異種SLMアンサンブルにおける多数決不能・回答欠損時の回答共有再推論

河野 尊[1,a)]　神野 敦哉[2,b)]　田口 開智[2,c)]　那須 渚[2,d)]　小比賀 亮仁[1,e)]　小板 隆浩[2,f)]

概要：異種小規模言語モデル 3 種のアンサンブルで，多数決不能または回答欠損が生じた設問に再推論を適用した．自己回答のみを使う方式 B1，生成元名を伏せて回答を共有する方式 A1，基準方式 B0 を GSM8K と ARC-Challenge 各 300 問で比較した．A1 と B0 の正答数差は GSM8K で +9 問，ARC-Challenge で −3 問だったが，多重比較補正後に有意差は確認されなかった．

1. はじめに

LLM の推論では，回答性能だけでなく GPU などの計算資源や利用コストも課題となる．限られた資源のもとで回答性能を高める方法の一つとして，複数の小規模言語モデル（SLM）の回答を組み合わせるアンサンブルが検討されている．ただし，3 モデルの回答がすべて異なると多数決できず，回答欠損時も集約情報が減る．Self-Consistency は同一モデルの複数推論経路を多数決し [1]，Multiagent Debate と ReConcile は複数エージェントの討論・合議で回答を決める [2][3]．一方，外部正解を使わない自己修正で性能が下がる例も報告されている [4]．そこで本研究は，多数決不能または初回回答欠損時に，基準方式 B0，自己回答のみを使う再推論方式 B1，匿名化した 3 モデルの回答を共有する再推論方式 A1 を比較し，初回集約を補えるか評価する．

2. 提案方式

Qwen3-8B（M1）[5]，Llama-3.1-8B-Instruct（M2）[6]，Mistral-7B-Instruct-v0.3（M3）[7] を用いた．未抽出回答は未回答（`null`）とし，M1 単体条件では同じ初回ログから M1 の回答を採用した．B0 は，有効回答が 2 件以上なら最多回答を採用し，同数時は較正用データの正答率順に選ぶ．有効回答が 1 件以下なら未回答とする．

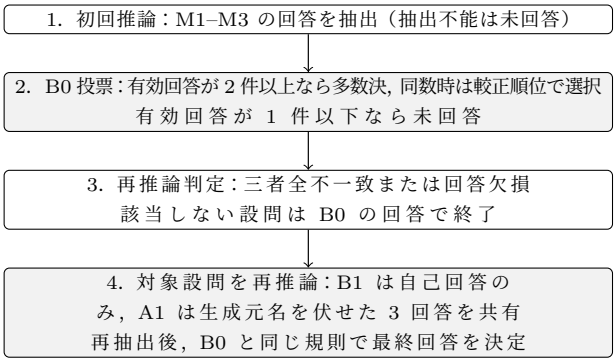


図 1　B0 の初回投票と B1・A1 の再推論手順

Fig. 1　Initial voting and re-inference procedure for B0, B1, and A1.

再推論条件は，初回回答が 3 件とも異なる場合または回答欠損がある場合とした．欠損時は残る 2 回答が一致していても対象に含め，欠損設問全体で再推論の影響を評価する．B1 には設問と各モデル自身の初回回答を提示し，A1 には設問と 3 回答を「候補 1」〜「候補 3」として M1〜M3 の順に提示した．生成元名を伏せてモデル名による推論バイアスを避け，未回答は利用不可と示した．初回の説明文や推論過程は共有せず，各モデルに 1 回再推論させた後，B0 と同じ規則で最終回答を決めた．提案方式の処理の流れを**図 1** に示す．

3. 評価方法

3 モデルを Q4_K_M 形式（4 bit 量子化）にし，NVIDIA T4 GPU 搭載の AWS EC2（`g4dn.xlarge`）上で実行した．GSM8K（`openai/gsm8k`）main/test[8] と ARC-Challenge（`allenai/ai2_arc`）test（以下 ARC-C）[9] を seed 42 でシャッフルし，較正用 100 問，同率判定用 50 問，評価用

[1] 同志社大学
Doshisha University
[2] 同志社大学大学院
Graduate School of Doshisha University
[a)] cguj1036@mail4.doshisha.ac.jp
[b)] ctwp0126@mail4.doshisha.ac.jp
[c)] ctwp0168@mail4.doshisha.ac.jp
[d)] ctwn0144@mail4.doshisha.ac.jp
[e)] akohiga@mail.doshisha.ac.jp
[f)] tkoita@mail.doshisha.ac.jp

© 1959 Information Processing Society of Japan

表 1　B0 からの正誤遷移と正答数差

Table 1　Correctness transitions from B0 and differences in correct counts.

