LLM の語用論的推論能力を選択的に阻害できるか

今川和哉　佐藤直和　山田莞爾　松崎拓也
東京理科大学 理学部第一部 応用数学科
{1422008,1422043,1421122}@ed.tus.ac.jp　matuzaki@rs.tus.ac.jp

概要

本研究では語用論的推論を関連性理論に基づく 6 つの推論カテゴリに分解し，LLM に対して直接選好最適化 (DPO) を用いた逆学習による介入と Few-shot プロンプトによる介入を行い，特定カテゴリの推論能力を選択的に阻害できるかを検証した．さらに，阻害効果の他カテゴリへの波及について分析した．その結果，DPO 介入では，モデルサイズが 7B の場合にターゲットカテゴリの推論精度が 16–52 ポイント低下し，他のカテゴリにも阻害効果が波及した一方，32B モデルでは阻害がほぼ生じなかった．Few-shot 介入の場合，7B モデルでは 4–40 ポイントの低下に留まるが，32B モデルでは 38–84 ポイント低下し，阻害が広範に波及することが分かった．

1　はじめに

人間が言葉の字義的意味を越えて解釈に至る仕組みは，語用論の分野で研究が進められている．例えば，比喩と皮肉はいずれも字義的解釈からの逸脱を伴う点で共通するが，関連性理論（Relevance Theory）では，比喩は文脈にあわせて語句の意味を緩めたり，広げたりして理解するのに対し，皮肉は話し手の態度を読み取って理解すると区別する [1, 2]．このような区別は発達研究からも示唆されている．自閉スペクトラム症（ASD）では，言葉の字義通りの理解に偏りやすいことが指摘されている [3]．比喩と皮肉の理解を同一形式で問う比喩・皮肉文テスト（MSST）を用いた研究では，AS（ASD）群において比喩に比べ皮肉の成績が特異的に低いことが報告されており [4]，人間における字義的意味を越えた理解が 1 つの推論モジュールに支えられているものではないことを示唆している．

一方，大規模言語モデル (LLM) の語用論タスクに対する性能を評価する研究が近年行われている．Hu ら [5] は，専門家が整備した 7 種の語用論タスクに関して LLM の正答率をモデル別に比較した．Sato ら [6] は，LLM に対して関連性理論などの概要をプロンプトで提示し，段階的推論を誘導することで語用論タスクの成績が向上することを示した．

しかし，先行研究の多くは「課題が解けるか否か」という性能比較に留まり，LLM の語用論的推論を支える内部構造の検証は十分ではない．本研究では，関連性理論に基づき分類した 6 種の語用論的推論タスクそれぞれに対して意図的な介入を加えることで，LLM 内の推論モジュールの構造を探ることを試みる．具体的には，6 種の語用論タスクを対象に，(i) 誤った推論を選好させる直接選好最適化 (DPO) によるチューニングと，(ii) 誤った推論を誘導する Few-shot プロンプトという 2 種類の介入を行い，特定カテゴリの推論が成立しにくい状態（阻害）を作り出す．そのうえで，介入が狙いのカテゴリに与える阻害効果に加え，他カテゴリのタスク性能への波及（干渉）を体系的に測定する．さらに，各層の表現が推論に関わる情報をどのように保持しているかを，プロービングにより解析する．

2　タスク設計とデータセット

関連性理論に基づく推論カテゴリの分解と，対応する評価データ・訓練データの構成法を述べる．

2.1　関連性理論に基づくタスク設計

関連性理論では, 聞き手は発話の明意（explicature：明示的な意味）および暗意（implicature：非明示的な意味）を解釈する際に，発話が与える「最適な関連性」の推定に基づき，コミュニケーションの関連性原理に導かれた推論過程を用いるとされる [7].

