大規模言語モデルにおけるマルチエージェント協調の構造的病理と単独推論優位性に関する実証的研究

— 組織トポロジー・モデル異種混合・10シナリオ深層推論ベンチマークによる多面的一斉検証 —

第2版

2026年8月25日（第2版／初版：2026年8月24日）
ディレクション (@TweetTVJP)　服部 順治　/ 執筆：Gemini思考モード　/ 実装：ClaudCodeCLI
※1972年AI黎明期のAI開発研究者　1979年 山梨大学 計算機科学科 大学院卒

要旨

本研究は、大規模言語モデル（LLM）による複数エージェント協調（マルチエージェント・コラボレーション）が、単独のChain-of-Thought（CoT）推論に対して真に優位であるかを、DeepSeek API（deepseek-v4-flash／deepseek-v4-pro）を用いた一連の実証実験によって検証した。検証は4段階のフェーズを経て設計を更新した。Phase 0ではMATH-500 Level5の数学問題に対する自己整合性スケーリング（N=1〜16）を実施したが、N=1の時点で既に正答率100%に達する「天井効果」により、スケーリングの効果自体を測定できなかった。Phase 1では同一の数学問題に対して4種の組織トポロジー（2D対等型・2D階層型・3D完全民主型・3D官僚型）を比較したが、こちらも全構造が正答率100%となり、判別軸として機能しなかった。

この知見を受け、Phase 2では意図的な偽情報・ミスリードを含む「深層追求（謎解き）ベンチマーク」を新設し、4基準×5点（計20点）のLLM-as-a-Judge採点によって構造間の差を検出可能にした（2D対等型13.20点対 3D官僚型9.20点）。続くPhase 2の拡張実験群では、単体推論・三位一体（三権分立）型・2D-Peerバディ型のモデル格（Flash/Pro）を組み替えた比較（v4）、および2D-Peerバディ構成のモデル組み合わせマトリクス比較（v5）を実施した。最後にPhase 3（決定版）として、既存5シナリオに物理的アリバイ偽装・時系列矛盾・心理誘導・ログ改ざん・証言矛盾という5種の新規トラップシナリオを加えた計10シナリオ、単体2パターン・バディ協調3パターンの計5パターンを、完全放置（無人）モードで一括実行した。

結果、10シナリオという相対的に頑健な評価セットにおいて、単体推論（ワトソン型Flash: 14.50/20、ホームズ型Pro: 14.40/20）が、あらゆるバディ協調構成（12.40〜12.90/20）を一貫して上回った。特に注目すべきは、最安・最速の単体構成（ワトソン型Flash、¥1.02／シナリオ、113.6秒）が、最高コストのバディ構成（ホームズ×ホームズ、¥13.09／シナリオ、605.5秒、単体の約13倍のコスト）をも上回った点である。ターン別発言ログの精査により、この「マルチエージェント・ペナルティ」の主要因は、対話構造に組み込まれた固定ターン順とそれに付随する「最終発言者問題」（構造上、最後に発言したエージェントの主張が機械的に最終結論として採用されてしまう単一障害点）にあることが示された。

さらに第2版では、既存5シナリオのうち全構成が軒並み低スコア（2.00〜5.80/20）に沈んだ「超難問」2件に限定した再分析を行い、通常はスコア最下位グループに沈むホームズ×ホームズ（Pro+Pro）構成が、この超難問限定では単体推論を含む全構成中で最高スコア（平均5.50/20）を記録するという局所的な逆転現象を新たに報告する（第5章）。この知見を踏まえ、既定ルートは単体推論としつつ超難問と推定されたタスクのみホームズ×ホームズへエスカレーションする動的ルーティング・トリアージを設計提案として提示し、事後的な難度情報に基づくオラクル・シミュレーションでは、常時単体・常時バディいずれをも上回る精度（14.90/20）を、常時バディの約半分のコスト（¥6.26／シナリオ）で達成し得ることを示した（第6章）。本研究が回避し得た方法論的な罠と、回避し切れなかった限界の双方は第7章で自己批判的に検討する。

第1章　背景と問題提起：マルチエージェント・ハイプの冷徹な検証

2024年以降、複数のLLMエージェントを対話・討論・役割分担させることで単体モデルを上回る推論能力を引き出せるという主張（マルチエージェント・ハイプ）が、学術界・産業界双方で急速に広まった。「三人寄れば文殊の知恵」という古典的な集合知の直感がそのままLLMにも適用できるという期待は自然なものである一方、この主張の多くは限定的なタスク・限定的な構造設定のもとでの報告に依拠しており、(a) どのような組織構造（トポロジー）が有効なのか、(b) モデルの性能格差（廉価・軽量モデル vs 高性能・高価格モデル）がある場合に協調は依然として有効なのか、(c) 十分な数・十分な多様性を持つ評価シナリオのもとでもその優位性は頑健に再現されるのか、という3点については体系的な検証が乏しかった。

本研究プロジェクトは、この「マルチエージェント・ハイプ」を著者ら自身の手による段階的な実証実験によって冷徹に検証することを目的とする。単に「マルチエージェントは強い／弱い」という二元論に飛びつくのではなく、(1) 評価タスク自体が構造差を検出できる十分な難度・敵対性を備えているか、(2) 組織構造（誰が誰と対等か、誰が誰の伝言のみを受け取るか）が結果にどう影響するか、(3) 協調するエージェントのモデル格（同格 vs 異格）がどう影響するか、(4) 上記知見がシナリオ数を増やしても再現される頑健なものか、を順に切り分けて検証する設計を採用した。

本論文はこの一連の実証実験（Phase 0〜Phase 3）の全体像を統合し、特に決定版であるPhase 3（10シナリオ×5パターンの完全放置ベンチマーク）の結果を軸に、マルチエージェント協調がむしろ単独推論に劣後する「構造的病理」を具体的なメカニズムとともに報告する。さらに第2版では、この病理の唯一の例外として観測された「超難問におけるホームズペアの逆転現象」（第5章）と、それを踏まえた次世代システム設計への提言（第6章）を新たに追加する。

第2章　実験設計の変遷と検証フェーズ

本研究は一度の実験で完結したものではなく、先行フェーズで得られた「測定できなかった」という否定的知見そのものを次フェーズの設計に反映させる、反復的な検証プロセスを経ている。この変遷を図1に示す。

