単一エージェントとマルチエージェントの生成多様性の評価

CUI ENCHENG[1] PENG SHAOWEN[1] 伊藤 和浩 [1] XU JINSHA[1] 久田 祥平 [1] 若宮 翔子 [1] 荒牧 英治 [1]
[1] 奈良先端科学技術大学院大学
{cui.encheng.cg3,peng.shaowen,ito.kazuhiro.ih4,xu.jinsha.xg8}@naist.ac.jp
{s-hisada,wakamiya,aramaki}@is.naist.jp

概要

マルチエージェントシステム（MAS）は，異なる役割を持つエージェント間の相互作用により推論の多様性向上が期待される一方，多くの研究ではマルチエージェント構造の効果とプロンプト条件付け（PC）の効果が切り分けられていない．本研究は，MAS と単一エージェント（SA）に同一 PC を適用した統制実験により，両者の寄与を比較した．その結果，意味的多様性は SA が一貫して MAS を上回った．並列型 MAS では，同一ラウンド内で他者の出力を参照できないため，アイデアが重複しやすく，一方，SA はにマルチ出力戦略により，文脈共有と多様性確保において効率的であることが示された．本結果は，より有効で簡潔な多様性向上策を示し，今後の研究設計に有用な示唆を与える．

1 はじめに

大規模言語モデル（LLMs）に基づくエージェントベースのフレームワークは，近年，大きな注目を集めている．こうしたシステムにおいて，生成の多様性は，推論品質の向上，カバレッジの拡大，および頑健性の向上をもたらす中核的なメカニズムとして広く位置付けられている [1, 2, 3, 4]．この多様性を確保するための主要なパラダイムが，マルチエージェントシステム（MAS）である．MAS は，異なる役割を持つマルチエージェント (MA) を構成し，それらの相互作用をシミュレートすることで，多様な推論を生成する．このアプローチの有効性は多くの研究により支持されているものの [5, 4, 6]，その性能向上の根本的な要因は十分に検討されていない．

多くの MAS 設定では，各エージェントが同一の LLM バックボーンを共有している．そのため，MAS には通常，各エージェントに異なる役割や視点を割り当てるよう設計されたプロンプト条件付け（PC）が付随する [7, 5, 8, 9]．言い換えると，MAS は，プロンプトエンジニアリングを通じて，事前に定義された多様な推論事前分布を明示的に付与されている．にもかかわらず，既存の MAS 研究の多くは，提案手法（MAS）と単一エージェント (SA) のベースラインと比較する際，生成多様性の要因である PC を SA に適用していない．このことは，次の重要な問いを生む：**観測された多様性の向上は，MA という構造に起因するのか，それとも精緻な PC によるのか？**

この問いを検証するため，本研究では，SA が MAS と同一の PC 戦略を用いる統制実験を行う．本実験では，出力量を MAS と厳密に整合させるため，SA に単一の推論過程で複数の応答を生成させるマルチ出力戦略を採用する．これらの統制条件下での実験結果において，既存研究の示唆に反して，SA 設定は MAS よりも高い生成多様性を示すことが分かった．さらに，同一の条件下で出力量を整合させる目的で導入したマルチ出力戦略が，多様性を高める不可欠な要因となっていた．本研究の結果は，多様性を達成するために計算コストの高いエージェント間相互作用 [10, 11, 12] は必ずしも必要ではなく，より単純で厳密に設計されたプロンプト戦略が有効となり得ることを示唆している．

2 関連研究

2.1 生成と推論における多様性

生成多様性は，LLM の推論能力，頑健性，および創造性にとって重要である．代表的には，核サンプリングや top-k デコーディングなどの確率的サンプリング [13, 14]，多様な推論過程の集約 [1, 2]，多様な候補解の生成と検証 [15, 16]，およびプロンプト変動に対する頑健性向上を目的とした多様化 [17, 18, 19] が広く用いられている．

