Single-SLM における 推論フレームワークの分析

関たいよう [1]，高野敏明 [1]
[1] 静岡理工科大学 情報学部
2218070.st@sist.ac.jp, takano.toshiaki@ieee.org

概要

近年，計算資源の制約や運用コストの観点から，ローカル環境で動作する小規模言語モデル（SLM）への注目が増している．しかし，SLM はパラメータ数の制約により推論能力が低く，複雑なタスクでハルシネーションを起こしやすい．大規模モデルではマルチエージェントの議論が有効とされるが，単一モデルの SLM（Single-SLM）での効果は自明でない．本研究では，Single-SLM の推論フレームワークが，推論の信頼性と多様性に与える影響を検証する．具体的には，2 つの推論タスクを用い，エージェント間で対話する議論型と，タスクを分割する分業型を比較した．さらに，発言のコサイン類似度から思考の多様性に与える影響を調査した．

1 はじめに

昨今，大規模言語モデル（LLM）は自然言語処理タスクにおいて人間と同等以上の性能を示し，単なる情報検索だけでなく，意思決定支援を行う自律型エージェントとしての応用が進んでいる．また，情報セキュリティの観点から，ローカル環境で動作する小規模言語モデル（SLM）も注目されている．しかし，SLM は信頼性と安定性に課題を抱えている．

言語モデルにおける信頼性の課題として，ハルシネーションと呼ばれる事実に基づかない嘘の生成があげられる．これらの問題の解決策として Single-Agent（SA）フレームワークの Chain-of-Thought（CoT）[1] や自己反省（Self-Reflection）[2] などの手法が提案されてきた．しかし，これらの提案手法ではモデルが自身の誤った初期回答に固執し，修正が機能しなくなる思考の退化（Degeneration-of-Thought: DoT）という問題が指摘されている．一方で，計算資源の限られた環境では GPT-4 などの LLM を動作させるのは難しい．家庭用計算機では，パラメータ数の少ないモデルの実装が現実的である．しかし，パラメータ数の少ない SLM は知識量や推論の安定性に課題がある．近年，これらの問題の解決手法として，Multi-Agent（MA）フレームワークが有効とされる．

本研究の目的は，S-SLM の MA システムの議論型フレームワークが有効に機能するかを明らかにすることである．モデル本来の生成と，エージェント同士を議論させる議論型，タスクを分割して順次処理する分業型を比較する．具体的には，常識推論とマルチホップ推論という性質の異なる 2 つの推論タスクを用いて正答率を比較する．また，議論中の発言の類似度で，推論フレームワークの違いが思考の多様性に与える影響を考察する．

2 関連研究と推論フレームワーク

LLM の推論能力向上とハルシネーション・DoT の抑制に向けて，これまで様々な手法が提案された．Liang ら [3] は，マルチエージェント（MA）に相手への批判を強制し議論させる Multi-Agent Debate (MAD) フレームワークを提案している．MAD フレームワークにより発散的思考を促進し，DoT を抑制できることを実証した．

Chen ら [4] は，GPT-4 や Claude など異なる LLM モデルを使用し，他者の推論を取り入れて回答を修正させる協調的議論，Reconcile を提案している．異なる学習過程を持つ複数のモデルにより，高い多様性を確保したが，高コストであるという課題が残る．これに対し，Wang ら [5] は，単一 LLM 内で複数のペルソナをシミュレートする Solo Performance Prompting（SPP）を提案した．SPP の手法はリソース制約下でも実用的な MA フレームワークが構築可能であることを示した．特に，具体的かつ複数の知識を正確に引き出す必要がある，知識集約型の問題において事実性の向上を示した．しかし，S-SLM

の MA システムは，モデルのバイアスが共通しているだけでなく，低パラメータという制約が残る．そのため，MA 化では DoT を回避できない可能性がある．

本研究では，Sanwal[6] が提唱する，推論を役割ごとの階層に分けて処理するアプローチ “Layered Chain-of-Thought（階層的な思考連鎖）” に着目した．計算資源の限られた S-SLM 環境下において，“議論” と “分業” のどちらの推論フレームワークが有効かを検証する．検証にあたり，性質の異なる 2 つのタスクを用いて，各手法が推論に与える影響を分析する．また，BART によるコサイン類似度を用いた思考の多様性を分析し，S-SLM に適したフレームワークを明らかにする．