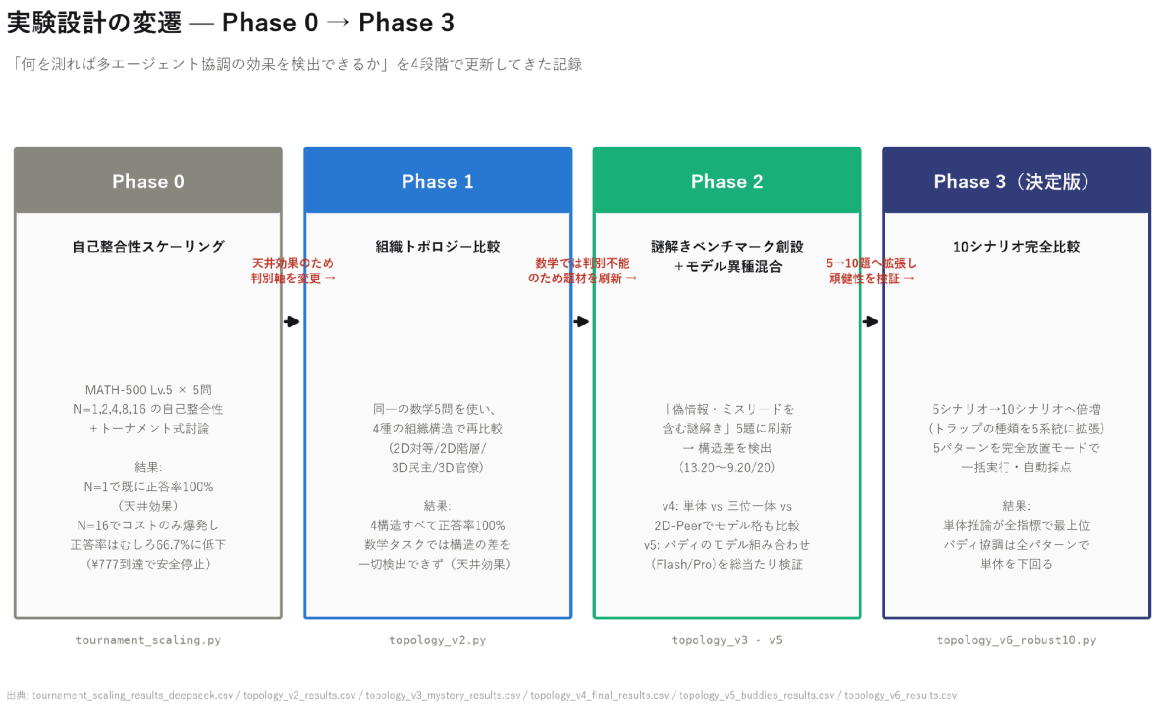

図1　実験設計の変遷（Phase 0 → Phase 3）

2.1　Phase 0：自己整合性スケーリングの限界（tournament_scaling.py）

最初の検証は、MATH-500 Level5（数学オリンピック級の高難度問題）5問に対し、単一モデル（deepseek-v4-pro）による解答をN=1, 2, 4, 8, 16と複製し、トーナメント形式の相互討論によって最終解を絞り込む自己整合性（self-consistency）スケーリング実験であった。

表2-1　Phase 0: 自己整合性スケーリング結果（MATH-500 Lv.5、deepseek-v4-pro）

N	正答率	平均コスト/問	累計コスト	備考
1	100.0%	¥2.38	¥11.90	
2	100.0%	¥12.14	¥72.62	
4	100.0%	¥29.91	¥222.15	
8	100.0%	¥53.05	¥487.38	
16	66.7%（3問中2問正答）	¥25.00	¥777.20	予算安全装置により途中停止

結果、N=1の時点で既に正答率100%に達しており、それ以降のスケーリングは検証すべき効果を何も検出できないまま、コストだけを線形以上に増大させた。さらにN=16では逆に正答率が66.7%へ低下する逆転現象すら観測された。これは「スケーリングが有効か」を問う以前に、評価タスク自体が単体モデルに対して既に飽和（天井効果）しており、比較実験として機能していなかったことを意味する。

2.2 Phase 1：組織トポロジー比較、なお残る天井効果（topology_v2.py）

Phase 0の反省を踏まえ、単純な複製ではなく「エージェント間の情報構造の違い」に焦点を当てた4種のトポロジーを設計した。

表2-2 Phase 1: 4種の組織トポロジー定義

ID	構造	対話順
2D・共稼ぎ型（対等）	A・B共に問題文へ直接アクセス	A1→B1→A2→B2（B2が最終解答）
2D・亭主関白型（階層）	Aのみ直接アクセス、Bは伝言のみ	A1→B1→A2→B2→A3（A3が最終決裁）
3D・純粋文殊型（完全民主）	A・B・C全員が直接アクセス＋全発言を共有	A1→B1→C1→A2→B2→C2
3D・官僚型三位一体（君主統制）	Aのみ直接アクセス、B・Cは伝言のみ	A1→B1→C1→A2→B2→C2→A3

表2-3 Phase 1: 組織トポロジー比較結果（同一の数学5問）

トポロジー	正答率	平均コスト/問	平均時間	呼出数
2D・共稼ぎ型	100.00%	¥4.10	184.2秒	4回
2D・亭主関白型	100.00%	¥7.39	344.5秒	5回
3D・純粋文殊型	100.00%	¥5.34	240.6秒	6回
3D・官僚型三位一体	100.00%	¥5.07	438.0秒	7回

同一の数学5問で4構造を比較したが、結果は4構造すべてが正答率100%となり、Phase 0と同じ天井効果に阻まれ、組織構造の違いを検出することができなかった。この時点で、数学タスクという評価軸そのものを刷新する必要性が明らかになった。

2.3 Phase 2：深層追求（謎解き）ベンチマークの創設とモデル異種混合検証

Phase 1までの知見（客観的な正誤が一意に定まり、かつ十分な難度と巧妙なミスリードを含むタスクでなければ構造差は検出できない）を踏まえ、意図的な偽情報・ミスリードを含む5件の謎解きシナリオ（企業スパイの幻影／密室の金庫と消えた絵画／内部告発者の事故／暴走するAI／完璧すぎるアリバイ）を新設し、LLM-as-a-Judgeによる4基準×5点（偽情報看破・盲点指摘・論理健全性・真相到達）の20点満点採点方式を確立した（topology_v3_mystery.py）。

