品質と網羅性に基づくデータ選択による言語モデルの効率的学習

Haocheng Zhu[1]　王 天奇[1]　鈴木 潤[1,2,3]
[1] 東北大学　[2] 理化学研究所　[3] 国立情報学研究所 LLMC
is-failab-research@grp.tohoku.ac.jp

概要

教師あり微調整（SFT）の学習データ選択は，学習コストを抑えつつ性能を確保するうえで重要である．本稿では，学習前後のモデルにおけるクロスエントロピー（CE）の変化から訓練事例の学習価値を推定し，埋め込み表現の類似度に基づく冗長性除去と組み合わせた選択手法を検討する．OLMo-2-1B の SFT データセットで実験した結果，単純な上位選択よりも，類似度除去を併用することで，少量データでもランダム選択と同等以上の性能が得られる傾向を確認した．また，類似度推定に用いるテキスト範囲や埋め込み構成の違いが結果に影響することを示す．

1　はじめに

近年，大規模言語モデル（LLM）は多様なタスクで高性能を示し [1]，実応用に向けた研究開発が進展している．実用化には事前学習後の教師あり微調整（Supervised Fine-Tuning; SFT）[2] が重要である一方，SFT データの規模 [3] は拡大しており，計算資源と学習コストが課題となる．さらに，データ量の増加が常に性能向上に直結するとは限らず，少量でも設計の良い事例が推論能力を大きく引き上げ得ること [4] や，ノイズの種類により学習効率や性能への影響が変わること [5] も報告されている．

このため，限られた学習予算の下で性能を引き出すデータ選択が重要となる．SFT データは人手作成や品質管理が施される場合が多く，Web 由来データに比べればノイズは抑えられているが，「有効な訓練事例」を単一指標で安定に同定することは容易ではない．評価軸が限定的だと，良質な事例を拾えても，ドメインや難易度など別側面の偏りが残り，サブセット全体の性能に影響し得る．

本稿では，SFT データ選択を（i）訓練事例の学習価値，（ii）選択後集合の網羅性の二観点から捉える．学習価値は，学習前後のモデルにおける各事例の損失（cross-entropy，CE）の変化により近似する．また，学習価値が高い事例でも互いに類似していれば冗長となり，限られたデータ量での学習効率を損なう可能性がある．

以上を踏まえ，CE 変化に基づくランキングと，埋め込み表現の類似度に基づく冗長性除去を組み合わせた選択手法を検討する．加えて，SFT データの構造に着目し，類似度推定に用いるテキスト範囲や埋め込み構成が結果に与える影響も分析する．

2　関連研究

代表的サブセット　訓練データの全量を用いずとも，高い性能を維持できる代表的サブセットが存在することが報告されている．この系譜の古典的手法として，Moore & Lewis [6] は，in-domain / 汎用言語モデル間のクロスエントロピー差（cross-entropy difference，CED）に基づき文を順位付けし，少量データでより適合した言語モデルが得られることを示した．近年では，Suzuki ら [7] が，学習前後の言語モデルにおける損失変化に基づいて訓練事例を評価し，元データより二〜三桁小さいサブセットでも元データ全体の約 90% に相当する性能が得られることを報告している．同研究では，サブセット構築前に重複除去を行っており，Lee ら [8] の手法が用いられている．一方で，順位付けにより上位に集まりやすい事例の偏りや，選択後集合の網羅性の扱いは，十分に明らかでない点が残る．

微調整におけるデータ選択　fine-tuning 段階のデータ選択は，限られた計算資源の下で性能を引き出すための重要な課題である．Liu ら [9] は既存手法を体系的に整理し，特徴抽出・評価基準設計・選択器評価という観点からデータ選択を整理する枠組みを提示した．

表 1　主要な実験条件と表記（本文中の略記に対応）．（b：抽出率，x：順位区間）

略記	品質選択	冗長性/分布制御（Filter）
Base	–	–
Fullset/Full	全量	なし
RND-b	ランダム抽出	なし
CED-x-b	CED ranking	なし
rCED-x-b	rCED ranking	なし
-Sim	rCED 上位 (Top)	類似度除去
-4096	–	長さフィルタ
-assistρ	–	役割比率フィルタ