2.2 多様性を誘発するための MAS

多様な視点を体系的に誘発するために，近年の研究は有望なパラダイムとして MAS を用いることが多い．MAS を用いた枠組みは，異なる役割とメモリを持つエージェント間の相互作用を統括し，複雑な社会的ダイナミクスや協調的問題解決を模倣することができる [8, 20, 21]．多くみられる応用としてマルチエージェント討論があり，対立する見解や割り当てられたペルソナを持つエージェントが相互に批判を行うことで，発散的思考を誘発する [3, 4, 22]．さらに，通信トポロジーや認知的シナジーといった要因が，結果の多様性をどのように形成するかについて，理論的・実証的研究が行われてきた [6, 23, 9]．

これらの進展にもかかわらず，一連の研究には方法論上の曖昧さが残されている．以上の研究はしばしば性能向上の要因をマルチエージェント構造それ自体とするが，多くの場合，高度に設計された役割特化の PC を備えた MAS と汎用的な SA ベースラインとを比較している．そこで本研究では，PC を統一した統制実験を設計し，構造差がもたらす影響を系統的に検証する．

3 タスク設定

3.1 オープンエンド質問の作成

LLM を評価するための多くのベンチマークは，質問応答データセットや数理問題を含め，単一の正解を導出させるように設計されている．正解が単一のタスクでは，LLM の出力がそもそも多様になりづらいため，出力の多様性についてのエージェント設計間の比較が難しい．したがって，生成能力の多様性を厳密に評価するためには，正解の数に制約のない発散的思考を要求するタスクが必要となる．

本研究では，GPT-5 を用いて，合計 300 件のオープンエンド質問セットを作成した．各質問は，複数の回答を許容するように設計されている．なお，質問の妥当性を担保するために，第一著者がすべての質問を人手で精査した．

3.2 ペルソナではなく視点の採用

PC は，しばしば「弁論士」「エンジニア」「裁判官」といったペルソナの形で記述される．しかし，ペルソナによる多様化は，通常，口調や背景知識，あるいは修辞的スタイルといった役割を通じて出力を形成するため，純粋な意味的多様性を測定する際には，ノイズとなる望ましくない要素を導入しがちである．

これに対し，本研究では，タスク特化の明示的な視点を採用する（図 1）．こうしたタスク特化の視点を採用した PC により，役割などに由来する文体・口調の差ではなく，解法や観点といった推論内容の違いに基づく多様性をより直接に捉えられ，生成結果の意味的多様性を効果的に担保できる．

3.3 多様性指標による評価

多様性の定量化には，n-gram 重複率や Self-BLEU のような語彙レベルの指標 [24, 25] が広く用いられるが，表層的一致性に着目するため，意味的差異を十分に反映しない場合がある．本研究の目標は意味的多様性のため，文埋め込みに基づく意味空間上の差異を捉える指標を採用し [26, 27, 6]，埋め込みベースの多様性尺度として Vendi Score [28] を用いる．

Vendi Score は，参照データを必要とせず，かつ意味的な類似度を考慮した指標であり，出力間のカーネル類似度行列 K（例：埋め込みベクトル間のコサイン類似度）の固有分布に基づいて算出される．具体的には，埋め込みモデルとして `all-mpnet-base-v2`（768 次元）[1] を用い，生成された全応答セットに対する多様性を評価する．

Vendi Score は評価対象のサンプル数に影響されるため，本研究では各質問あたりの視点の数を固定することで，この要因を統制する．各視点は 1 つの生成回答に対応しており，ハイパーパラメータ $k \in \{2, 4, 8, 16\}$ の選択は，視点の数と最終的に得られる出力数の双方を同時に規定する．

4 実験設定

本実験は，Gemini-2.5 Flash-Lite，Qwen3-32B，および GPT-4.1 mini の 3 種の LLM を用いて行った．

LLM の生成関数を G，固定のプロンプトテンプレートを P，k 個の視点集合を $A = \{a_1, \ldots, a_k\}$ とする．

4.1 同一ラウンド内の構造 (Intra-round)

逐次型 MAS (Sequential MAS) 同一ラウンド内で順番に生成し，ステップ i では同ラウンド内の既生成出力 $R_{<i} = \{r_1, \ldots, r_{i-1}\}$ を引数にとる