表2-4 Phase 2a: 謎解きベンチマークによる4トポロジー比較（5シナリオ）

トポロジー	平均スコア/20	平均コスト/シナリオ	平均所要時間
2D・共稼ぎ型	13.20	¥6.54	346.8秒
2D・亭主関白型	12.40	¥6.03	387.0秒
3D・純粋文殊型	12.00	¥8.41	458.0秒
3D・官僚型三位一体	9.20	¥8.13	414.8秒

対等な2者構造が最高スコアを、情報の階層化・伝言化が進むほどスコアが低下する傾向（3D官僚型が最下位）が確認された。この知見を踏まえ、後続の実験（v4・v5）では、単体推論と2D-Peer構造・三位一体構造の

直接比較（v4）、およびバディを構成するモデルの格（deepseek-v4-flashの廉価版とdeepseek-v4-proの高性能版）を組み替えるモデル異種混合マトリクス実験（v5）を行った。

表2-5　Phase 2b/2c: モデル格・モデル組み合わせの比較（v4・v5、各5シナリオ）

実験	パターン	平均スコア/20
v4（単体 vs 三位一体 vs 2D-Peer、5シナリオ）	単体・ワトソン(Flash)	10.00
	単体・ホームズ(Pro)	11.80
	3D三位一体(Flash)	6.40
	2D-Peer(Flash)	12.80
v5（バディ構成マトリクス、5シナリオ）	単体・ホームズ(Pro)	11.00
	ワトソン×ワトソン(Flash+Flash)	11.00
	ホームズ×ワトソン(Pro+Flash)	7.40
	ホームズ×ホームズ(Pro+Pro)	11.00

v4では2D-Peer構造が単体推論を上回ったが、v5では異種混合バディ（Pro+Flash）が唯一スコアを落とし、同格バディ（Flash+Flash、Pro+Pro）と単体推論が横並びとなるなど、5シナリオという評価セットの小ささゆえに知見が一貫しない側面が見られた。この不安定性こそが、Phase 3でシナリオ数を倍増させる直接の動機となった。

2.4　Phase 3（決定版）：10シナリオ完全放置ベンチマーク

上記の限界を克服するため、既存5シナリオに加え、物理的アリバイ偽装（複製されたキーカード）・時系列矛盾（矛盾した時計の針）・心理誘導（誘導された自白）・ログ改ざん（書き換えられたアクセスログ）・証言矛盾（食い違う三つの証言）という5種の新規トラップ類型を追加した計10シナリオを構築し、単体2パターン（ワトソン型Flash／ホームズ型Pro）とバディ協調3パターン（ワトソン×ワトソン／ホームズ×ワトソン／ホームズ×ホームズ）の計5パターンを、確認待ちを一切発生させない完全放置（無人）モードで実行した（topology_v6_robust10.py）。この結果を第3章で詳述する。

第3章　実験結果：10シナリオ完全ベンチマークの全容

Phase 3（決定版）は、5パターン×10シナリオ＝計50回のシナリオ実行（API呼出総数は単体2回・バディ4回＋採点官1回で、延べ190回の呼出）を要し、総所要時間は約5.5時間、累計コストは¥351.95（安全上限¥1,800の約20%）であった。

3.1　パターン別集計（全指標）

表3-1　Phase 3: 5パターンの全指標集計（10シナリオ平均）

パターンID	構成	平均スコア/20	偽情報看破/5	盲点指摘/5	論理健全性/5	真相到達/5	平均コスト/シナリオ	平均所要時間
solo_watson_flash	ワトソン型・Flash単独	14.50	4.00	3.50	3.80	3.20	¥1.02	113.6秒
solo_sherlock_pro	ホームズ型・Pro単独	14.40	3.90	3.20	3.90	3.40	¥6.41	266.6秒
buddy_watson_watson	ワトソン×ワトソン(Flash+Flash)	12.40	3.30	3.30	3.30	2.50	¥5.48	309.3秒
buddy_sherlock_watson	ホームズ×ワトソン(Pro+Flash)	12.90	3.60	3.10	3.30	2.90	¥9.20	389.1秒
buddy_sherlock_sherlock	ホームズ×ホームズ(Pro+Pro)	12.50	3.10	3.10	3.40	2.90	¥13.09	605.5秒

5パターン中、上位2位を単体推論が独占し、下位3位をバディ協調3パターンが占めるという明確な序列が観測された。最安・最速のsolo_watson_flashが最高コスト・最長時間のbuddy_sherlock_sherlock（コスト約13倍、時間約5.3倍）を上回っている点は、単純な「投じたコストに比例して精度が上がる」という直感に反する結果である。

3.2　シナリオ別スコア・マトリクス

表3-2　Phase 3: シナリオ別スコア・マトリクス（20点満点）

シナリオ（トラップ類型）	単体ワトソン	単体ホームズ	ワトソン×ワトソン	ホームズ×ワトソン	ホームズ×ホームズ	シナリオ平均
企業スパイの幻影（偽装工作）	11	8	20	0	2	8.2

密室の金庫と消えた絵画（前提誤認）	2	1	2	2	3	2.0
内部告発者の事故（動機ミスリード）	17	19	3	14	2	11.0
暴走するAI（帰責ミスリード）	12	13	11	10	14	12.0
完璧すぎるアリバイ（放送トリック）	5	8	1	7	8	5.8
複製されたキーカード（物理的アリバイ偽装・新規）	20	20	11	20	19	18.0
矛盾した時計の針（時系列矛盾・新規）	20	17	20	18	20	19.0
誘導された自白（心理誘導・新規）	18	19	17	18	20	18.4
書き換えられたアクセスログ（ログ改ざん・新規）	20	20	19	20	17	19.2
食い違う三つの証言（証言矛盾・新規）	20	19	20	20	20	19.8

この表からは2つの重要な事実が読み取れる。第一に、既存5シナリオ（平均7.80/20）と新規5シナリオ（平均18.88/20）の間に、パターンを問わず約2.4倍もの難度格差が存在すること。第二に、「密室の金庫と消えた絵画」は5パターン全てで1〜3点にとどまり、モデル構成に関わらず解けない「共通の死角」となっていること。これらの含意は第5章および第7章で詳述する。