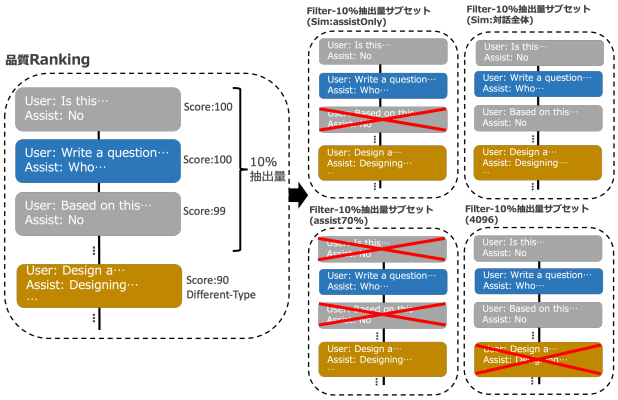

図 1　提案手法の概観

3　手法

本研究では，SFT 訓練データ $\mathbf{D}$ から予算比率 b に応じた部分集合 $\mathbf{S}$ を抽出し，限られた学習予算下でも性能を維持・向上できるかを検討する．基本方針は，(i) 訓練事例ごとの学習価値をスコア化して順位付けし，(ii) 上位に集中しやすい冗長性を抑えて集合としての網羅性を確保することである．

3.1　品質スコア：CED と rCED

各訓練事例 x に対して，base モデルと SFT 後モデルのクロスエントロピーを $\mathrm{CE}_{\mathrm{base}}(x)$，$\mathrm{CE}_{\mathrm{sft}}(x)$ とする．学習価値の指標として，学習によりどれだけ損失が下がったかを

$$\mathrm{CED}(x) = \mathrm{CE}_{\mathrm{base}}(x) - \mathrm{CE}_{\mathrm{sft}}(x) \tag{1}$$

で定義する．$\mathrm{CED}(x)$ が大きいほど，当該事例が SFT により「より説明しやすくなった」と解釈できる．

一方で，$\mathrm{CED}(x)$ は $\mathrm{CE}_{\mathrm{base}}(x)$ の絶対値に依存し得る．例えば，学習前から $\mathrm{CE}_{\mathrm{base}}(x)$ が小さい事例では，差分が小さく見積もられ，学習価値を過小評価する可能性がある．そこで本研究では，差分を初期値で正規化した相対指標として

$$\mathrm{rCED}(x) = \frac{\mathrm{CE}_{\mathrm{base}}(x) - \mathrm{CE}_{\mathrm{sft}}(x)}{\mathrm{CE}_{\mathrm{base}}(x)} \tag{2}$$

を併用する．以降，CED または rCED により事例を降順に並べ，品質ランキングを得る．

3.2　類似度に基づく冗長性除去

品質ランキングの上位には，同形式・同内容の訓練事例が集中しやすい．このような同質化が起こると，限られた予算内で学習できる入力・出力パターンが狭まり，結果として汎化性能が伸びにくくなる可能性がある．そこで本研究では，ランキングで候補を絞った後に，埋め込み表現の類似度に基づいて冗長な事例を除外する．

各事例 x を埋め込み $\mathbf{e}(x)$ に写像し，2 事例の類似度をコサイン類似度

$$\mathrm{sim}(x_i, x_j) = \frac{\mathbf{e}(x_i)^\top \mathbf{e}(x_j)}{\|\mathbf{e}(x_i)\|\,\|\mathbf{e}(x_j)\|} \tag{3}$$

で定義する．ランキング順に事例を走査し，既選択集合 S に対して

$$\max_{x' \in S} \mathrm{sim}(x, x') \geq \tau$$

を満たす事例は冗長として除外する（τ は閾値）．これにより，学習価値を保ちつつ，選択後集合の網羅性を補うことを狙う．

3.3　SFT データ特徴に基づくフィルタ

SFT データは user/assistant からなる対話形式であり，事例の構造に起因してランキングや類似度除去の挙動が変化し得る．本研究では，主手法（品質ランキング＋冗長性除去）を補助する制約として，以下のフィルタを併用する．

