallow-math-v2>

データ教師が OSS または OSS-h である  $X_m \in \{0, 1\}$ , STEM データのパープレキシティ  $X_p$ , STEM データの推論過程+応答の長さ  $X_l$  の最大 4 つを説明変数とする切片つき線形回帰モデル

$$Y_b = \beta_{b,c} X_c + \beta_{b,m} X_m + \beta_{b,p} X_p + \beta_{b,l} X_l + \alpha_b \quad (1)$$

を考えた。なお日英 STEM データを併用した実験 10 は、 $X_m, X_p, X_l$  は日英の平均とした。次に、説明変数が 1 個から 4 個のすべての組み合わせ 15 通りをモデル候補として、推論過程を学習していない実験 8 を除く 9 個の実験結果を対象として最尤推定を行った。最後に、各モデル候補について小サンプル補正つき赤池情報量基準 AICc [41] を計算し、AICc が最小のモデルで採用された説明変数を、実験結果を最もよく説明する実験因子の組み合わせとして報告した。

教師 LLM のバリエーション (R1, OSS, OSS-h の 3 通り) が存在するのは英語 STEM データのみであるため、本分析は英語ベンチマークのみを対象とした。日本語 STEM データおよび日本語 Chat データの教師 LLM は OSS に固定されており、他の実験因子との効果を分離して評価できないためである。

## E 推論過程の定性分析

**Question:** What is the least total number of pens that can be distributed into 49 packages such that each package contains a different number of pens, and no package is left empty?

**DeepSeek-R1-0528 (R1):** I need to find the least total number of pens to distribute into 49 packages, each with a different number of pens, and no package empty. [...] Let me calculate that: the sum of the first n positive integers is  $n(n+1)/2$ . So for  $n=49$ , it's  $49*50/2 = 49*25 = 1225$ . So is 1225 the answer? But the problem says [...], but I think 1225 should work. But let me make sure. [...], so 1 to 49 should be the smallest possible sum. [...] I think I'm overthinking. But let me see if there's a smaller sum possible. [...] Another idea: perhaps the pens are distributed. [...]  $49*25 = 50*25=1250$ , minus  $25=1225$ , yes. So I think that's it.

**gpt-oss-120b medium (OSS):** We need to find minimal total number of pens such that we can have 49 packages, each with distinct positive integer number of pens, none empty. [...] That's just the sum of the first 49 positive integers:  $1+2+...+49 = 49*50/2 = 1225$ . But maybe they ask [...] So answer 1225. Check if any other constraints like each package must have at least one pen, distinct counts. So yes. Thus answer: 1225 pens.

**gpt-oss-120b high (OSS-h):** We need to interpret the problem: "What is the least total number of pens that can be distributed into 49 packages such that each package contains a different number of pens, and no package is left empty?" [...] The minimal sum is achieved by using the smallest possible distinct positive integers: 1,2,3,...,49. Sum =  $49*50/2 = 1225$ . So answer is 1225 pens. [...] But perhaps there is a twist: maybe the packages are identical? [...] But maybe they want to consider that the packages are unlabeled, [...] Thus answer: 1225. But maybe they want to consider that [...] Thus answer: 1225. [...] Thus final answer: 1225. Ok. I'll answer: 1225.

図 1: 教師 LLM ごとに合成された推論過程の例。

図 1 に、数学の設問 (指示文) に対してそれぞれの教師 LLM が合成した推論過程を示す。すべての教師 LLM に共通する定性的な特徴として、問題の定式化、解法の探索、別解や前提の再検討、回答の検算によって推論過程が構成されること、および “but” のような談話標識を介して各段階を行き来することが挙げられる。DeepSeek-R1-0528 (R1) および、思考の深さを high に設定した gpt-oss-120b (OSS-h) では、別解・前提の再検討および回答の検算が頻繁に現れる。これにより “Thus answer: 1225. But maybe ...” のような特定の表現が繰り返し出現するとともに、しばしば 1 万トークン以上の長い推論過程が生成される。また、推論過程+応答のパープレキシティは OSS よりも R1 および OSS-h のほうが低い (表 1)、すなわち長い推論過程は続きを予測しやすい特徴がある。これは、R1 や OSS-h が前提の再検討や回答の検算において同じような表現を繰り返すためだと思われる。

コーディングの設問に対する推論過程は、解法の探索に類似コードが現れる点を除けば、数学の設問に対する推論過程と類似の構成だった。科学の多肢択一式設問に対する推論過程は、選択肢の逐次検討および、関連知識の想起を行う様子が観察された。