3　タスク実験

本章では，S-SLM 環境下における推論フレームワークの違いを，特性の異なる 2 つの推論タスクで検証する．

3.1　データ設定

解の導き方による違いを分析するため，jcommonsenseqa-v1.3[7] と JEMHopQA[8]，2 つの日本語データセットを採用した．

JCommonsenseQA（JCQA）　常識推論能力を問う選択式問題．学習済みのデータから知識を集約し，答えを導く形式．本実験ではランダムに抽出した 100 問× 30 セットを行い，その平均正答率を評価した．

JEMHopQA（JEQA）　マルチホップ推論を必要とする問題．入力に与えた正解の根拠となるデータ（derivation）情報合成を行わせる形式と，derivation を入力に与えずに学習済み知識を集約し回答する形式．2 つの形式でそれぞれ 120 問× 30 セットを行い，その平均正答率を評価した．

3.2　実験条件

モデルには `ELYZA-llama-3-8b` を使用．計算資源の制約を考慮し，モデルを 4bit 量子化して推論を行った．生成パラメータは多様性を確保するため，`temperature` = 1.0，`top-`p = 0.9 とした．本研究では，推論プロセスにおけるエージェント間の議論と，エージェント間の分業に着目し，図 1 の 4 つの推論フレームワークを定義した．全てのエージェントが 1 度発言することを 1 ラウンドとする．

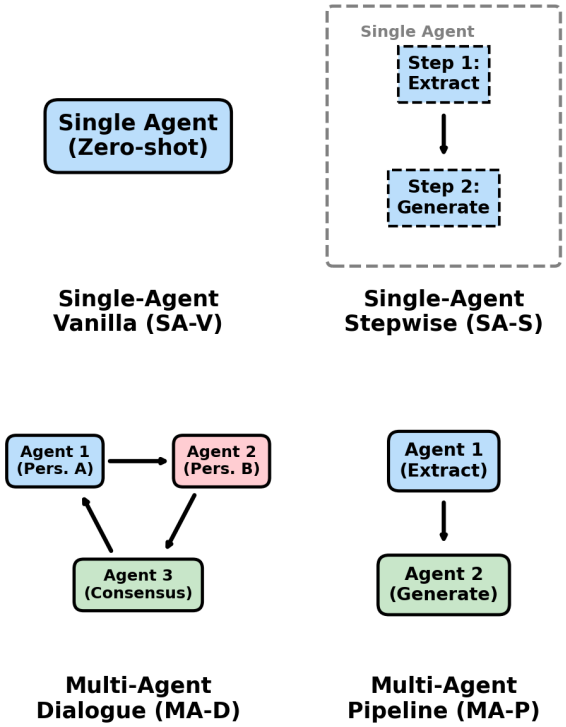

図 1: 4 つの推論フレームワーク

1. **Single Agent Vanilla (SA-V)**：ペルソナを指定せず，Zero-shot で回答させるベースライン（1 左上）．
2. **Single Agent Stepwise (SA-S)**：MA における議論プロセスをエージェント内の思考ステップとして模倣し，順次的に処理を実行して最終的な回答を導出する（1 右上）．
3. **Multi-Agent Dialogue (MA-D)**：3 体のエージェント（異なるペルソナ）が 1〜3 ラウンドの議論を行い，最終的な回答を導出する（1 左下）．
4. **Multi-Agent Pipeline (MA-P)**：議論を行わず，タスクを情報抽出, 回答生成などの明確なステージに分割し，各ステージを専任のエージェントが担当して処理する（1 右下）．

3.3　結果と考察

各推論フレームワークにおける 30 回試行の平均正答率（Accuracy）を表 1 から表 3 に示す．

結果，MA-D は SA-V の平均正答率を上回ることができず，表 1 および表 2 において，MA-D はラウンドを重ねるごとに平均正答率が低下した．これは，S-SLM 構成では全エージェントが同一の知識ベースとバイアスを共有しているため，議論を行っても多様な視点が生まれなかったことが要因と考

表 1: JCQA 各手法の平均正答率

Architecture	Accuracy (%)
SA-V	81.67
SA-S	76.23
MA-D (Round 1)	82.73
MA-D (Round 3)	81.30
MA-P	**87.23**