テキスト範囲　SFT では通常，対話全体（user+assistant）を入力として与える一方で，損失計算と更新の主な信号は assistant トークン上に乗る（user 側はラベルをマスクする）．このため，assistant 応答に定型句・同型パターンが多い場合，出力側の重複は「ほぼ同一の更新」を繰り返し，少量予算下で学習効率を損ない得る．

以上を踏まえ，類似度計算に用いるテキスト範囲として，対話全体（user+assistant）に加え assistant 出力のみ（assistOnly）も検討する．対話全体は文脈も含めた重複を除去して対話条件の網羅性を確保する一方，user が長い事例では文脈差が支配的となり，同一（または類似）の応答であっても類似度が下がる可能性がある．assistOnly はこの影響を避け，

更新信号が集中する出力側の冗長性を優先的に抑制することを狙う．なお user のみ（userOnly）は入力側に限定され，出力側の冗長性を直接抑えにくいため，本研究では扱わない．

埋め込み構成　複数発話からなる事例に対し，(i) 発話表現の平均により文表現を得る Avg-pooling（avg），(ii) 対話を連結して単一系列として表現する All-in-one（aio）を比較する．強く反映する情報が異なるため，冗長性除去の結果にも影響し得る．

長さ・役割比率　最大系列長を超える事例を除外する長さフィルタ，および user と assistant のトークン比率（assistant 比率）に基づく制約を扱う．これらは品質ランキングや類似度除去とは独立に適用でき，分布の偏りや過度な冗長性を緩和する．

4　実験

本章では，SFT における代表的サブセット抽出を対象に，品質ランキングと冗長性除去，および SFT データ特有の制約が性能に与える影響を検証する．

4.1　実験設定

データセット　OLMo-2-1B の公式 SFT データセット（**D**）`tulu-3-sft-olmo-2-mixture-0225` [10] を用いる．各サンプルは user / assistant からなる対話であり，選択はサンプル単位で行う．予算比率は $b \in \{10, 30, 50\}$（%）とし，$|\mathbf{S}|/|\mathbf{D}| = b$ を満たす部分集合 **S** を構成する．

モデル・学習・評価　ベースモデルは OLMo-2-1B[11] とし，選択した **S** に対して SFT を行う．学習は `open-instruct`[12] の公式実装[1)]に準拠する．評価は OLMo の公式評価基盤 `OLMES` [13] を用い，IFEval [14], BBH [15], DROP [16], MMLU [17], GSM8K [18], MATH , PopQA [19] , TruthfulQA [20] の 8 タスクで評価し，平均スコア（Mean）を報告する．また，Full 比（%）は Fullset-SFT の Mean を 100 とした相対値として算出する．

比較対象　比較対象として，Fullset（$b = 100\%$），Random（予算 b のランダム抽出），Ranking（品質ランキングに基づく順位抽出），Ranking+Filter（Ranking 後に冗長性除去を適用）を用いる．

4.2　品質ランキングによる抽出

品質指標として CED および rCED を用い，各事例のスコアで **D** を降順に並べ，順位に基づき抽出する．CE は標準的な SFT と同様，**対話全体（user+assistant）を入力**しつつ，**assistant トークンのみ**を損失計算の対象として算出した（user はラベルをマスク）．

表 2　rCED-top-10%・4096 設定で、役割比率 ρ における平均スコア（Mean）と Full 比（%）

閾値（ρ）	10%	20%	30%	40%	60%	70%	80%	90%
Mean	27.6	28.0	27.5	27.3	28.1	28.4	28.0	27.0
Full 比	84.64	85.83	84.36	83.90	86.08	87.09	86.00	82.69

順位区間（Top / Middle / Tail）　順位を $r = 1, \ldots, N$（$r = 1$ が最上位），選択数を $k = \lfloor bN \rfloor$ とする．Top は上位から k 件（$1 \le r \le k$），Tail は下位から k 件（$N - k + 1 \le r \le N$）を取る．Middle は中央順位 $c = \lfloor (N + 1)/2 \rfloor$ を中心に前後ほぼ半分ずつ取り，合計 k 件となる連続区間を採用する．