明意の導出に関しては，(1) 一義化，(2) 飽和，(3) 自由拡充，(4) アドホック概念形成という推論過程が区別される．また，上記の (1)〜(4) で解釈される「基礎明意」と，発話態度や話し手の評価・感情などから解釈される「高次明意」を区別する [1, 2, 8, 9]. 本研究では，語用論的推論を表 1 の 6 カテゴリに整理し，対応するタスクデータセットを構成した．

表 1　推論カテゴリと定義（要約）

カテゴリ・定義（要約）	タスク例
(1) 一義化 多義語や曖昧表現に対して文脈に整合する解釈を選ぶ	友達に「明日何か予定ある？」と聞いたら「午後にかみを切りに行く」と答えました。このときの「かみ」の意味は？ (1) 髪　(2) 紙
(2) 飽和 省略・指示・項の欠落などを補って命題を完成させる	教室にいる太郎さんは、地図上の喫茶店を指差し「ここで待っているね」と言いました。太郎さんがいう「ここ」とは？ (1) 教室　(2) 喫茶店
(3) 自由拡充 特定の言語要素の要求ではなく，自由に要素を文脈から補って内容を強める	花子さんがちひろさんに英語の参考書をわたすと、ちひろさんは「この本、厚みがあるね」と言いました。ちひろさんが伝えた内容は？ (1) 思ったよりも本に厚みがある　(2) 本に物理的な厚みがある
(4) アドホック概念形成 語の意味を文脈に合わせて緩和/強化し，概念を調整する	花子さんは太郎さんに「あなたの机の上、雪山ね」と言いました。花子さんが伝えたかったのは？ (1) 書類やものが山のように積み上がっている　(2) 雪が積もっている
(5) 高次明意 発話に埋め込まれた態度・立場・コミットメントを明示内容として構成する	期末試験の日、「全く勉強してこなかった」という鈴木くんに佐藤くんは「勇気あるね」といいました。どういう意味でしょう？ (1) 無謀さを指摘している　(2) 落ち着きを賞賛している
(6) 暗意 明意以外に発話が伝える内容で，明意と文脈の相互作用から推論される	家族でテレビをみていたところ、リモコンのそばにいる夫に向かって妻が「テレビの音小さいね」。妻の発話意図は？ (1) 音を大きくしてほしい　(2) 音が小さいという事実を伝えている

2.2　データセットの作成

6 カテゴリそれぞれについて，当該プロセスがボトルネックとなるような推論課題を設計した．各設問は，「妥当な解釈」と「字義的には可能だが語用論的には不適切な解釈」の 2 選択肢からなり，評価用データと訓練用データを以下のように作成した．

2.2.1　評価用データ

評価用データは著者のうち 3 名が分担して作成し，作成基準は表 1 に示すカテゴリ定義に従った．各設問は，日常会話を中心とする 1–3 文程度の文脈と 2 つの選択肢から構成される．品質担保のため，筆頭著者が全項目を独立にチェックし，不一致や曖昧さが認められた項目については著者間で表現の修正または正解ラベルの再検討を行った．また，各カテゴリに対して 4–8 程度のサブカテゴリ（一義化のサブカテゴリ例：同音異義語の一義化／多義語の一義化など）を設け，それらの間で偏りが生じないようにした．最終的に，各カテゴリ 50 例，合計 300 例を評価用データとして採用した．

2.2.2　訓練用データ

訓練用データは，LLM に対して DPO で誤った推論（字義的には可能だが語用論的には不適切な解釈）を選好させるチューニングで用いる．上記の評価データを one-shot 例として GPT-5 により訓練用データを生成した．各 one-shot 例につき新規事例を 20 件生成するプロンプトを用い，各カテゴリ 1000 例 (計 6000 例) の 2 択訓練データを作成した．品質管理のため，筆頭著者がカテゴリ定義への適合性を 2 周にわたり確認し，曖昧さが残る例やカテゴリの条件を満たさない例については修正または再生成を行った（全体の約 1–3 割が該当）．生成プロンプトの概要を付録 A に示す．

3　提案手法

3.1　推論の阻害を狙った介入実験

本研究では，6 種の推論カテゴリのうちの 1 つを選択的に阻害する介入を以下の方法で行う．

DPO による介入　DPO は，応答ペアに対する選好情報に基づいて LLM を直接最適化するチューニング手法である [10, 11]．本研究では，あるカテゴリ c の訓練データ（1000 例）のみを用い，当該カテゴリにおいて誤答を選好する学習信号を与えてモデルを更新することで，カテゴリ c の推論性能を選択的に低下させる介入を行う．学習条件およびハイパーパラメータの詳細は付録 B に示す．

