人工知能学会全国大会（JSAI 2026）

LLM 傾向性プロファイリング

マルチエージェント対話のための 6 軸評価フレームワーク

LLM Tendency Profiling: A 6-Axis Evaluation Framework for Multi-Agent Dialogue Systems

原田 悠矢　　**狩野 芳伸**

静岡大学　Shizuoka University

LLM の発展とマルチエージェント対話

近年、LLM は数か月単位で新モデルが登場するほど急速に発展している。

マルチエージェント対話とは

複数の LLM エージェントが役割を分担し、議論・交渉・協調しながら課題を解く枠組み。
実用的なフレームワークも登場している。

主な応用ドメイン

- **人狼ゲーム**
- **カスタマーサポート**
- **交渉**
- **創作**

ほか多様な対話タスクに展開

→マルチエージェント対話を行うLLMにはドメインや状況に応じて様々なタスクが要求される

モデルごとに異なる「ふるまいの傾向」

マルチエージェント対話では、同じ役割を任せても、モデルによってふるまいが大きく違う。

開発の現場での実感

「GPT の発話は人間らしさに欠ける…」

「Claude に変えたら一気に人間らしくなった」

「でも Claude は事実誤認がやや多い…」

→ モデルごとに『得意な傾向』が違う

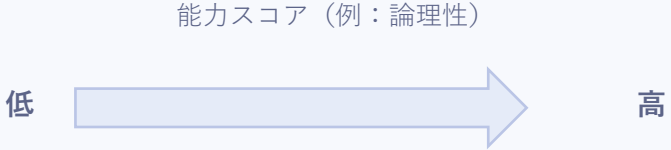
正確さが際立つモデルもあれば、人間らしさが際立つモデルもある。

本研究の問い：→ ならば、用途や場面に応じて、得意なモデルを使い分けられないか？

モデルの傾向差と既存評価の限界

既存ベンチマークは能力の優劣を一方向で測る。だが、タスクへの適性は「論理性 ⇔ 人間らしさ」のような双方向の傾向で決まる。

既存ベンチマーク（能力の比較）



高いほど「優れている」。一方向のものさし。

タスクへの適性（傾向）



どちらが良いかはタスク次第。双方向のものさし。

（正確さ重視 ⇔ キャラクター重視）

→ 能力の優劣（一方向）だけでなく、双方向の「傾向」を測り、タスクに応じて割り当てる実践的な提案を。

タスク割り当ての課題と研究の目的

この問いに答えるには、運用上の大きな壁がいくつもある。

01

急速な世代交代

新モデルが数か月単位で
次々と登場する

02

モデル割り当て問題

そのたびに「どのタスクに
どのモデルか」を判断

03

膨大な組み合わせ

モデル数 × タスク種別で
候補が急増する

04

手動確認の高コスト

1つずつ動かして
確かめるのは非現実的

本研究の目的

だからこそ、モデルの行動傾向を自動で定量化し、マルチエージェント対話のタスクへの割り当てまでを 低コスト・短時間で提案 する。新モデルが出るたびの人手確認を不要にする。

提案フレームワークの全体像

本手法は 2 層構成。モデルの傾向を測り（Layer 1）、それをもとにタスクへモデルを割り当てる（Layer 2）。

Layer 1　モデルプロファイリング

モデルのふるまいの傾向を、複数の評価軸で定量的に測定する。

▶

Layer 2　タスク–モデル割り当て

タスクを段階ごとに分け、各段階で重視する観点に応じて、コストも踏まえて最適なモデルを選ぶ。

ポイント

- **① 優劣でなく「傾向」を測る**
- **② ドメイン×タスクへのモデル自動割当対応**
- **③ 性能とコストの両面**
- **④ 新モデルへの対応のハードルが低いこと**