4.3　冗長性除去と分布制御の設計

Ranking で得た候補を先頭から走査し，既選択集合との最大コサイン類似度が閾値 τ 以上となる事例を除外する（greedy filtering）．基本設定は $\tau = 0.90$ とする．埋め込みは（i）対話全体（user+assistant）または（ii）assistant のみ (assistOnly) を対象とし，構成方法は Avg-pooling（avg） と All-in-one（aio）を比較する．

加えて，SFT 事例の単純統計が選択後集合の性質に与える影響を見るため，以下の追加制約を検証する：(i) 最大長フィルタ：最大系列長を超える事例を除外．(ii) 役割比率フィルタ：各事例の user / assistant トークン数を $T_{\text{user}}, T_{\text{asst}}$ とし，assistant トークン比率

$$\rho = \frac{T_{\text{asst}}}{T_{\text{user}} + T_{\text{asst}}}$$

が所定の条件を満たす事例のみを採用する．

5　結果と考察

本章では，第 4 章の設定に基づき，(i) 品質スコアに基づくランキング選択の挙動，(ii) 類似度に基づく冗長性除去および追加制約（長さ・役割比率）の効果，を整理する．性能指標としては 8 タスク平均（Mean）を主に扱う．全タスクの評価結果は付録 A にまとめて示す．

5.1　ランキング選択の分析

まず CED と rCED を比較すると，いずれのサブセット比率においても，rCED に基づく選択は CED

1)　https://github.com/allenai/open-instruct

表 3　10%抽出量における Similarity filtering の設定比較

Group	Embedding	assist70%	4096	Mean	Full 比 (%)
Baselines					
RND-10%	–	–	–	28.6	87.75
rCED-top-10%	–	–	–	28.0	85.94
Ablations（rCED-top-10% + Similarity filtering, $\tau = 0.9$**）**					
aio	対話全体	–	–	28.4	87.03
aio	対話全体	✓	–	29.0	88.96
aio	assistOnly	–	–	27.7	84.93
aio	assistOnly	✓	–	29.1	89.41
aio	assistOnly	✓	✓	28.9	88.65
avg	assistOnly	✓	✓	**29.5**	**90.47**
avg	assistOnly	✓	–	28.8	88.27
avg	assistOnly	–	–	28.2	86.46
avg	対話全体	✓	–	28.7	88.15
avg	対話全体	–	–	28.0	85.86

に比べて概ね高いスコアを与える傾向が観測された（付録 A 参照）．この結果は，第 3 章で述べた動機づけ—CED は $\mathrm{CE}_{\mathrm{base}}(x)$ の大小により差分が過小評価され得るのに対し，rCED は相対化によりその影響を緩和する—と整合的である．

次に順位区間を見ると，データ量に依らず Top は Tail より高いスコアを示し，順位付け自体が「より有効な事例を上位に集める」性質を持つことが示唆される．しかし同時に，Top に基づく学習結果は Random を明確には上回らず，さらに Middle が一貫して最も高いスコアを与える傾向が確認された（付録 A 参照）．この現象は，ランキングが捉えている“品質”の軸が限定的であり，実際に Top 側の事例を目視すると，assistant 出力が極端に短い定型応答など，構造の近い事例が偏って含まれる傾向が見られた．このような同質化は，限られた予算下での有効な更新量を減らし，結果として性能を押し上げにくくする要因になり得る．したがって，Top が Random や Middle に及ばないのはランキングの有効性が否定されるというよりも，ランキング単体では選択後集合の網羅性を十分に担保できないことの表れである．

5.2　冗長性除去と追加制約の併用効果

前節の観察を踏まえ，本節では Top 領域に生じやすい冗長性を抑制できるかを検証する．表 3 に示す 10% 条件では，類似度除去を併用することで Top サブセットの性能が概ね改善し，複数の条件で Random と同等水準に近づく（あるいは上回る）結果が得られた．これは，ランキングが集めやすい「似た形式・近い内容の事例」を間引くことで，限られたデータ量でもより多様な入出力パターンが残り，有効更新が増えた可能性を示唆する．