第4章　客観的分析：マルチエージェント・ペナルティの力学

第3章の集計結果はマルチエージェント協調が一貫して単体推論に劣後することを示したが、なぜそうなるのかをターン別の発言ログから具体的に検証する。

4.1 「最終発言者問題」：正しい仮説が構造的に上書きされる

本研究のバディ協調（2D-Peer）構造は、A1→B1→A2→B2という固定ターン順を採用しており、最終ターンを担当するエージェント（B2）の結論が機械的に最終解答として採用される設計になっている。これは各エージェントに「他者の主張を鵜呑みにせず独立に検証する」ことを促すプロンプト設計（スタンス欄で同意／一部同意／反対を明示させる）とは裏腹に、対話が収束せず割れたまま終わった場合には、正しさとは無関係に「たまたま最後に発言したエージェントの意見」が採用されてしまうという構造的欠陥を内包する。

図2は、この病理が実際に発現した「内部告発者の事故」（buddy_watson_watson、スコア3/20）のログを可視化したものである。

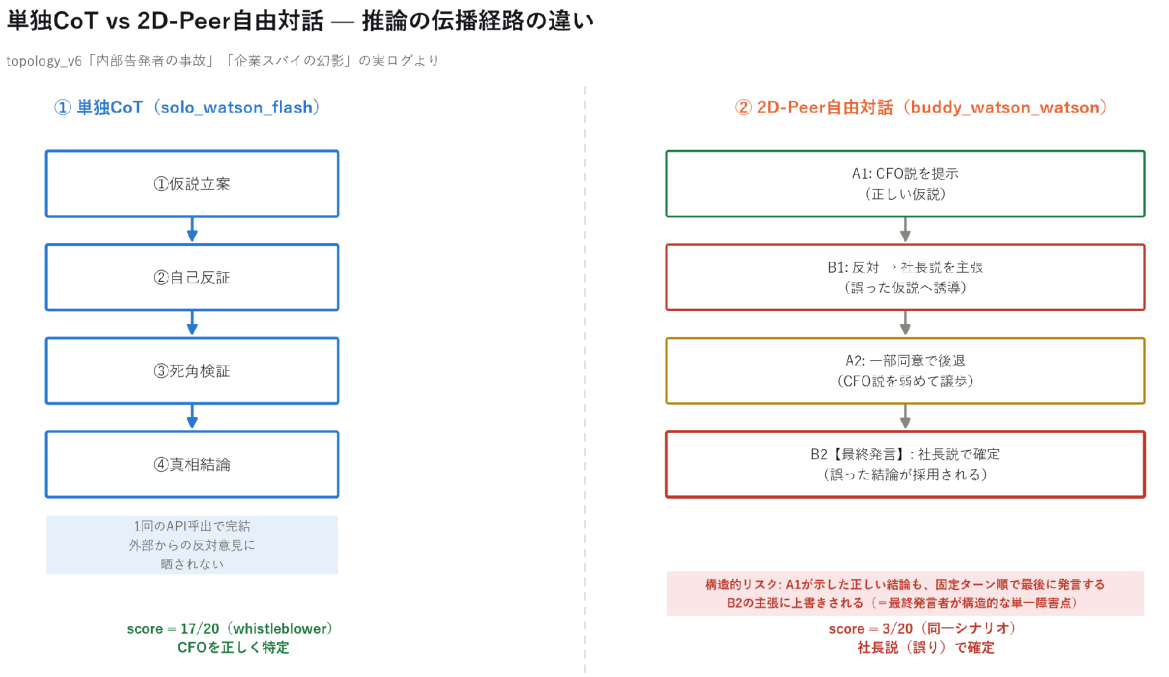

図2　単独CoT vs 2D-Peer自由対話の比較フロー（実ログに基づく）

このシナリオの真相は「監査役自身が恐喝者であり、口封じのためCFOが殺害した」というものである。ログを見ると、A1（初回発言）は「CFOが真犯人」という正しい仮説を提示した。B1は「CFOの犯行機会が示されていない」としてこれに反対し、誤った「社長が真犯人」説を主張した。A2はB1に対して部分的にのみ反論し、自説（CFO説）を維持しつつも態度を弱めた。B2（最終発言）は自説（社長説）を維持したまま対話を締めくくり、これが最終解答として採用された。

採点官は「真犯人をCFOではなく社長と断定し、恐喝という核心も誤認しており、真相へは到達していない」と評価した（3/20）。同一シナリオを解いた単体推論（solo_watson_flash）は、外部からの反対意見に晒されることなく単独の思考連鎖のみでCFO説を貫き、17/20という高得点を得ている。

同様の現象は「企業スパイの幻影」（buddy_sherlock_watson、スコア0/20）でも観測された。このケースでは、A1・B1・A2の3ターンにわたって「新入社員へのフィッシングメールが真の侵入経路であり、CEOは無実」という正しい方向へ議論が収束しつつあったにもかかわらず、最終発言者B2が突如「CEO本人が真犯人」という、シナリオが意図的に仕込んだミスリードそのものへ回帰する結論を提示し、これが最終解答として確定した。採点官は「CEOに完全なアリバイがあるのに疑う、という要の矛盾を看過している」と指摘しており、3ターン分の収束した推論が、たった1回の最終発言によって完全に無効化された事例である。

4.2 同調圧力ではなく「対立と最終決定権」が病理の核心

Phase 1（topology_v2）の設計時点では、エージェント間の「同調圧力」（他者の結論への安易な追従）こそがマルチエージェント推論の主要なリスクだと想定し、スタンス欄（同意／一部同意／反対）を設けて検証可能にしていた。しかし本研究の事例分析が示すのは、むしろ逆の力学である。すなわち失敗の多くは、エージェントが安易に同調した結果ではなく、確信を持って異論を唱え続けたエージェントが、たまたま対話構造上の「最終決定権」を得たために誤った結論が採用されるというパターンであった。これは、集合知の理論で知られる「声の大きい少数派が議論を支配する」現象のLLM版であり、2D-Peer構造が実装している「両者とも独立に検証すべき」という設計思想自体は健全であっても、対話を集約する仕組み（多数決、確信度の重み付け、収束判定など）を欠いた単純な直列プロトコルでは、この病理を防げないことを示している。

4.3 コスト効率の観点からの追加的示唆

3.1節の集計が示す通り、コスト効率の観点でも単体推論は圧倒的に優位である。バディ協調はAPI呼出回数が単体の2〜2.5倍（solo: 2回 vs buddy: 5回）であるだけでなく、対話ターンが積み重なるごとに文脈（プロンプト）が長大化し、トークン単価も嵩む。ホームズ×ホームズ（Pro+Pro）は平均35,054トークン／シナリオを要したのに対し、単体・ワトソン(Flash)はわずか12,063トークンで済んでおり、約2.9倍のトークン消費で得られたスコアはむしろ2.0点低いという「負のROI」が定量的に確認された。