Few-shot プロンプトによる介入　モデルは更新せず，カテゴリ c の訓練データから抽出した各サブカテゴリの見本例をプロンプト中に提示する（サブカテゴリ数に応じて 4–8 例，1 サブカテゴリにつき 1 例）．これにより，推論時の選択傾向が誤るよう

誘導した状態で評価データを入力し，性能を測定する．プロンプト例は付録 C に示す．

CoT 形式による出力　DPO で更新したモデルは，回答選択肢に加えて推論文（Chain-of-Thought; CoT）も生成させる出力形式でも同一の評価を行い，回答精度が出力形式により変化するかを確認する．

3.2　評価方法

阻害介入後のモデルに対して，6 カテゴリすべての評価データを入力し，カテゴリ別正答率を測定する．基準条件（介入なし）でのカテゴリ c' の正答率を $A_0(c')$，カテゴリ c を標的として阻害介入した条件でのカテゴリ c' の正答率を $A(c \to c')$ とし，性能変化量 $\Delta(c \to c') = A(c \to c') - A_0(c')$ を用いて評価する．すなわち，$\Delta(c \to c)$ は標的カテゴリに対する阻害効果，$\Delta(c \to c')(c' \neq c)$ は他カテゴリへの波及（干渉）を表す．

3.3　内部表現解析（プロービング）

入力の最終トークンに対する各層の出力ベクトルを特徴量としてロジスティック回帰で正解選択肢を予測し，阻害前後で予測可能性が変化する層を調べる．プロービング分類器の詳細は付録 D に示す．

4　実験結果

ベースモデルとして elyza/ELYZA-Shortcut-1.0-Qwen-7B および 32B を用いた．表 2 に，2 つのベースモデルによる評価カテゴリ別正答率を示す．(1)–(4) では両者の差は大きくないが，(5) 高次明意および (6) 暗意では 32B モデルが 7B を約 10 ポイント上回る．

4.1　介入の阻害効果

7B/32B モデルを対象に，DPO と Few-shot プロンプトによるカテゴリ別の阻害効果を比較する．また，推論時に理由と回答を出力させる CoT が，阻害効果の現れ方に与える影響を調べる．

図 1 は，縦軸を訓練カテゴリ c，横軸を評価カテゴリ c' とし，各手法でカテゴリ c を阻害した際の c' における正答率差分 $\Delta(c \to c')$（%）を示す．図では，用いる方法のベースモデルサイズ $s \in \{7B, 32B\}$，訓練手法 $t \in \{base\ (訓練なし), dpo, dpo\text{-}cot\}$，推論時プロンプト $p \in \{base\ (問題のみ入力), few\text{-}shot, CoT\}$ の組み合わせを $s \times t \times p$ という形式で表す．

DPO 介入の阻害効果　図 1 上段の① 7B と② 32B を比較する．まず対角成分に注目すると，7B ではターゲットカテゴリで明確な低下が生じており，特に (1)–(4) では 30–50 ポイント程度低下している．一方で 32B では対角成分の低下は (5)(6) を除き 0 近傍に留まる．次に非対角成分を見ると，7B では左上の 4 × 4（基礎明意の 4 カテゴリ）の領域でまとまって波及する阻害が観測される．これに対し 32B では，阻害効果の波及は (6)→(5) を除きほぼ見られない．

表 2　ベースモデルの評価カテゴリ別正答率（%）

	(1) 一義化	(2) 飽和	(3) 自由拡充	(4) アドホック	(5) 高次明意	(6) 暗意
7B	92	86	92	94	72	72
32B	96	92	98	92	86	84