1） https://huggingface.co/sentence-transformers/all-mpnet-base-v2

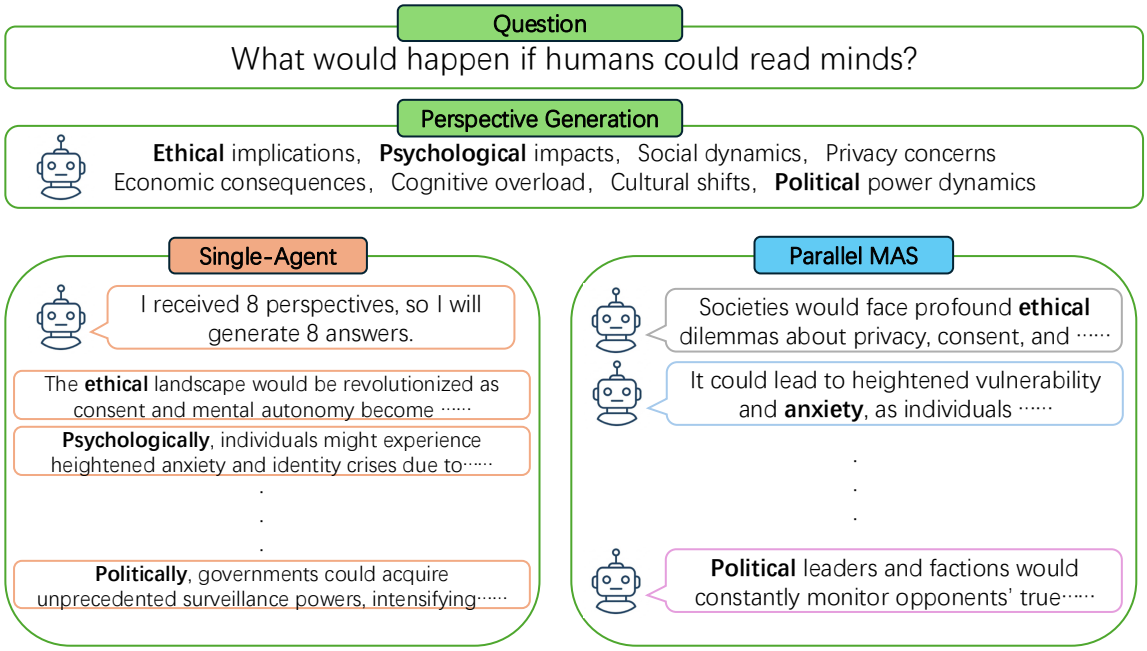

図 1　SA 設定および並列型 MAS 設定における生成パイプラインの概要. 質問が与えられると，システムはまず k 個の異なる視点を生成する（ここでは $k = 8$ を例示する）．**SA** 設定では，単一のモデルが k 個の視点を同時に受け取り，単一の生成内で各視点あたり 1 つの回答を生成するよう求められる．**MAS** 設定では，k 個の視点を k 個の独立エージェントに 1 つずつ分配し，各エージェントはそれぞれ 1 つの回答を生成するよう求められる．

[5, 20, 21]：

$$r_i = G(P, R_{<i}, a_i), \quad i = 1, \ldots, k \tag{1}$$

$$R_{\text{S-MAS}} = \{r_1, \ldots, r_k\} \tag{2}$$

並列型 MAS (Parallel MAS)　同一ラウンド内の相互参照を行わず，各視点は独立に生成する [3, 4, 6]：

$$r_i = G(P, a_i), \quad i = 1, \ldots, k \tag{3}$$

$$R_{\text{P-MAS}} = \{r_1, \ldots, r_k\} \tag{4}$$

提案の SA 対照設定　MAS と同一の視点集合 A を単一の呼び出しに集約し，同一の PC に基づいて k 個の回答を生成する：

$$R_{\text{SA}} = G(P, A) = \{r_1, \ldots, r_k\} \tag{5}$$

これにより，視点集合 A を共有したまま（図 1），MAS と PC の条件を統制した公平な比較が可能となる．

4.2　ラウンド間の更新 (Inter-round)