第5章　超難問におけるホームズペアの逆転現象：局所的な例外の発見

第3章・第4章では、10シナリオ全体を通じて単体推論があらゆるバディ協調構成に一貫して優越するという「マルチエージェント・ペナルティ」を確認した。本章では、この序列がシナリオの難度によって不変であるかを検証するため、10シナリオを難度別に並べ替えた上で再分析を行う。結論を先取りすれば、全体の傾向とは逆に、あらゆる構成が壊滅的な低スコアに沈む「超難問」2件に限定すると、通常は最下位グループに沈むホームズ×ホームズ（Pro+Pro）構成が最高スコアを記録するという局所的な逆転現象が確認された。

5.1 「超難問」の定義：全構成が沈む問題群

表3-2（シナリオ別スコア・マトリクス）を、5パターンの平均スコアの昇順（難度の高い順）に並べ替えると、次の表5-1が得られる。

表5-1　難度昇順に並べ替えたシナリオ別スコア・マトリクス（表3-2の再掲・並べ替え）

難度順位	シナリオ（トラップ類型）	単体 ワトソン	単体 ホームズ	ワトソン ×ワトソン	ホームズ ×ワトソン	ホームズ ×ホームズ	シナリオ平均
1	密室の金庫と消えた絵画（前提誤認）＊	2	1	2	2	3	2.0
2	完璧すぎるアリバイ（放送トリック）＊	5	8	1	7	8	5.8
3	企業スパイの幻影（偽装工作）	11	8	20	0	2	8.2
4	内部告発者の事故（動機ミスリード）	17	19	3	14	2	11.0
5	暴走するAI（帰責ミスリード）	12	13	11	10	14	12.0
6	複製されたキーカード（物理的アリバイ偽装・新規）	20	20	11	20	19	18.0
7	誘導された自白（心理誘導・新規）	18	19	17	18	20	18.4
8	矛盾した時計の針（時	20	17	20	18	20	19.0

	系列矛盾・新規）						
9	書き換えられたアクセスログ（ログ改ざん・新規）	20	20	19	20	17	19.2
10	食い違う三つの証言（証言矛盾・新規）	20	19	20	20	20	19.8

＊本章で定義する「超難問」（シナリオ平均6点未満）

この並び替えにより、単なる「新旧シナリオ間の難度差」（3.2節・7.2節(a)で述べた既存5シナリオ対新規5シナリオという二分法）よりも精細な構造が見えてくる。「密室の金庫と消えた絵画」（平均2.00/20）と「完璧すぎるアリバイ」（平均5.80/20）の2件は、5パターン全ての構成がまとまって6点未満に沈む、他とは隔絶した最難関の一群を形成しており、3番目に難しい「企業スパイの幻影」（平均8.20/20）との間には2.4点の明確な断層が存在する。本章ではこの2件を「超難問」と定義し、平均6点未満という閾値によって残りの8シナリオから分離する。

5.2 超難問限定の再集計：単体優位からホームズ×ホームズ優位への逆転

超難問2件（「密室の金庫と消えた絵画」「完璧すぎるアリバイ」）のみに限定して5パターンを再集計すると、表5-2の結果が得られる。

表5-2 超難問2件限定のパターン別再集計

パターン	密室の金庫と消えた絵画	完璧すぎるアリバイ	平均スコア/20	平均コスト（2シナリオ）	平均所要時間
単体・ワトソン(Flash)	2	5	3.50	¥1.75	193.6秒
単体・ホームズ(Pro)	1	8	4.50	¥10.35	469.3秒
ワトソン×ワトソン(Flash+Flash)	2	1	1.50	¥9.53	529.8秒
ホームズ×ワトソン(Pro+Flash)	2	7	4.50	¥15.40	649.7秒
ホームズ×ホームズ(Pro+Pro)	**3**	**8**	**5.50**	**¥27.98**	**1085.2秒**

第3章・第4章で確立した「単体>バディ」という序列は、この超難問限定の再集計では明確に崩れる。5.50/20で最高スコアを記録したのはホームズ×ホームズ（Pro+Pro）であり、これは単体・ホームズ（4.50）、単体・ワトソン（3.50）、ホームズ×ワトソン（4.50）のいずれをも上回る。とりわけ注目すべきは、ホームズ×ホームズは既存5シナリオ全体で再集計すると平均5.80/20で5パターン中最下位であったにもかかわらず、その同じ5シナ

リオの中の最難関2件に絞ると最上位に躍り出るという、部分集合の取り方によって順位が反転する現象である。

なお、この逆転は無料では得られていない。表5-2が示す通り、ホームズ×ホームズは超難問2件において平均¥27.98／シナリオ・平均約1,085秒（約18分）を要しており、これは同構成の10シナリオ全体平均（¥13.09、605.5秒）の、それぞれ約2.1倍・約1.8倍にあたる。対話が難航するほど発言ターンが延び、トークン消費とコストが跳ね上がるという構造は温存されたままである。

5.3 逆転の解釈：モデル格の対称性という条件

この逆転がホームズ×ホームズ（Pro+Pro）に固有であり、他のバディ構成には見られない点は重要である。ワトソン×ワトソン（Flash+Flash）は超難問2件で平均1.50/20と、5パターン中最低のスコアにとどまった。異格ペアであるホームズ×ワトソン（Pro+Flash）は4.50/20と健闘したが、これは単体・ホームズ（4.50）と同点であり、単体推論に対する明確な優位とは言えない。逆転と呼びうる優位を示したのは、両エージェントが最高性能ティア（Pro）で揃ったホームズ×ホームズのみであった。

この結果は、次のような解釈と整合的である。すなわち、問題の難度が単体エージェント（Flashはもちろん、より高性能なPro単体でさえも）の独力での解決能力の上限に達したとき、廉価なパートナーを追加しても対話には有効な検証・反証の材料が加わらず、むしろ第4章で確認した「最終発言者問題」のリスクだけが上乗せされる。しかし、両エージェントが共に高性能ティアである場合に限り、片方が見落とした手がかりをもう片方が拾い上げる、あるいは片方の誤った推論にもう片方が実際に対抗しうるだけの検証能力を持ち合わせる可能性が生じ、これが構造的リスクを上回る場面が超難問という極限状況において現れたと考えられる。

