通信トポロジは議論を左右する：マルチエージェント LLM における正解率と多様性の分析

本田純也 [1]　坂本航太郎 [2]　浮田純平 [2]　小島武 [2]
岩澤有祐 [2]　松尾豊 [2]
[1] 豊橋技術科学大学　[2] 東京大学
{kotaro.sakamoto,iwasawa,matsuo}@weblab.t.u-tokyo.ac.jp　{honda.junya.qt}@tut.ac.jp

概要

複数の大規模言語モデル（LLM）エージェントに議論させるマルチエージェント議論は，単一モデルの誤りを相互に訂正できる一方，「強い単一エージェント＋多数決」を上回らない場合も報告されている．本稿は，議論のプロトコルを固定したまま通信トポロジと議論の深さのみを制御し，正解率と多様性の時間発展を系統的に比較する．リング，スター，完全グラフ，Erdős–Rényi，スモールワールド，スケールフリーの 6 種を対象に，各ラウンドで近傍要約を共有しつつ<final>を出力して多数決で集約した．語彙指標（n-gram エントロピー等）と埋め込み空間における分散を解析すると，(1) 疎なリングは混合が遅く初期改善が鈍い，(2) ハブを持つグラフは 1–2 ラウンドで高精度に達しやすい，(3) 語彙多様性と意味多様性は乖離し得る，という傾向が得られた．さらに，日本語タスクへ置き換える際の指標設計上の注意点を議論する．

1　はじめに

大規模言語モデル（LLM）は多様な自然言語処理タスクで高性能を示す一方，幻覚や推論誤りが残る．近年は，複数の LLM を**エージェント**として並列に走らせ，議論・相互批判・多数決などで最終出力の信頼性を高める枠組みが広く検討されている [3, 4, 5, 7].

しかし，多数の研究が前提とする「全員が全員の発言を毎ラウンド共有する（完全グラフ）」は，計算コストが高いだけでなく，過度な同調（意見の早期収束）を招き得る．一方で，実際の会議・査読・意思決定では，情報共有経路（誰が誰と話すか）すなわち**通信トポロジ**が固定ではない．にもかかわらず，トポロジと議論深さ（ラウンド数）を系統的に比較し，正確度と多様性のトレードオフを定量化した報告は限られている．

本稿では，議論プロトコル（プロンプト，更新規則，集約則）を固定し，**トポロジと深さのみ**を変えた比較を行う．特に，(i) 疎な規則グラフ（リング），(ii) ハブ中心（スター），(iii) 完全グラフ，(iv) ランダム（Erdős–Rényi）[2]，(v) スモールワールド（Watts–Strogatz）[6]，(vi) スケールフリー（Barabási–Albert）[1] を対象に，多様性の変化を語彙・意味の両面から追跡する．

本稿の貢献

- マルチエージェント議論を**通信グラフ上の同期更新**として定式化し，トポロジとラウンド数の影響を切り分けて比較する．
- 語彙指標（n-gram エントロピー等）と埋め込み空間の意味分散を併記し，多様性の指標間の乖離を明示する．
- 日本語タスクへ置き換える場合に生じる評価上の落とし穴（表記ゆれ・分かち書き・敬体の同調など）を整理し，実験設計の指針を示す．

2　問題設定：通信グラフ上のマルチエージェント議論

2.1　通信トポロジ

エージェント集合を V，通信路を E とするグラフ $G = (V, E)$ を考える．各エージェント $i \in V$ は近傍 $\mathcal{N}(i) = \{j \mid (i, j) \in E\}$ からのみ情報を受け取る（無向グラフを基本とする）．本稿では典型的な 6 種類のトポロジを扱う：Ring（各ノードが局所近傍のみと接続），Star（中心ハブと周辺），Clique（全結合），ER（$G(n, p)$），WS（スモールワールド），BA（スケールフリー）である [2, 6, 1].