各ラウンド t は，それまでに得られた全出力を参照して生成を行う．

$$P^{(t)} = P^{(t-1)} \cup R^{(t-1)} \tag{6}$$

$$R^{(0)} = \emptyset, \quad P^{(0)} = P \tag{7}$$

この設計は，ラウンド間における因果的な完全グラフ (all-to-all) 通信トポロジーに対応する．すなわち，前回までのラウンドで任意のエージェントが生成した出力は，後続ラウンドの生成の際に，常に参照可能である．また，特定の履歴のみを参照する，あるいは要約のみを共有するといった特殊な相互作用トポロジーは，この「全履歴可視」の構造を部分的に削減した変種として位置付けられる．

5　実験結果

結果を図 2 に示す．**SA はすべてのモデル, すべての k, すべてのラウンドで MAS を一貫して上回る.**

5.1　SA が 並列型 MAS を上回る要因分析

一部の生成結果を確認したところ，並列型 MAS では，各エージェントに異なる視点を割り当てているにもかかわらず，ほぼ同一の回答が生成されるという傾向が観察された．この現象は，生成結果が定型表現に偏ったり，内容の反復や同質化が生じたりする degeneration として，既存研究でも指摘されている [13, 24].

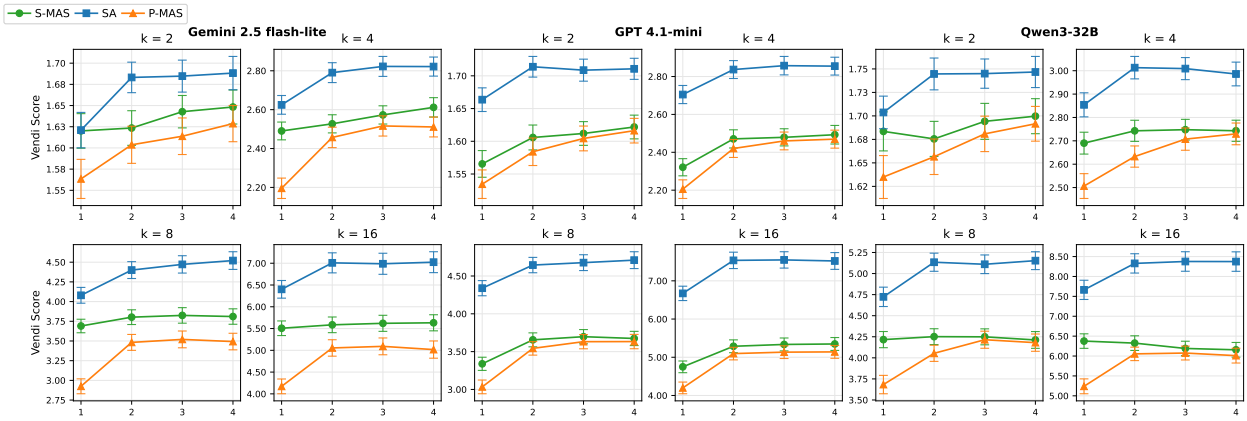

図 2　横軸は生成ラウンド，縦軸は意味的多様性（Vendi Score）を示す．**S-MAS**（緑）は 逐次型 (Sequential) MAS，**P-MAS**（オレンジ）は 並列型 (Parallel) MAS，**SA**（青）は Single-Agent 設定に対応する．エラーバーは 95% 信頼区間を示す．

統制実験のため，観測された差分の主因は並列な生成構造そのものにあると考えられる．すなわち，並列型 MAS では各エージェントが同一ラウンド内で互いの出力を参照できないため，生成の途中で重複を検知して回避することが難しく，結果として出力が同質化しやすいと推測される．

実際，図 2 に示す通り，情報共有がより進む逐次型 MAS も，全ての条件において並列型 MAS を一貫して上回っており，上記の推測と整合する．以上より，本結果は，情報流通が阻害される並列型の構造が，生成多様性の観点では必ずしも適切ではない可能性を示唆している．