一方で，埋め込みの構成（avg / aio）や類似度計算に用いるテキスト範囲（対話全体 / assistant-only）によって結果は変動した．すなわち冗長性除去の効果は単なる後処理ではなく，「何を同一視して除外するか」という表現設計に依存する．本実験範囲では，avg による埋め込みに加え，長さ制約と比率制約を併用した条件が最良となり，10% のデータ量で Fullset の約 90% 水準に到達し，Random 抽出（10%）も上回った（表 3）．

役割比率フィルタ　役割比率フィルタ単体の挙動を表 2 に示す．assistant トークン比率の閾値により性能は変動し，本実験では assist70% が最良であった．また全体傾向として，assistant 側の比率が相対的に高い条件（assist60–assist80%）は，user 側の比率が高い条件（assist10–assist40%）より高いスコアを与えることが多い．SFT では学習信号の大半が assistant 側にあるため，出力が十分な情報量を持つ事例を優先することが性能に寄与した可能性がある．ただし，assistant 比率は単なる長さ制約ではなく，「説明が多い応答／短い定型応答」「ユーザ文脈の長さ」など，訓練事例の構造そのものに介入する．そのため，単一の比率帯が常に最良になるとは限らず，網羅性と偏りのトレードオフとして理解するのが自然である．実際，表 3 では，比率制約を加えた上で類似度除去を行うことで，役割比率フィルタ単体の最良値を上回っており，分布の調整（比率・長さ）と冗長性の抑制（類似度除去）は補完的に機能する可能性が高い．

6　おわりに

本研究では，教師あり微調整（SFT）における代表的サブセット抽出に着目し，学習前後の損失変化に基づく品質ランキングと，埋め込み類似度に基づく冗長性除去を組み合わせたデータ選択を検討した．実験より，品質ランキングは有効である一方，上位偏重は同質化を招き網羅性を損ない得ることを確認した．さらに，冗長性除去と長さ・役割比率などの制約を併用することで，限られた予算下でも性能の安定化に寄与し得ることを示した．本研究の知見が，SFT におけるデータ選択を品質と網羅性の両面から設計する際の一助となることを期待する．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011-35（fundamental research），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」の支援を受けたものです．また，本研究成果の一部は，九州大学情報基盤研究開発センター研究用計算機システムの「一般利用」，ならびに，「ABCI 3.0 開発加速利用」の支援を受けて産総研及び AIST Solutions が提供する ABCI 3.0 を利用して得られたものです．

参考文献

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025.

[2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[4] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025.

[5] Alex Havrilla and Maia Iyer. Understanding the effect of noise in llm training data with algorithmic chains of thought, 2024.

[6] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In Jan Hajič, Sandra Carberry, Stephen Clark, and Joakim Nivre, editors, **Proceedings of the ACL 2010 Conference Short Papers**, pp. 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[7] Jun Suzuki, Heiga Zen, and Hideto Kazawa. Extracting representative subset from extensive text data for training pre-trained language models. **Information Processing & Management**, Vol. 60, No. 3, p. 103249, 2023.

[8] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[9] Ziche Liu, Rui Ke, Yajiao Liu, Feng Jiang, and Haizhou Li. Take the essence and discard the dross: A rethinking on data selection for fine-tuning large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 6595–6611, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[10] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.

[11] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025.

[12] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. How far can camels go? exploring the state of instruction tuning on open resources, 2023.

[13] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations, 2025.

[14] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023.

[15] Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. Big-bench extra hard, 2025.

[16] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, 2019.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[18] 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, 2021.

[19] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023.