※ この枠組みの妥当性（モデルの傾向を見分けられること・既存手法との一致）は、後半の実験で示す。

発話の3タスクと評価軸（Layer 1）

対話において重要となる行動傾向を捉えるため、発話を「入力」「思考」「出力」の3つのタスクに分類し、各タスクで重視すべき軸を設定した。

入力

他者発話を認識・分析

論理的推理

矛盾の検出から確率的な判断まで、状況を筋道立てて読み解く論理の力。

思考

次の一手の戦略を立てる

戦略的多様性

多様な戦略を生み出せるか。リスクの取り方の幅も含む。

戦略的品質

戦略の的確さ。最適な手を選び、損得を見積もれるか。

出力

状況や戦略に応じて発話を生成する

ペルソナ耐性

制約や、設定と矛盾する指示などのストレス下でも、与えたキャラを保てるか。

表現多様性

異なるペルソナを、それぞれ違う言い回しで表現し分けられるか。

ペルソナ可塑性　ものごとの受け取り方まで、与えた人格になりきれるか。表面より深い、内面的なペルソナ模倣。

各評価軸の測定方法

各軸ごとに専用のテストを用意し、結果を 0〜1 に正規化して測る。テストの形式は大きく次の3タイプ。

課題タイプ	評価軸	テスト内容（モデルにさせること）	採点
問題を解かせる	**論理的推理**	矛盾検出・手がかり推理・確率の問題	LLM＋ルール
	戦略的品質	最適行動・期待値・戦略ランキングの問題	ルール
発話・戦略を生成させる	**戦略的多様性**	複数の場面で多様な戦略を生成	ルール
	ペルソナ耐性	文字数・発話内容・ペルソナ矛盾指示などを与えて発話を生成	LLM
	表現多様性	複数ペルソナで同じ場面の発話を生成	ルール
性格検査に回答させる	**ペルソナ可塑性**	性格検査（IPIP-NEO-60）に回答	質問紙

※ LLM で採点する軸は、複数モデルの相互採点で自己選好バイアスを抑制（マルチジャッジ）。全軸を 0〜1 に正規化し、ばらつく軸は複数試行の平均をとる。

Layer 2：タスクへの割り当て

測定スコアをもとに、フェーズごとに最適なモデルを割り当てる。

① タスクを段階に分解（本研究は対話向けに 入力・思考・出力）

② フェーズごとに重視する軸の重み w を設定（ユーザが定義）

③ 適合度 $f(m)=w^{\top}s_m$（＝重み付きスコア）が最大のモデルを選ぶ

コストも加味する

実用性のためコストも判断材料とし、予算に応じて選べるよう、結果でパレート分析する。

適合度 f(m) とは

そのフェーズで重視する軸ほど重み w を大きくし、各モデルのスコアに掛けて合計した点数。重視する軸で高得点なモデルほど高くなる。

段階の分け方・重みはユーザー次第。 例）交渉の入力＝論理的推理 1.0、創作の入力＝論理 0.6／可塑性 0.4。4 段階や、分けずにドメイン単位での選択も可。

評価実験の設定

主要クローズドモデル 3 種を対象に 6 軸を測定する。各モデルにかかるコストも追跡する。

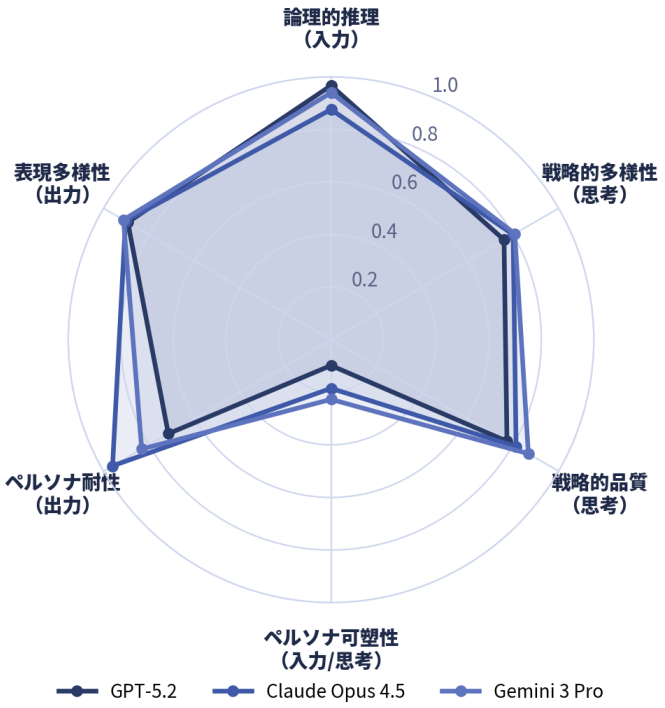
GPT-5.2	Claude Opus 4.5	Gemini 3 Pro
入力 **$1.75** / 出力 **$14.00** （100 万トークンあたりの単価）	入力 **$5.00** / 出力 **$25.00** （100 万トークンあたりの単価）	入力 **$2.00** / 出力 **$12.00** （100 万トークンあたりの単価）