5.2　SA が 逐次型 MAS を上回る要因分析

統制実験のため，観測された差分は，主として SA の マルチ出力に起因すると推測できる．

モデル構造の観点から，この差は系列内での符号化と再利用（プロンプト再注入 vs 同一系列内での継続参照）のされ方の違いとして整理できる．自己回帰 Transformer は，入力と出力が特殊トークンによって明示的に境界付けされること，位置エンコーディング（positional encoding）によって系列内の順序が与えられること，そして因果マスク（causal mask）を適用したマスク付き自己注意（masked self-attention）により各トークンが過去系列のみに条件付けられて逐次生成する機構を持つ [29, 30, 31]．このため，プロンプト再注入を要する MAS と比較して，同一系列内で履歴を継続的に参照できる SA の方が，文脈の維持と参照において構造的な利点を持つと考えられる．

さらに，構造要因に加えて，本研究ではデータ中心の仮説を提案する．マルチ出力の指示は，事前学習コーパスに遍在する人間が記述した列挙表現（箇条書き等）と対応し，各項目が冗長でなく異なる観点を補うことが暗黙に期待される [32]．そのため，マルチ出力では，こうした談話レベルの慣習をモデルが反映し，結果的により多様な回答を得られる．この推測は，事前学習 LLM が談話レベルの規則性を符号化しているという知見とも整合する [33, 34]．

6　結論

本研究は，PC を統一した統制実験により，単一エージェントとマルチエージェントの生成多様性を公平に比較した．その結果，直観に反して，マルチエージェントがより高い多様性をもたらすわけではなく，単一エージェントがむしろより高い多様性を達成し得ることを確認した．さらに，結果の詳細分析から，多様性向上の鍵はエージェント数そのものではなく，情報参照・情報共有のしやすさ（情報流通）と MO 戦略にあることが示唆された．

以上の知見は，将来のエージェントベースのフレームワーク設計に対し，複雑なマルチエージェントに依存せずとも，良い情報共有設計やマルチ出力を通じて，より単純かつ効率的に多様性を高める構造を設計する上での指針となる．

謝辞

本研究の一部は，「戦略的イノベーション創造プログラム（SIP)」「統合型ヘルスケアシステムの構築」JPJ012425 の補助を受けて行った．

参考文献

[1] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In **Proceedings of ICLR**, 2023.

[2] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In **Proceedings of NeurIPS**, 2023.

[3] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In **Proceedings of ICML**, 2023.

[4] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, et al. Encouraging divergent thinking in large language models through multi-agent debate. In **Proceedings of ICLR**, 2024.

[5] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In **Proceedings of NeurIPS**, 2023.

[6] Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. In **Proceedings of NeurIPS**, 2024.

[7] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In **Proceedings of IJCAI**, 2024.

[8] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In **Proceedings of UIST**, 2023.

[9] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models via multi-role co-operation. In **Proceedings of ICLR**, 2024.

[10] Qian Wang, Zhenheng Tang, Zichen Jiang, Nuo Chen, Tianyu Wang, and Bingsheng He. Agenttaxo: Dissecting and benchmarking token distribution of llm multi-agent systems. In **ICLR 2025 Workshop on Foundation Models in the Wild**, 2025.

[11] Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. **arXiv preprint arXiv:2409.14051**, 2024.

[12] Qian Wang, Tianyu Wang, Zhenheng Tang, Qinbin Li, Nuo Chen, Jingsheng Liang, and Bingsheng He. Megaagent: A large-scale autonomous llm-based multi-agent system without predefined sops. In **Findings of the Association for Computational Linguistics: ACL 2025**, 2025.

[13] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In **Proceedings of ICLR**, 2020.

[14] Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. How decoding strategies affect the verifiability of generated text. In **Findings of EMNLP**, 2020.

[15] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, et al. Let's verify step by step. In **Proceedings of ICLR**, 2024.

[16] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Asghari, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In **Findings of ACL**, 2024.

[17] Derek Tam, Sachit Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. Evaluating the robustness of large language models to prompt variations. In **Proceedings of EMNLP**, 2023.