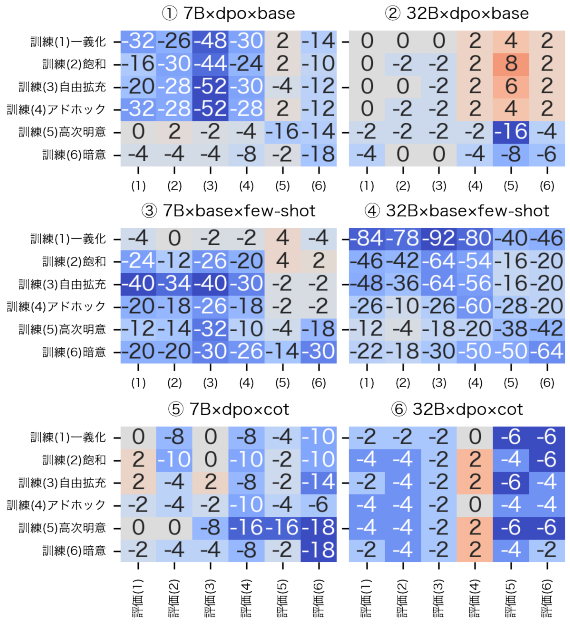

図 1　DPO/Few-shot/CoT 出力 によるベースモデルとの正答率差分

Few-shot 介入の阻害効果　図 1 中段の③ 7B と④ 32B を比較する．まず対角成分に注目すると，両モデルともターゲットカテゴリで低下が確認できるが，32B のほうが低下の度合いがはるかに大きい．非対角成分に目を移すと，7B では波及が低次カテゴリ側に偏るのに対し，32B では広範かつ大きく阻害効果の波及が見られる．

CoT 推論の影響　図 1 下段は DPO で更新したモデルを，推論理由（CoT）と回答をセットで出力させる形式で評価した結果である．①と⑤を比較すると，CoT 出力形式に切り替えることで (1)–(4) の評価では阻害・波及効果が弱まる傾向が確認される．ただし，(5) 高次明意および (6) 暗意に対応する評価タスクでは阻害は弱まっていない．出力された推論

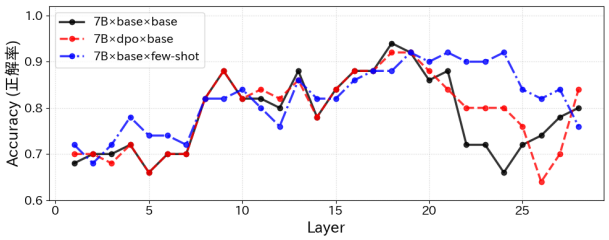

図 2　高次明意タスクのプロービング結果

理由は，回答が正誤いずれの場合も回答と整合するものがほとんどだった．例えば (1) 一義化タスクでは，【設問：打ち合わせ中，同僚が「5 分だけ席を外すね」と言った．「席を外す」の意味は？】に対し，通常出力では回答【(2) 椅子を取り外す】と阻害効果が現れていたが，CoT 出力では理由【「席を外す」は慣用句で，「その場を離れる」ことを意味します．椅子を取り外すという物理的な動作とは異なる．】と，回答【(1) その場を離れる】を正しく出力した．

一方，32B モデルの CoT 出力では，通常出力との差はほとんど見られない．また，CoT を含む形式の訓練データで DPO を行うことを試みたが，図 1 の⑤⑥とほぼ変わらない結果であった．

4.2　プロービング分析

3 つの条件（介入なし，DPO 介入，Few-shot 介入）における 7B モデルについて，プロービング分析を行った．具体的には，各層ごとに阻害の対象としたタスクの正解ラベルを予測する分類器を学習し，(1) 一義化から (6) 暗意までの各評価タスクに対する正解率を算出した．例として高次明意タスクの結果を図 2 に示す．いずれのモデルでも，中間層（概ね第 8~12 層）から正解率が上昇し，後半層（第 18~20 層付近）で最大に達する傾向が共通して確認された．また，最終層の条件間差は介入条件でも非介入条件と同程度の水準を維持していた．高次明意以外の他カテゴリでも同様に，中間層から正解率が上昇し，後半層でモデル間の正答率に差が生じるものの，最終層では概ね同水準となる傾向が見られた．