2.2 同期ラウンド型の議論プロトコル

各ラウンド $t = 1, \ldots, T$ で，エージェント i は (1) タスク入力，(2) 前ラウンドの近傍メッセージ（要約）を受け取り，応答 $\mathbf{o}_i^{(t)}$ を生成する．応答には近傍へ送る短い要約と，解答部分<final>を含める．最後に集約器が $\{\mathbf{o}_i^{(T)}\}_{i \in V}$ から最終解 $\hat{y}$ を決める．分類・選択問題では**多数決**が自然であり，数値解答では正規化（単位・丸め）後の一致判定を用いる．

ここで重要なのは，深さ T は推論コスト（トークン消費・API 費用・遅延）にほぼ比例し，トポロジは各ラウンドの受信量（コンテキスト長）を支配することである．したがって，**精度だけでなく**，多様性の維持とコストの観点からもトポロジ設計が必要になる．

3 多様性の定義と測定

マルチエージェント議論では，「多様性が高いほど探索が進む」一方，「多様性が高過ぎると合意形成ができない」という直観がある．しかし，多様性は観測量であるため，測り方により結論が変わり得る．本稿では語彙と意味の 2 種類の指標を併用する．

3.1 語彙多様性

各ラウンド t における全エージェントの出力（または要約部）を連結し，トークン列 $w_{1:L}$ を得る．unigram 分布 p_1 および n-gram 分布 p_n に対し，

$$H_n(t) = -\sum_{g \in \mathcal{G}_n} p_n(g) \log p_n(g) \tag{1}$$

を定義する（$\mathcal{G}_n$ は観測された n-gram 集合）．また，型トークン比（Type–Token Ratio）等も併用できる．

3.2 意味多様性

語彙指標は表記ゆれや言い換えに敏感であり，「言い換えが増えただけ」を多様性増と誤認することがある．そこで，各エージェント出力 $\mathbf{o}_i^{(t)}$ を埋め込み $\mathbf{e}_i^{(t)} \in \mathbb{R}^d$ に写像し，エージェント間のペア距離

$$D(t) = \mathrm{median}_{i \neq j}\left(1 - \cos(\mathbf{e}_i^{(t)}, \mathbf{e}_j^{(t)})\right) \tag{2}$$

を意味多様性の指標として用いる．$D(t)$ は「同じ意味に収束しているか」を比較的直接に反映する．

4 実験：トポロジと深さの比較

4.1 設定

基本設定としてエージェント数を $n = 8$，ラウンド数を $T \leq 10$ の浅い議論を想定する．各エージェントには（モデルやシステムメッセージの違いを含む）異なる初期バイアスが入るように設定し，初期解が一致しない状況から議論を開始する．タスクは，短い推論・計算を要する問題群を例にとり，最終解の正解率と多様性の推移を観測する．

4.2 主結果

図 1 は典型例として，GSM-Plus 系の算術推論タスクにおける挙動を示す（図は提出時に差し替え）．観測された傾向をまとめると次の通りである．

疎な規則グラフは混合が遅い　Ring では情報伝播の最短距離が長く，誤り訂正に必要な反例・指摘が全体に行き渡るまでラウンドを要する．このため，初期ラウンドでの正解率改善が鈍い傾向がある．

ハブを持つトポロジは少ラウンドで合意しやすい　Star や BA では，ハブが近傍情報を集約しやすく，1–2 ラウンドで多数決が安定しやすい．一方，Clique のように全結合で毎ラウンド全情報を共有すると，早期に合意が形成されるが，多様性（特に意味多様性）が急速に低下し得る．

語彙多様性と意味多様性は乖離し得る　語彙指標 $H_n(t)$ が上昇していても，$D(t)$ が低下する場合がある．これは，エージェントが同一結論に向かって言い換え・説明追加を行い，表層が多様化しつつ意味が収束するケースに対応する．したがって，多様性制御や「収束し過ぎ」の検知には，語彙指標のみでは不十分である可能性が高い．

設計上の含意　以上より，(a) 浅いラウンドでの改善を重視する場合はハブ型（Star/BA）やスモールワールド（WS）が有利になり得る，(b)Clique はコストが高いだけでなく，早期同調により探索を打ち切る危険がある，(c) 早期停止（T の自動決定）には語彙指標より意味指標が有用，といった指針が得られる．