表 2: JEQA 各手法の平均正答率 derivations あり（情報合成）

Architecture	Accuracy (%)
SA-V	**88.67**
MA-D (Round 1)	67.69
MA-D (Round 2)	61.08
MA-D (Round 3)	58.44

表 3: JEQA 各手法の平均正答率 derivations なし（知識集約）

Architecture	Accuracy (%)
SA-V	29.35
SA-S	20.08
MA-D (Round 1)	23.78
MA-D (Round 3)	25.03
MA-P	**36.28**

えられる．また，表 2 において MA-D が SA-V の平均正答率を大きく下回ったのは，事実性が重視されるタスクにおいて，ペルソナによる批判や検証がかえって事実を歪める事象が確認された．

表 1 と表 3 において，SA-S の平均正答率は JCQA で 76.23%,JEQA で 20.08%まで低下し，SA-V を大きく下回った．小規模モデルが 1 つのエージェント内で思考プロセスを展開すると，自身の生成した不正確な中間推論に注意が向き，修正ができなくなってしまうことが確認された．本実験では両タスクを通じて，MA-P が一貫して最も高い平均正答率を記録した．JCQA では平均 87.23%，難易度の高い JEQA では平均 36.28%を達成し，ベースラインである SA-V を t 検定により有意に上回った（$p < 0.0001$）ことを確認した. SA-S と MA-P は手順を分けて考えるという点で共通しているが，その実装形態において異なる．SA-S が全ての思考を一つのコンテキストに詰め込んだのに対し，MA-P はプロンプトを物理的に分割し，前段の出力結果のみを次段に渡す．これにより，推論過程で生じたハルシネーションがコンテキストから遮断され，後段の推論はクリーンな情報のみに基づいて推論したと考えられる．

4 議論構造と発言多様性の分析

前章の実験で，S-SLM では議論の効果が推論に対し有効でないことが示唆された．本章では，この要因を深掘りするため，推論フレームワークの違いがエージェント間の思考の多様性を示す尺度として，BART によるコサイン類似度を用いて分析する．

4.1 データ設定と比較手法

本実験では，匿名表現の規制・生成 AI の教育利用・高等教育の費用負担の価値観により意見が分かれる 3 つの議題を設定した．また，これらの議題に対して多様な観点からの発言を促すため，7 つの議論用ペルソナと調整役を定義した．各エージェントには，ペルソナに基づき，日本語で 300 文字程度の意見を述べるよう指示を与えた．S-SLM における情報の流れが思考の多様性に与える影響を検証するため，3 つの推論フレームワークを定義した．

1. **Simple Relay (Relay)**：過去の議論履歴（全エージェントの発言）をそのまま入力として受け取り，次の発言を生成する単純なリレー形式．
2. **Moderated Dialogue (Dialogue)**：調整役が介在する形式．入力には直前のエージェントの発言に加え，調整役が生成したこれまでの意見の要約と中立的な問いかけが含まれる．
3. **Summary Panel (Panel)**：個別の発言履歴を受け渡さず，調整役が生成した意見の要約と問いかけのみを入力とする形式．

各フレームワークで，3 つの議題，2 つの議論用ペルソナと調整役を実装し，全 126 パターンを実施した．

4.2 評価指標

実験では思考の多様性を示す尺度として，日本語文章をベクトルに変換できる SentenceTransformer のモデル `paraphrase-multilingual-MiniLM-L12-v2` を使ったコサイン類似度を用いた．そこで，以下の 3 つの指標を定義した．

- **直前自己類似度**: エージェントのラウンド N の発言と同じエージェントの直前の発言（ラウンド $N-1$）の類似度．これが高い場合，議論が

進展していないことを示す．

- **自己ループ度**: 現在の発言と，自身の過去の全発言履歴中の最大類似度．これが高い場合，エージェントは新しい情報を生成できていないとみなせる．
- **ペルソナ同調度**: 他エージェントの直近の発言との類似度．これが高い場合，コンテキストの汚染が起こっていると見なせる．

4.3 結果と考察

図 2 に各ワークフレームのコサイン類似度が閾値（0.9）を超えたラウンドの出現割合を，図 3 に各ワークフレームにおける 3 つの指標の分布を示す．

図 2 から，Relay はすべての指標において他の 2 つのフレームワークに比べて発言の多様性が乏しいことが分かる．さらに，類似度 0.9 以上の自己ループ度は全体の 41.9%に達した．これは，発言の約 4 割が過去の繰り返しであることを意味する．S-SLM 環境下では，ペルソナを変えても推論回路が同一である．そのため議論が長引くとコンテキスト内の自身の発言に注目し，新しい情報を生成できなくなる．結果，DoT が発生していることが確認された．この結果が，前章の実験でも同様に発生し MA-D の平均正答率が向上しなかったと推察される．