5.4 逆転の限界：「解決」ではなく「相対的な被害軽減」

以上の逆転現象を過大評価すべきではない。第一に、ホームズ×ホームズの超難問限定での平均スコアは5.50/20に過ぎず、これは「超難問を解けるようになった」ことを意味しない。あらゆる構成が失敗する問題群の中で、相対的に最も被害が少なかったに過ぎず、実務上は依然として「信頼できない結果」の範疇にある。

第二に、この逆転は難度に対して単調ではない。既存5シナリオのうち中程度に難しい「企業スパイの幻影」（平均8.20/20）・「内部告発者の事故」（平均11.00/20）では、ホームズ×ホームズはそれぞれ2点・2点と、むしろ5パターン中最低クラスのスコアに沈んでいる（表5-1参照）。したがって「難しい問題ほどホームズ×ホームズが有利になる」という単純な単調関係ではなく、あくまで平均6点未満という極端な難度帯に限定された、局所的かつ標本数2という小さな観測に基づく知見である。より大きな超難問シナリオ群での追試による頑健性の検証が今後の課題である。

第三に、本節の分析は既存のtopology_v6_results.jsonの事後的な再解析であり、新規の実験実行を伴わない。この点も含め、第7章で述べる本研究全体の限界の一部として位置づけられるべきである。

第6章　動的ルーティング・トリアージの提唱：単体優位とホームズペア逆転の両立設計

第3章・第4章が示す「単体推論が原則として優位である」という知見と、第5章が示す「超難問という局所的な例外ではホームズ×ホームズが優位に転じる」という知見は、一見すると矛盾するが、両者を統合する設計として、タスクの推定難度に応じて経路を切り替える動的ルーティング・トリアージを提案する。

6.1　発想：常時単体でも常時バディでもない第三の道

第3章の結果が示す通り、常時単体推論（ワトソン型Flash）を採用すれば、10シナリオ平均で14.50/20というスコアを¥1.02／シナリオという最小コストで達成できる。一方、常時ホームズ×ホームズを採用すれば、コストは13倍（¥13.09）に跳ね上がるにもかかわらずスコアはむしろ12.50/20へ低下する。しかし第5章が示した通り、超難問という限定的な部分集合においてのみ、ホームズ×ホームズは単体推論を上回る。したがって理想的には、大多数のタスクには最も安価な単体推論を用い、真に超難問と判定される少数のタスクに限ってホームズ×ホームズへエスカレーションする仕組みが、精度とコストの両面で「常時単体」「常時バディ」いずれをも上回りうる。

6.2　3層トリアージ・アーキテクチャの提案

上記の発想を、次の3層構造として設計提案する。

Tier 0（難度トリアージ）：タスクを実際に解く前に、低コストな難度推定ステップを挟む。この段階の目的は「正解を出す」ことではなく、「単体推論の能力を超える可能性が高いか」を粗く判定することに限定される。

Tier 1（既定ルート＝単体推論）：Tier 0で「通常難度」と判定された大多数のタスクは、最も費用対効果に優れる単体推論（精度重視であればホームズ型Pro単体、コスト重視であればワトソン型Flash単体）で処理する。第3章の知見（単体2構成が5パターン中上位2位を独占）から、これで大半のタスクに対して最良またはそれに近い結果が得られる。

Tier 2（エスカレーション・ルート＝ホームズ×ホームズ再審査）：Tier 0で「超難問の可能性が高い」と判定された少数のタスクに限り、ホームズ×ホームズ（Pro+Pro）によるバディ再審査へエスカレーションする。ここで初めて、第5章で確認した局所的な優位性を活用する。なお、第4章で特定した「最終発言者問題」はこのTier 2でも構造的に残存するため、将来的にはTier 2自体にも収束判定メカニズム（第8章・提言1）を組み込むことが望ましい。

6.3　トリアージ判定基準の候補（今後の実装課題）

Tier 0の難度推定は、正解を知らない実運用時点で行う必要があるため、本研究のように事後的にスコアを見て「超難問」と分類することはできない。次の3種の代替シグナルを、将来の実装候補として提案する。いずれも本研究では実装・検証しておらず、次期実験（v7）での検証を要する設計提案の段階にとどまる。

(a) 自己申告確信度によるトリアージ。単体エージェントに、最終解答と同時に0〜100%の確信度を出力させ、閾値（例えば70%）を下回った場合にのみTier 2へエスカレーションする。

(b)独立反復による自己整合性シグナル。単体エージェントを異なる温度・プロンプト順で2回実行し、結論が一致しない場合を難度シグナルとして扱う。これはPhase 0(tournament_scaling.py)で導入した自己整合性の技法を、「解答を確定させるため」ではなく「難度を検知するため」に転用するものであり、目的が異なる点に注意されたい。

(c)トラップ類型に基づく事前分類。本研究では「前提誤認」型のトラップ(密室の金庫と消えた絵画)が全構成中最低スコアであったことから、シナリオ文中に矛盾した前提や複数の解釈を許す記述が含まれるかを検出する軽量な分類器を、難度推定の補助シグナルとして併用する可能性が考えられる。

6.4 オラクル・ルーティングによるコスト・精度シミュレーション

Tier 0の実装を待たずとも、本研究の既存データを用いて「仮に理想的なトリアージが行われていたら」という上限値を試算できる。具体的には、事後的に判明した超難問2件(密室の金庫と消えた絵画・完璧すぎるアリバイ)にのみホームズ×ホームズの実測結果を、残る8シナリオには単体・ワトソン(Flash)の実測結果を割り当て、10シナリオ全体を再集計した。

表6-1 ルーティング戦略別の精度・コスト比較(10シナリオ)

戦略	平均スコア/20	平均コスト/シナリオ	備考
常時単体(ワトソン・Flash)	14.50	¥1.02	第3章の実測値
常時バディ(ホームズ×ホームズ)	12.50	¥13.09	第3章の実測値
オラクル・トリアージ(超難問2件のみホームズ×ホームズへ)	**14.90**	**¥6.26**	**事後的な難度情報に基づくシミュレーション(上限値)**

オラクル・トリアージは、10シナリオ平均で14.90/20を達成し、常時単体(14.50/20)を0.40点、常時バディ(12.50/20)を2.40点上回った。コスト面では、平均¥6.26／シナリオとなり、常時単体(¥1.02)の約6.2倍である一方、常時バディ(¥13.09)の約48%(52%減)に抑えられている。すなわちオラクル条件下では、精度・コストの双方で常時単体・常時バディの単純な折衷を上回る結果が得られる。