5 日本語タスクに置き換える際の論点

ここでは，日本語タスクへの適用について議論する．

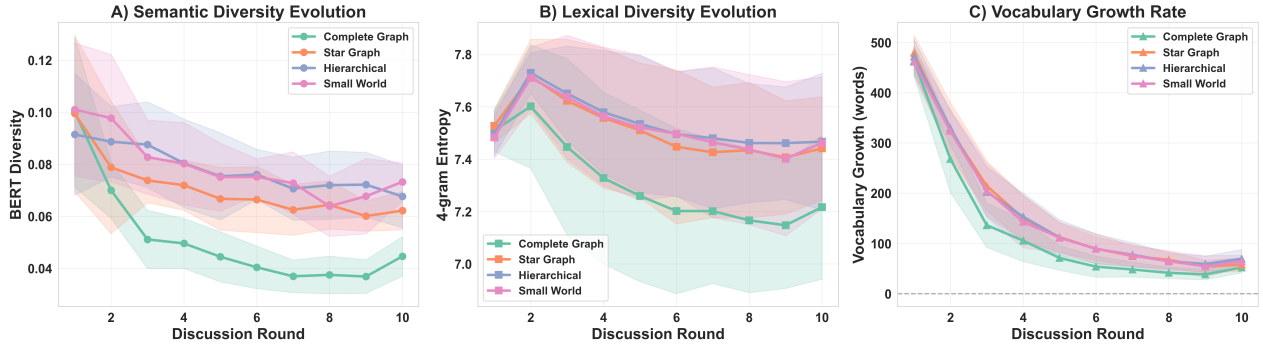

図 1　通信トポロジ別の正解率（左）・語彙多様性（中）・意味多様性（右）の推移の例（プレースホルダ）．主要な観察は，リングの遅い混合，ハブ型の速い合意，語彙と意味の乖離である．

5.1　評価（正解率）の難しさ

日本語では，(i) 表記ゆれ（漢字／ひらがな／カタカナ，数字表記），(ii) 助詞・敬体の違い，(iii) 省略・照応により，同一意味でも文字列一致が取りづらい．選択式タスクでは影響が小さいが，生成式（要約・説明・自由記述）では**正解判定器**自体がボトルネックになる．このため，(a) 正規化ルール（数字・記号・全半角）を明示した上での一致判定，(b) 人手または LLM 判定器によるペア判定，(c) 情報抽出型の自動評価（必要スロットの充足）など，タスクに応じた評価設計が必要である．

5.2　語彙多様性が過大評価されやすい

語彙指標 $H_n(t)$ はトークナイザに依存する．日本語 BPE は，漢字 1 文字・ひらがな連続・外来語などでサブワード分割が変わり，表記ゆれがそのまま多様性として計上されやすい．したがって，日本語で語彙多様性を議論する場合は，

- 形態素解析に基づく正規化（原形・品詞）を通した上で n-gram を作る，
- 数字・固有表現・記号の正規化を行う，
- 可能なら読み（かな）に落とした指標も併記する，

などの工夫が望ましい．

5.3　意味多様性のモデル選択

意味指標 $D(t)$ は埋め込みモデルの言語能力に依存する．英語中心の埋め込みを流用すると，日本語の言い換えや省略を適切に距離に反映できない危険がある．多言語対応埋め込み（または日本語特化埋め込み）を用い，可能なら人手確認用の最近傍例を併記することで，指標の妥当性を担保できる．

5.4　日本語における「同調」とトポロジ

日本語の丁寧体・婉曲表現は，議論が進むにつれて文体が均質化しやすい．このとき語彙指標はむしろ低下するが，結論の多様性（賛否）は維持されている場合がある．また，敬体での「相手への配慮」が強いプロンプトは，同調（迎合）を誘発する可能性がある．したがって，日本語設定では，(a) 役割（批判役・検証役・反例探索役）を明示する [5]，(b) 反対意見の提出を**義務化**する，(c) トポロジとしては Clique よりも「部分的に隔離された探索」と「少数の橋渡し」を併せ持つ WS のような構造を検討する，といった設計が有効だと考えられる．

6　おわりに

本稿は，マルチエージェント LLM の議論を通信グラフ上の同期更新として捉え，通信トポロジと議論深さが正解率と多様性に与える影響を整理した．疎な規則グラフは混合が遅く，ハブ型は少ラウンドで合意しやすい一方，完全グラフでは早期収束により探索が止まる可能性がある．また，語彙多様性と意味多様性は乖離し得るため，早期停止や探索度合いの判断には意味指標の併用が重要である．日本語タスクでは表記ゆれ・トークナイザ依存・同調様式が異なるため，語彙指標の正規化と多言語埋め込みの選定が特に重要になる．

参考文献

[1] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. **Science**, 286(5439):509–512, 1999.

[2] Paul Erdős and Alfréd Rényi. On random graphs I. **Publicationes Mathematicae Debrecen**, 6:290–297, 1959.

[3] 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 the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)**, 2024.

[4] 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**, pages 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[5] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. **arXiv preprint arXiv:2405.06373**, 2024.

[6] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. **Nature**, 393(6684):440–442, 1998.

[7] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. Chain of agents: Large language models collaborating on long-context tasks. **arXiv preprint arXiv:2406.02818**, 2024.