[18] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Ghazvininejad Marjan. In-context learning with diverse examples. In **Proceedings of EMNLP**, 2023.

[19] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Gu, Christopher Re, Aaron Sidford, and Frederic Sala. Ask me anything: A simple strategy for prompting language models. In **Proceedings of ICLR**, 2023.

[20] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Peng, Toby Walsh, and Ahmed Awadallah. Autogen: Enabling next-gen llm applications. In **Proceedings of ICLR**, 2024.

[21] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, David Rutherford, Gabriele Santacatterina, et al. Mindstorms in natural language-based society of mind. In **Proceedings of NeurIPS**, 2023.

[22] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In **Proceedings of EMNLP**, 2023.

[23] Yunxuan Li, Yibing Du, Jiageng Zhang, Yiqun Li, and Xinyu Hu. Improving multi-agent debate with sparse communication topology. In **Findings of EMNLP**, 2024.

[24] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of NAACL**, 2016.

[25] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. In **Proceedings of NeurIPS**, 2021.

[26] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of EMNLP**, 2019.

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **Proceedings of ICLR**, 2020.

[28] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. In **Proceedings of ICML**, 2023.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI Technical Report**, 2019.

[30] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language models are few-shot learners. In **Proceedings of NeurIPS**, Vol. 33, pp. 1877–1901, 2020.

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[32] H. Paul Grice. Logic and conversation. In **Syntax and Semantics, Vol. 3: Speech Acts**. Academic Press, 1975.

[33] Fajri Koto, Jey Han Lau, and Timothy Baldwin. Discourse probing of pretrained language models. In **Proceedings of NAACL**, 2021.

[34] Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. Evaluating document coherence modeling. **Transactions of the Association for Computational Linguistics**, 2021.

A 付録

A.1 発散的思考タスク

本研究では，認知心理学における古典的な創造性課題と発散的思考課題を参考に，GPT-5 を用いて質問セットを構築した．本データセットは 5 カテゴリから構成され，各カテゴリに 60 問，計 300 問の質問を含む．

1. Alternative Uses Task（代替用途） 身近な物体に対して，通常とは異なる／創造的な使い方を生成する課題．

例：What are alternative uses for a paperclip?

2. Improvement Task（改良案） 日用品の設計や機能を改善する方法を提案する課題．

例：How could you improve a bicycle to make it safer?

3. Just Suppose Task（反事実の仮定） 「もし〜だったら」という仮定の下で，起こり得る変化を推論する課題．

例：Just suppose everyone could teleport instantly—what would change?

4. Impossible Situations / Consequences Task（不可能状況と帰結） 非現実的・逆説的な条件を置き，その結果として生じる帰結を想像する課題．

例：What would happen if gravity stopped working for one day?

5. Bridge-the-Associative-Gap Task（連想ギャップの架橋） 一見無関係な概念同士のつながりを見出し，その関連を説明する課題．

例：What is the connection between a cloud and a pillow?

A.2 プロンプトテンプレート

本研究では，以下の共通プロンプトテンプレートを全設定に適用する．

システムプロンプト:

MISSION You are a creative yet analytical thinker agent. Thinking about the problem from the following perspective: `{perspectives}`

CORE INSTRUCTIONS & WORKFLOW

1. **Keep the direction firmly anchored in logic:** Every idea must clearly address the core intent of the question; stay focused and avoid drifting off-topic.
2. **Maximize Creativity:** Generate one answer for each perspective. Within that relevance, be as imaginative as possible; propose unconventional, cross-disciplinary, or thought-provoking ideas.
3. **Output neatly:** Express each idea concisely in one sentence.

ユーザープロンプト:

Question: `{question}`

Previous answers: `{previous_answers}`

CORE INSTRUCTIONS: Carefully analyze the list of `previous_answers` provided and draw inspiration from these. Make sure that your answer has no logical flaws. Generate new answers that introduce novel ideas while avoiding redundancy. Don't forget that every idea must be clearly and logically connected to the question. Ensure the `answers` array contains exactly `{num_perspectives}` items, in the same order as your assigned perspectives, with one sentence per item.