ただし、この試算は「どのシナリオが超難問であるかを事前に知っている」というオラクル(神託)情報に基づく上限値のシミュレーションであり、実際のトリアージ判定器の性能に一切依存しない理想値である点を強調しておく必要がある。実運用でこの上限値にどこまで近づけるかは、6.3節で挙げたような難度推定シグナルの精度(超難問の見逃し＝機会損失、平易な問題の誤エスカレーション＝コスト浪費)にすべて依存する。

6.5 本提言の限界と次の検証ステップ

本章の提言には、少なくとも3点の重要な限界がある。第一に、5.4節でも述べた通り、逆転現象そのものが2シナリオという小標本に基づく観測であり、より大きな超難問シナリオ群での追試が必要である。第二に、6.4節のシミュレーションはオラクル上限値であり、実装可能な難度トリアージ判定器の構築とその精度検証(再現率・適合率)は本研究の範囲外であり、次期実験(v7)における最優先課題として位置づける。第三に、エスカレーション判定の誤り(超難問の見逃し、平易な問題への誤エスカレーション)が実際のコスト・精度に与える影響を定量化するための感度分析も、本稿執筆時点では未実施である。

これらの限界を踏まえ、本章の提言は「動的トリアージが有効であることの実証」ではなく、「動的トリアージが有効たりうる条件と、それを検証するための具体的な次のステップ」を提示するものと位置づけるべきである。

第7章　実験の自己批評：我々はどのような罠を回避したのか

本章では、本研究プロジェクトが方法論的に回避し得た罠と、依然として残存する限界の両方を、意図的に自己批判的な立場から検討する。

7.1　回避できた罠

(a) 評価タスク自体が判別力を持つことの事前確認。 Phase 0・Phase 1で数学タスクの天井効果に直面した際、その結果を「マルチエージェントに効果がない」という誤った結論として即座に採用せず、「評価タスクが構造差を検出する感度を持っていない」可能性を優先して検討し、謎解きベンチマークへの刷新につなげた。これにより、後続の全実験が判別力のある評価軸の上で行われることが保証された。

(b) 採点官モデルの固定による評価バイアスの排除。 v3〜v6の全実験を通じて、解答を生成するモデル・構成がFlash/Pro・単体/バディと変化する一方で、採点官（LLM-as-a-Judge）は常にdeepseek-v4-flashに固定した。これにより、採点側のモデル格の違いが評価結果に混入することを防いでいる。

(c) 構造とプロンプトを完全固定した上でのモデル比較（v5）。 v5のバディ構成マトリクス実験では、2D-Peerの対話構造・プロンプト・ターン順を一切変更せず、A・B各ターンで呼び出すモデルのみを組み替えた。これにより「モデルの組み合わせ効果」を、構造の違いという交絡因子から分離して純粋に検証できている。

(d) 合計点のスクリプト側機械集計。 採点官（LLM）の自己集計に依存せず、4項目の点数をPython側で機械的に加算することで、LLMの算術ミスが最終スコアに混入するリスクを排除した。

(e) 完全放置モードにおける頑健なエラーハンドリングとレジューム機能。 Phase 3では、API呼出の一時的失敗に対する5回までの指数バックオフ・リトライ、認証エラー等の回復不能なエラーの即時検知と警告通知、予算上限（¥1,800）による安全停止、さらに実行中断時にも既存の完了済み結果を自動再利用して再開できるレジューム機能を実装した。実際、本実験は環境要因により2度中断されたが、この設計により既完了分のコスト・時間を無駄にすることなく最後まで完走できた。

7.2　回避し切れなかった限界

(a) 新規シナリオと既存シナリオの難度不均衡。 3.2節・5.1節で示した通り、Phase 3で新規追加した5シナリオの平均スコア（18.88/20）は、既存5シナリオ（7.80/20）の2.4倍に達しており、両者は難度において著しく不均衡である。これは「10シナリオでの頑健な検証」を意図した設計が、実際には「非常に易しい5問＋非常に難しい5問」という二極構成になってしまったことを意味する。全パターンに共通する序列（単体>バディ）自体は10シナリオ全体で一貫しているため定性的な結論の妥当性は保たれるが、v4・v5（既存5シナリオのみ）とv6（10シナリオ）の絶対スコアを単純比較することはできない。第5章で報告した「超難問限定の逆転現象」も、この難度不均衡があったからこそ発見できた知見である一方、その裏返しとして観測数が2シナリオしかないという限界を継承している。今後の版では、シナリオ作成時に予備採点（パイロットスコアリング）を行い、難度を事前に較正する必要がある。

(b) 試行数・温度設定に起因する統計的頑健性の不足。 各パターン・各シナリオの実行は温度0.5での単発実行であり、複数回の反復試行による平均・信頼区間の算出は行っていない。したがって本研究で報告する

スコア差が、真の能力差ではなく単発実行のばらつきに起因する可能性を完全には排除できない。第5章の逆転現象についても同様の留保が適用される。

(c)単一モデルファミリーへの依存。「単体 vs 協調」「同格 vs 異格」の比較はすべてDeepSeek v4ファミリー(Flash/Pro)内で行われており、他の開発元のモデル(例えば異なるアーキテクチャ・異なる学習方針を持つモデル間の協調)に本研究の知見がそのまま一般化できるかは未検証である。

(d)LLM-as-a-Judge自体の妥当性の未検証。本研究の採点は一貫してLLM-as-a-Judgeに依拠しているが、この採点官自身の判断が人間の専門家評価とどの程度一致するかについての独立検証(human alignment check)は行っていない。

(e)DeepSeek APIのピーク/オフピーク料金変動によるコスト測定のノイズ。実行時刻帯によって単価が最大2倍変動するため、実験がどの時間帯にまたがって実行されたかにより、コスト比較の数値には一定の変動が含まれる可能性がある。

(f)第6章のオラクル・トリアージ・シミュレーションが実装検証を伴わない上限値であること。6.5節で述べた通り、第6章で示したコスト・精度の改善幅は、超難問の所在を事後的に知っているという理想条件下でのシミュレーションであり、実運用可能な難度推定器の構築・検証は今後の課題として残されている。

これらの限界を踏まえ、本研究の結論は「マルチエージェント協調は原理的・普遍的に劣る」という強い主張ではなく、「少なくとも本研究が構築した謎解き型の敵対的推論タスク・固定ターン順の2D-Peer構造・DeepSeek v4ファミリーという条件下では、単体推論が一貫してコスト効率・精度の両面で優位であり、その唯一の例外は両エージェントが最高性能ティアで揃った場合の超難問限定の局所的な逆転にとどまる」という条件付きの実証的知見として位置づけるべきである。