[20] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

A　実験結果（詳細スコア）

本付録には，本文で扱った全ての設定について，8 タスクの個別スコアと平均スコア（Mean）をまとめる．

表 4　全実験結果：8 タスク個別スコアと平均スコア

Setting	IFEval	BBH	DROP	MMLU	GSM8K	MATH	PopQA	TruthQA	Mean	Full 比 (%)
Reference										
Base (OLMo-2-1B)	20.0	30.8	36.2	29.6	12.7	6.7	12.3	36.8	23.1	–
Fullset-SFT	48.8	29.7	34.4	37.7	44.7	10.5	14.5	40.4	32.6	100.00
10%Ranking										
RND	32.5	30.2	34.9	30.1	39.7	7.1	15.0	39.3	28.6	87.75
CED-top	25.9	29.5	34.2	38.6	31.4	5.8	13.3	40.4	27.4	84.01
CED-mid	34.4	30.8	33.5	29.8	39.7	6.8	15.0	38.9	28.6	87.77
CED-tail	21.8	30.8	35.5	35.8	26.0	9.0	14.4	36.4	26.2	80.41
rCED-top	24.6	30.9	35.6	39.7	37.7	5.0	13.4	37.3	28.0	85.94
rCED-mid	40.5	30.7	33.5	27.8	39.0	7.8	16.0	41.2	29.6	90.66
rCED-tail	25.1	30.7	34.4	30.0	39.9	6.0	15.0	37.3	27.3	83.73
10% Ranking(rCED)+Filter										
Sim-aio-対話全体	26.1	31.5	34.9	40.4	39.3	4.9	13.9	35.8	28.4	87.03
Sim-aio-assistOnly	27.2	30.8	35.8	36.5	35.1	5.2	13.9	36.9	27.7	84.93
Sim-aio-対話全体-assist70%	30.5	31.6	34.2	21.7	50.0	6.7	15.0	42.4	29.0	88.96
Sim-aio-assistOnly-assist70%	29.2	31.2	34.5	23.5	50.2	6.6	14.4	43.6	29.1	89.41
Sim-aio-対話全体-4096-assist70%	29.0	30.6	33.5	22.3	50.7	7.2	14.3	43.5	29.0	88.96
Sim-avg-assistOnly-4096-assist70%	30.3	31.4	34.4	23.5	51.5	6.7	14.4	43.6	29.5	90.47
Sim-avg-対話全体-assist70%	29.6	30.6	34.2	20.6	51.3	7.3	14.9	41.4	28.7	88.15
Sim-avg-assistOnly-assist70%	30.5	30.8	33.6	21.9	49.1	6.5	14.7	43.1	28.8	88.27
Sim-avg-対話全体	24.4	30.4	35.9	38.8	36.3	5.5	13.7	38.9	28.0	85.86
Sim-avg-assistOnly	27.9	31.1	36.1	36.1	37.6	5.8	13.8	37.0	28.2	86.46
30%Ranking										
RND	41.1	30.0	34.8	34.4	42.3	8.3	14.3	39.3	30.6	93.79
CED-top	43.4	29.1	34.1	37.6	36.1	5.7	14.2	38.4	29.8	91.51
CED-mid	40.5	30.1	34.4	31.2	40.7	8.2	14.4	40.3	30.0	91.97
CED-tail	26.6	31.4	34.2	31.9	42.4	9.7	14.5	38.6	28.7	87.94
rCED-top	35.1	29.9	34.6	34.8	38.7	7.9	13.3	39.0	29.2	89.50
rCED-mid	46.0	30.1	33.4	32.0	41.1	8.8	15.2	39.6	30.8	94.42
rCED-tail	34.9	30.4	33.5	28.2	40.9	7.6	14.3	39.7	28.7	88.01
50%Ranking										
RND	45.3	30.3	34.5	36.4	44.4	9.4	14.6	39.6	31.8	97.60
CED-top	47.3	29.6	34.4	36.1	39.9	6.2	14.1	39.2	30.9	94.65
CED-mid	45.3	30.1	34.2	33.5	42.2	8.2	14.3	41.4	31.1	95.51
CED-tail	33.8	30.8	33.8	34.6	44.5	9.3	14.2	38.6	29.9	91.88
rCED-top	44.5	30.1	34.4	34.9	42.0	8.9	13.5	40.0	31.0	95.20
rCED-mid	47.9	29.2	33.9	33.9	42.3	9.5	15.1	40.6	31.5	96.76
rCED-tail	40.9	30.4	33.5	33.6	42.2	7.9	14.6	40.8	30.5	93.51