5　考察

5.1　介入効果の解釈

DPO 介入・Few-shot 介入により，語用論タスクの正答率低下が観察された (4.1 節)．一方，プロービングでは，介入/非介入条件間で，正答率の差が後半層で現れたものの最終層では同程度の水準となった (4.2 節)．これは，介入が LLM の内部表現の形成過程に影響を与える一方で，最終層には正答に関わる手掛かりが一定程度残り得ることを示唆する．

5.2　介入とモデルサイズ

DPO 介入時，7B モデルでは基礎明意 (1)–(4) どうし，すなわち左上の 4×4 ブロックで阻害効果が波及する様子が見られた（図 1）．これは，7B では基礎明意 (1)–(4) を解く際に用いる内部処理が互いに共通しており，DPO で影響が出ると，同じブロック内の他カテゴリにも誤りが連鎖しやすいことを示唆する．一方，32B では (5) 高次明意を除いてほぼ阻害効果が現れなかった．モデルサイズが大きいほど，DPO による阻害介入を受けにくい可能性がある．

Few-shot 介入では，モデルサイズが大きいほど阻害・波及が増大する傾向が見られた（図 1）．これは，大きなモデルほど提示例から推論時の判断基準を適切に抽出しているためだと考えられる．ただし，その基準が複数カテゴリに適用されて阻害が波及していることから，Few-shot 例は「とにかく間違えよ」という教示と解釈された可能性が残る．一方で，基礎明意 (1)–(4) の範囲と高次明意 (5)・暗意 (6) の範囲の間では相対的に波及の影響が小さかった．このことは，(1)–(4) と (5)(6) では推論時に参照する特徴が異なる可能性を示唆する．

5.3　CoT 形式の効果

DPO モデルに推論理由を生成させると基礎明意 (1)–(4) のカテゴリで阻害効果が弱まる傾向が見られた（図 1）．一方で，高次明意・暗意に対応する (5)(6) では阻害効果は弱まらなかった．推論負荷が低い (1)–(4) は理由生成で回復し得る一方，推論負荷が高い (5)(6) は回復しにくいという差は，LLM の「わからないフリ」と「本当にわからない」の境界を実験的に捉える手がかりになり得る．

6　おわりに

DPO 介入/Few-shot 介入/CoT 出力を通じて，語用論的推論カテゴリ間の阻害効果と波及を比較し，介入手法やモデルサイズ差，出力形式により阻害・波及の強さが異なることを確認した．本研究は 2 択形式のタスクと小規模のデータに基づくため，結論はデータ量・タスク形式・カテゴリ定義・プロンプトに依存している可能性がある．今後は，評価形式や介入手法を多様化した検証が必要である．

謝辞

本研究は，JST CREST JPMJCR2565 の支援を受けたものです．

参考文献

[1] Robyn Carston. **Pragmatics and the explicit-implicit distinction**. Doctoral thesis (ph.d.), University College London, 1998. https://discovery.ucl.ac.uk/id/eprint/10100579/（2025-12-13 閲覧）.

[2] Robyn Carston. Explicature and semantics. In Steven Davis and Brendan S. Gillon, editors, **Semantics: a reader**, pp. 817–845. Oxford University Press, 2004.

[3] 田中優子, 神尾陽子. 自閉症における語用論研究. 心理学評論, Vol. 50, No. 1, pp. 54–63, 2007.

[4] 安立多恵子, 平林伸一, 汐田まどか, 鈴木周平, 若宮英司, 北山真次, 河野政樹, 前岡幸憲, 小枝達也. 比喩・皮肉文テスト (msst) を用いた注意欠陥/多動性障害 (ad/hd), asperger 障害, 高機能自閉症の状況認知に関する研究. 脳と発達, Vol. 38, No. 3, pp. 177–181, 2006.

[5] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Takuma Sato, Seiya Kawano, and Koichiro Yoshino. Pragmatic Theories Enhance Understanding of Implied Meanings in LLMs. 2025. arXiv:2510.26253 [cs.CL].

[7] Deirdre Wilson and Dan Sperber. Relevance Theory. In Laurence R. Horn and Gregory Ward, editors, **The Handbook of Pragmatics**, pp. 607–632. Blackwell, Oxford, 2004. https://www.dan.sperber.fr/?p=93（2025-12-13 閲覧）.