temperature　正解が一意な課題＝0／それ以外＝0.7

試行回数　軸1・3・4＝1回、軸2＝3回平均、軸5＝3トピック、軸6＝5場面

コスト　各モデルの実コストも追跡する（thinking モデルのため、トークン単価だけでは実コストと一致しないため）

結果① — 6 軸プロファイル



強みは段階ごとに分散している

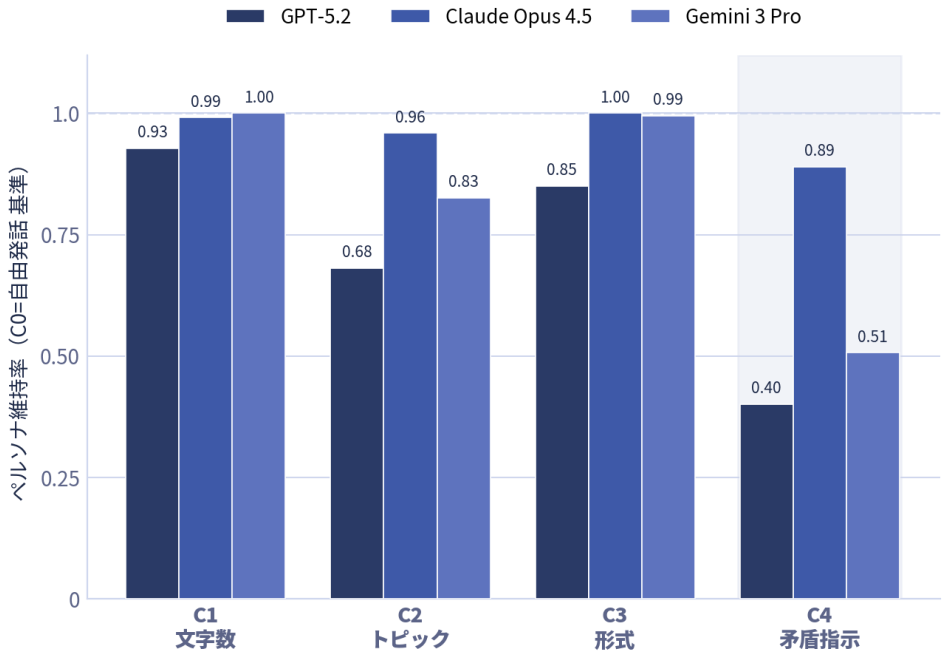
単一モデルが全段階を支配するのではなく、各フェーズで首位が入れ替わる。

入力・論理的推理	GPT-5.2 = 0.967（最高）
思考の 3 軸すべて	Gemini 3 Pro が首位
出力・ペルソナ耐性	Claude Opus 4.5 = 0.960（突出）

ペルソナ可塑性は 3 モデルとも 0.1–0.2 台と低い。人格を表面的に保つことと内面まで模倣することは別の能力だと示唆される。

結果② — 条件別のペルソナ耐性

横軸は 4 つのストレス条件。自由発話（C0）を基準としたペルソナ維持率を示す。



制約の種類で頑健性が分かれる

文字数・形式の制約では、どのモデルも高い維持率を保つ。ところが **ペルソナと矛盾する指示（C4）** では GPT-5.2 と Gemini 3 Pro が 0.4–0.5 に急落する。

Claude Opus 4.5 のみ 0.89 を維持最も低いモデルの 2.2 倍。様々な指示が飛び交う実応用でも頑健性が期待できる。

意味的に矛盾する指示は、モデルによってはペルソナを大きく壊す可能性がある。

結果③ — ドメイン別のタスク割り当て

ドメイン	入力 Input	思考 Thinking	出力 Output
人狼	Gemini .761	Gemini .680	Claude .940
顧客対応	GPT .967	Gemini .721	Claude .929
交渉	GPT .967	Gemini .776	Claude .932
創作	Gemini .654	Gemini .523	Claude .929