図 3 から，直接的な発言履歴の受け渡しを行っていない Panel が最もペルソナ同調度の分布が低い値に偏っている．また，図 2 から，Dialogue および Panel は自己類似度や自己ループ度の高類似度出現ラウンドは 10%以下に留まった．このことからもコンテキストの分離は S-SLM において発言の多様性を確保し DoT を抑制する効果があると考えられる．したがって，コンテキストを意図的に遮断する分業型の推論フレームワークが，S-SLM の特性に適していると考えられる．

5 おわりに

本研究では，計算資源に制約のある S-SLM 環境下において有効な推論フレームワークを明らかにするため，特性の異なる 2 つのタスクを用いた評価実験を行った．加えて，BERT を用いたコサイン類似度により，議論構造と発言の多様性を定量的に分析した． 実験の結果，S-SLM 同士の “議論” は DoT やノイズの増幅を招き，SA 以下の性能に留まることが確認された．対照的に，タスクを機能単位で分割し，独立したプロンプトで順次処理を行う “分業型” フレームワークは，常識推論およびマルチホップ推論の双方において，ベースラインを有意に上回る平均正答率を達成した．

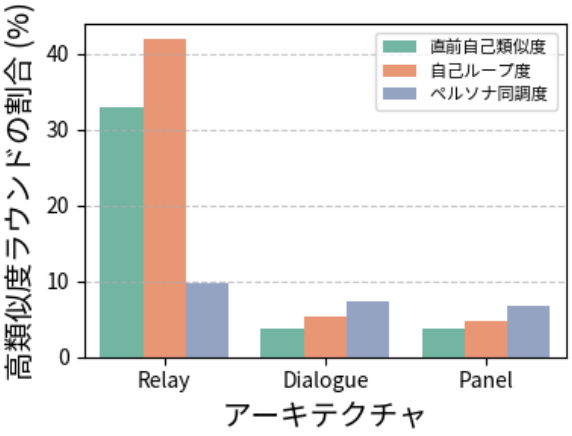

図 2: 高類似度出現割合

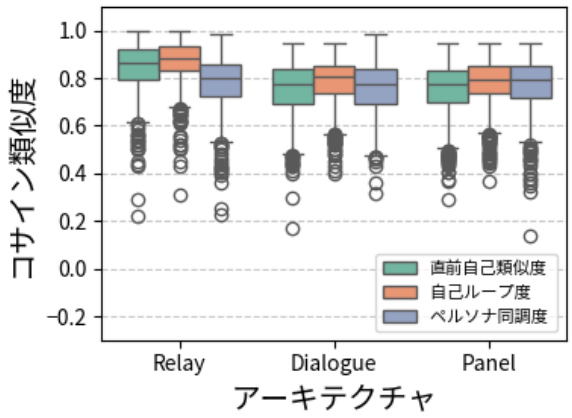

図 3: 各推論フレームワークの類似度

結論として，S-SLM の推論能力を最大化する鍵は，コンテキスト分離によるノイズの遮断にある．S-SLM の MA システムにおいて，不用意な履歴共有を伴う議論は DoT を誘発し，推論品質を低下させるリスクが高い．したがって，コンテキストを物理的に分断するフレームワークこそが，S-SLM の潜在能力を引き出し，推論の質を向上させるアプローチであることを示唆した．

参考文献

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837, 2022.

[2] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In **Advances in Neural Information Processing Systems**, Vol. 36, 2023.

[3] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 17889–17904, 2024.

[4] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7066–7085, 2024.

[5] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 257–279, 2024.

[6] Manish Sanwal. Layered chain-of-thought prompting for multi-agent llm systems: A comprehensive approach to explainable large language models, 2025. https://arxiv.org/abs/2501.18645.

[7] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[8] Aoi Ishii, Masato Mita, Yuiki Minahara, Yugo Murawaki, and Daisuke Kawahara. JEMHopQA: A Japanese explanation-generating multi-hop question answering dataset. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 10032–10045, Singapore, December 2023. Association for Computational Linguistics.