データ	方式	正 → 正	正 → 誤	誤 → 正	誤 → 誤	B0 との差
GSM8K	B1	31	4	7	11	+3
	A1	33	2	11	7	+9
ARC-C	B1	20	7	1	3	−6
	A1	24	2	0	4	−3 ※

対象は GSM8K 53 問，ARC-C 31 問．正誤遷移は回答確定問で集計し，B0 との差は未回答を不正答とした正答数差．※ ARC-C/A1 は未回答 1 件を遷移分類から除外（B0 は正答）．

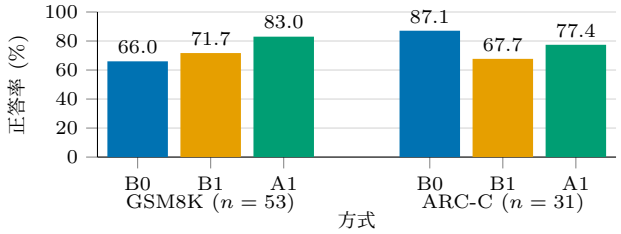


図 2　再推論対象の最終正答率

Fig. 2　Final accuracy on questions selected for re-inference.

300 問を重複なく抽出した．較正用データで正答率が最高のモデルを選び，同率時は同率判定用を加えた 150 問で再評価した．回答欠損時の規則は推論後に定め，生ログから再集計した．

生成 API は `temperature=0`，`max_tokens=8192` とし，Qwen3 の thinking を無効にした．`Final Answer:` 以降から回答を抽出し，GSM8K では `\boxed{}` 内の数値を優先して桁区切りを除去した．ARC-C では枠内の選択肢ラベルを大文字小文字を区別せず抽出した．生成上限に達しても抽出できれば採用し，抽出できない回答は `null` とした．未回答は再推論対象に含め，正答率の分母に含めたが正答には数えなかった．

各方式の正誤を同一設問ごとに比較し，正誤の不一致数に基づく正確 McNemar 検定を用いた．B0・B1・A1 の全ペア比較を 2 データセットで行った計 6 件には，多重比較による偶然の有意差を抑えるため，Holm 法で p 値を補正した（未回答は不正答とした）．

4.　実験結果

較正用 100 問での M1，M2，M3 の正答率は GSM8K で 90/79/15%，ARC-C で 91/87/65%であり，いずれも M1 を回答選択モデルとした．初回の有効回答が 3 件とも異なる設問は GSM8K で 45 問，ARC-C で 20 問，初回欠損は 8 問，11 問であり，再推論対象は計 53 問，31 問となった．**表 1** に B0 からの正誤遷移，**図 2** に再推論対象の最終正答率を示す．

5.　考察

A1 と B0 の正答数差は GSM8K で +9 問（誤答から正答 11 問，逆方向 2 問），ARC-C で −3 問（誤答から正答 0 問，逆方向 2 問，未回答 1 件）だった．全 300 問の正答率は GSM8K で B0/B1/A1 が 92.0/93.0/95.0%，ARC-C で 90.7/88.7/89.7%；M1 単体は 91.7/93.7%だった．再推論対象では A1 が B1 を GSM8K で 6 問，ARC-C で 3 問上回った．

多重比較補正後，いずれの比較でも有意差は確認されなかった（最小の補正後 p 値は 0.135）．再推論対象 1 問は正答率の 1.9 ポイント（GSM8K），3.2 ポイント（ARC-C）に当たる．回答欠損時の投票規則は結果を見た後に定めたため，再推論の効果を確定できない．ARC-C では多数決不能 20 問中 18 問を B0 が正答しており，改善余地が小さかった可能性がある．

6.　おわりに

多数決不能または回答欠損時に 3 モデルの回答を共有する再推論 A1 を評価したが，B0 との有意差は確認できなかった．今後は問題数とモデル構成を増やし，回答順序の影響も検証する．

参考文献

[1] X. Wang et al., ”Self-Consistency Improves Chain of Thought Reasoning in Language Models,” ICLR, 2023.
[2] Y. Du et al., ”Improving Factuality and Reasoning in Language Models through Multiagent Debate,” ICML, PMLR, vol. 235, pp. 11733–11763, 2024.
[3] J. C.-Y. Chen, S. Saha, and M. Bansal, ”ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs,” ACL, pp. 7066–7085, 2024.
[4] J. Huang et al., ”Large Language Models Cannot Self-Correct Reasoning Yet,” ICLR, 2024.
[5] A. Yang et al., ”Qwen3 Technical Report,” arXiv:2505.09388, 2025.
[6] Meta AI, ”Llama 3.1 Model Card,” 2024.
[7] Mistral AI, ”Mistral-7B-Instruct-v0.3 Model Card,” 2024.
[8] K. Cobbe et al., ”Training Verifiers to Solve Math Word Problems,” NeurIPS, 2021.
[9] P. Clark et al., ”Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge,” arXiv:1803.05457, 2018.

© 1959 Information Processing Society of Japan