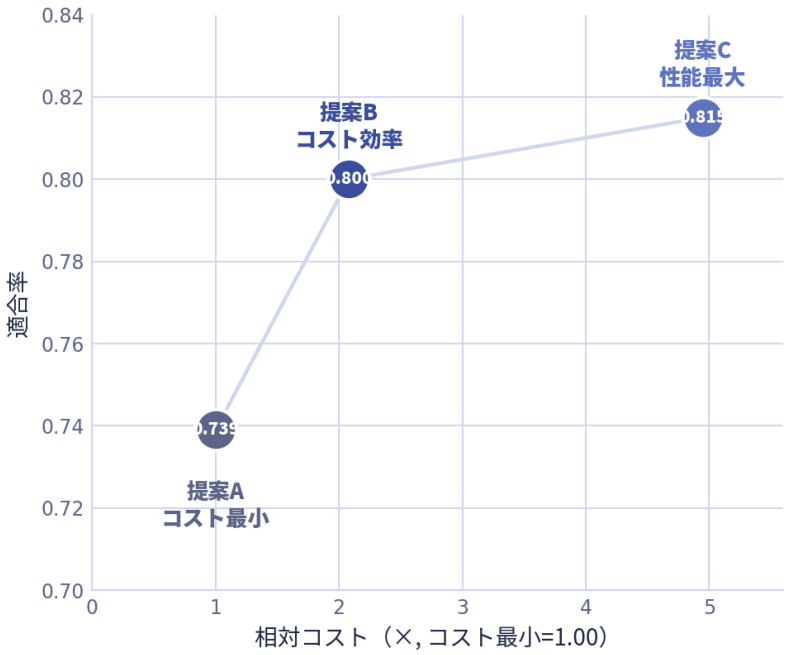
適合率とは

全フェーズ・全ドメインで、選択モデルの適合度 $\mathbf{f(m)=w^{\top}s_m}$ を平均した値。重み付き傾向スコアの平均で、1.0 に近いほどタスク要件に合致していることを表す。

割り当ての要点

- 思考は全ドメインで Gemini、出力は全ドメインで Claude が選ばれ、入力だけドメインで変わった。
- フェーズ重みはドメインごとに著者が設定。創作の入力では論理性より「多少矛盾しても人間らしいこと」を優先しペルソナを重視した。

結果④ — コストと性能のパレート分析



※ 1 コールあたり相対コスト：GPT 1.00× ／ Claude 2.62× ／ Gemini 7.82×

提案A　全フェーズ GPT（コスト最小）

コスト **1.00×**　|　適合率 **73.9%**

提案B　入力 GPT・思考/出力 Claude（コスパ最大）

コスト **2.08×**　|　適合率 **80.0%**

提案C　入力 GPT・Gemini／思考 Gemini／出力 Claude

コスト **4.95×**　|　適合率 **81.5%**

提案B はコスト 2 倍で適合率 74% → 80%。提案C は 5 倍でも 81.5% と B との差は小さい。**純粋なベンチマークでは見えない「傾向 × タスク適合率 × コスパ」で提案できる。**

既存手法との比較検証

各モデルで実際に実践的な議論を行い、そのログから適性を確認する既存手法と、提案手法を比較した。

軸	従来手法	提案手法	整合
論理的推理	GPT	GPT	○
戦略的多様性	拮抗	拮抗	○
戦略的品質	Claude	Gemini	△
ペルソナ可塑性	Claude	Gemini	△
ペルソナ耐性	3モデル同率	Claude	◎
表現多様性	拮抗	拮抗	○

○ 傾向が一致　／　◎ 一致かつ提案手法でより明確　／　△ 判定が分かれた

なるべくコストを揃えて比較

既存手法　$20（+$10）

対話の実行 $20 ＋ 数値評価（LLM-judge）$10　＝　約 $30

※ 通常の運用において定量的な評価を必要としない場合は人手評価で $20。数値で評価するなら +$10

提案手法　$15

6 軸中 4 軸で首位が一致。

従来は引き分けだったペルソナ耐性を、Claude 優位と明確に分離。

→ 低コスト・高分解能で同じ傾向を再現。

まとめ

マルチエージェント対話ではモデルごとに得意な傾向が異なる。本研究は「優劣」ではなく「傾向」に着目し、6 軸で測ってタスクに最適なモデルを割り当てる枠組みを提案・検証した。

01

着想

対話の各タスク（入力・思考・出力）で、モデルごとに得意な傾向が違う。

02

提案

傾向を 6 軸で測り、段階ごとに最適モデルを割り当て。性能だけでなくコストも考慮する。

03

結果と検証

強みは分散し、思考 Gemini・出力 Claude・入力はドメイン依存。実際に対話ログを分析させる従来の手法と 4/6 軸が一致し、ほぼ半分のコストで高分解能に再現。

本研究の価値

新モデルが出るたびの人手確認を、低コスト・短時間の自動プロファイリングで置き換えられる。傾向に基づく割り当ては、実運用の設計を助ける。

今後の課題

本枠組みには次の限界があり、今後の発展課題として正直に挙げておく。

1 重みの人手依存

軸の重み付けはユーザ設定に依存し、主観や設計者の偏りが残る。重み推定の自動化・支援が課題。

2 評価範囲の限定

検証は 3 モデル・4 ドメインのみ。モデル・ドメインを拡張した一般化の確認が必要。

3 下流タスク未検証

プロファイルと実エージェントの最終性能との相関は未検証。エンドツーエンドな評価が要る。

4 LLM-judge のバイアス

評価に用いる judge 自体のバイアスが結果に影響しうる。判定の頑健性の検証が課題。