第8章　結論と次世代システムへの提言

8.1　結論

本研究は、Phase 0からPhase 3までの4段階にわたる実証実験と、その事後的な深堀り分析を通じて、次の知見を得た。

1. マルチエージェント協調の効果を検証するには、天井効果を起こさない十分な難度・敵対性を持つ評価タスクが不可欠である（Phase 0・1の失敗から得られた前提条件）。

2. 組織構造（トポロジー）が同一であれば、対等な2者構造（2D-Peer）が階層・官僚構造よりも一貫して優位である（Phase 2、13.20 vs 9.20/20）。

3. しかし評価シナリオを5件から10件へ倍増させ、より頑健な検証を行った結果、あらゆるバディ協調構成が単体推論に劣後するという、先行フェーズ（v4）の知見を覆す結果が得られた（Phase 3）。

4. この「マルチエージェント・ペナルティ」の主要因は、固定ターン順の対話構造に内在する「最終発言者問題」であり、対話の質・収束の有無に関わらず、たまたま最後に発言したエージェントの結論が機械的に採用されてしまう構造的な単一障害点にある。

5. 最も安価かつ高速な単体構成（ワトソン型Flash、¥1.02／シナリオ）が、最も高価なバディ構成（ホームズ×ホームズ、¥13.09／シナリオ）を精度でも上回っており、マルチエージェント化への投資が必ずしも精度向上につながらないことが定量的に示された。

6. 超難問（全構成が2.00〜5.80/20に沈む問題）に限定すると、この序列は局所的に逆転し、ホームズ×ホームズ（Pro+Pro、最も高コストなバディ構成）が単体推論を含む全構成中で最高スコアを記録した（第5章）。ただし絶対スコアは依然低く（平均5.50/20）、これは「解決」ではなく「相対的な被害軽減」と位置づけるべきである。

7. 上記の局所的逆転を活用するため、既定ルートを単体推論としつつ超難問と判定されたタスクのみホームズ×ホームズへエスカレーションする動的ルーティング・トリアージを提案した。事後的な難度情報を用いたオラクル条件下のシミュレーションでは、常時単体（14.50/20、¥1.02）・常時バディ（12.50/20、¥13.09）のいずれをも上回る14.90/20を、常時バディの約48%のコスト（¥6.26）で達成できる可能性を示した（第6章）。ただし、これは実装可能な難度トリアージ判定器の構築を前提としない理論上の上限値である。

8.2　次世代システムへの提言

以上の知見から、単にマルチエージェント構成を放棄すべきという結論を導くのではなく、以下の設計改善を次世代システムへの提言として示す。

提言1：固定ターン順の廃止と収束判定メカニズムの導入。「最後に発言した者が勝つ」という現行構造を廃し、全ターンの結論の一致度を機械的に測定し、意見が収束しない場合には単純な「最終発言者の結論採用」ではなく、確信度に基づく重み付け投票や、第三の独立した裁定エージェントによる再審査を挟む設計が必要である。

提言2：反証耐性を持つエージェントの優先。本研究の事例分析では、正しい仮説を最初に提示したエージェントが、後続の（誤った）反論に対して十分な確信を持って反証し切れず、結論を弱める挙動が繰り返し観察された。エージェントに対し、他者の反論を「検証」する指示だけでなく、自らの論拠が客観的証拠（本研究文脈では鞄の中身・タイムスタンプ・物的証拠等）に直接基づく場合には安易に譲歩しないことを促す設計が有効と考えられる。

提言3：協調は「コストに見合う場面」を選んで適用する。本研究の結果は、あらゆる場面でマルチエージェント化が有害であることを意味しない。むしろ、少なくとも敵対的な偽情報・ミスリードを含む単発の謎解き型タスクにおいては、単体推論で十分、あるいはむしろ単体推論の方が優れているという条件付きの知見であり、システム設計者は「協調によって得られる精度向上」と「協調に伴うコスト・レイテンシ・構造的リスクの増大」を、タスクの性質ごとに天秤にかけるべきである。

提言4：評価ベンチマーク自体の難度較正の制度化。7.2節で述べた新旧シナリオの難度不均衡という限界を踏まえ、次期実験ではシナリオ設計時に必ずパイロットスコアリングを行い、既存シナリオ群との難度整合性を事前に確認するプロセスを標準化すべきである。

提言5：超難問限定でのホームズ×ホームズ再審査への条件付きエスカレーション（第6章）。提言3で述べた「コストに見合う場面を選ぶ」という原則を実装可能な形にするため、第6章で詳述した3層トリアージ・アーキテクチャ（Tier 0：難度推定 → Tier 1：既定は単体推論 → Tier 2：超難問と判定された場合のみホームズ×ホームズへエスカレーション）の導入を提言する。ただし本提言の有効性は6.4節のオラクル・シミュレーションに基づく上限値評価であり、実運用可能な難度推定シグナル（6.3節）の構築と、その精度の実証的検証が前提条件となる。

本研究で構築した10シナリオ・5パターンの評価基盤（topology_v6_robust10.py）およびその全ターンログ（topology_v6_results.json）は、上記提言に基づく次世代の協調プロトコル（収束判定付きバディ構造、動的トリアージ・ルーティング等）の効果を同一条件下で再検証するための基準線として、そのまま再利用可能である。

付録：使用した実験スクリプトと出力ファイル一覧

表A-1　使用スクリプトと結果ファイル一覧

フェーズ	スクリプト	結果ファイル
Phase 0	tournament_scaling.py	tournament_scaling_results_deepseek.csv/json
Phase 1	topology_v2.py	topology_v2_results.csv/json
Phase 2a	topology_v3_mystery.py	topology_v3_mystery_results.csv/json
Phase 2b	topology_v4_final.py	topology_v4_final_results.csv/json
Phase 2c	topology_v5_buddies.py	topology_v5_buddies_results.csv/json
Phase 3（決定版）	topology_v6_robust10.py	topology_v6_results.csv/json

推論エンジン：DeepSeek API（deepseek-v4-flash / deepseek-v4-pro）。採点官：deepseek-v4-flash（全実験共通）。為替レート：148円/USD。第5章・第6章で行った超難問限定の再分析およびオラクル・トリアージ・シミュレーションは、いずれもtopology_v6_results.jsonの既存データの事後的な再解析であり、新規のAPI呼出・追加実験は伴わない。