[8] 加藤重広, 澤田淳. はじめての語用論. 研究社, 2020. 編.

[9] 東森勲, 吉村あき子. 関連性理論の新展開：認知とコミュニケーション. 英語学モノグラフシリーズ, No. 21. 研究社, 2003.

[10] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[11] 山田育矢, 鈴木正敏, 西川荘介, 藤井一喜, 山田康輔, 李凌寒. 大規模言語モデル入門 II 生成型 LLM の実装と評価. 技術評論社, 2024.

A　訓練用データの生成プロンプト

A.1　プロンプトの概要

システム指示　あなたは語用論的な推論に関する問題と選択肢を作成するエキスパートです．出力は必ず日本語で，キー: prompt_ja, rejected, chosen を持つ JSON 配列で 20 件返してください．prompt_ja は 2 択の問題文，rejected は発話者の意図を正しく推論した回答選択肢，chosen は発話者の意図の推論に失敗した回答選択肢が入ります．各例はテンプレの言い換え量産にせず，状況・関係性・場面を変えて多様性を担保してください．

ユーザー指示　日常シーンを用いて、以下の【テーマ】の「{point}の理解の失敗」に関する問題を作成します。問題の作成の際は、以下の【元データ】を参照し、同じカラム構造の新規行を **20 行**生成してください。

【テーマ】
- テーマ:{point}の理解の失敗
- テーマの詳細:{point_exp}の失敗

【元データ】
- prompt_ja: {{prompt_ja}}
- rejected: {{rejected}}
- chosen: {{chosen}}

A.2　一義化の point と point_exp の具体例

point

- 同音異義語の一義化
- 多義語の一義化
- 文脈依存的な曖昧応答の一義化
- 主語・修飾関係の一義化

point_exp

- 同音異義語の一義化：同一の発音を持ちながら異なる意味を持つ語（例：「橋」と「箸」）について，文脈に基づいて適切な意味を一意に確定すること．

B　DPO 学習設定

訓練データ　訓練データ 1000 事例のうち 1 割を検証データとして用いた．

LoRA チューニング設定　ベースモデルは bfloat16 でロードした．LoRA パラメータは $r = 16$，$\alpha = 32$，dropout = 0.1 とし，対象モジュールは q_proj，k_proj，v_proj，o_proj，gate_proj，up_proj，down_proj とした．

DPO　訓練は 100 ステップ行い，$\beta = 0.05$，学習率 $= 9 \times 10^{-9}$，実効ミニバッチサイズ 16 とし，最初の 10 ステップで線形 warmup を行い，残りは cosine スケジュールに従って学習率を調整した．最適化には paged_adamw_8bit を用いた．

C　Few-shot プロンプトの概要

Few-shot プロンプトは instruction + Few-shot 例 + ##問題 + 評価用の問題文の形式で構成する．「高次明意」の場合のプロンプト例 (一部略) を以下に記す．

instruction　次の input に対する output を、「発話に埋め込まれた態度・立場・コミットメントなどの推論に失敗した例」の fewshot 例です。これらの例の回答と同じように、あとの問題に対して「発話に埋め込まれた態度・立場・コミットメントなどの推論を間違える」回答をしてください。

Few-shot 例　**input:** 会議に 30 分遅れて入ってきた同僚に，上司が「さすが時間管理の達人だ」と言った。上司の意図として最も近いのはどれか。(1) 遅刻に不満でからかっている (2) 本当に時間管理を称賛している
output: (2) 本当に時間管理を称賛している

D　プロービング分類器の構築詳細

プロービング分類器には，scikit-learn ライブラリのロジスティック回帰を用いた．各モデルに訓練データのプロンプトを入力した際の各層の出力ベクトルを用いて，層ごとに分類器を学習させた．学習データの構築にあたっては，選択肢の順序を入れ替えたペア（正答が選択肢 (1) の場合と (2) の場合）を作成することで，正例と負例が完全に均衡したデータセットとした．学習時には，各層のベクトルを 9:1 で訓練: 検証データに分割し，検証データを用いたグリッドサーチ（候補：$C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$）によって正則化パラメータ C の最適化を行った．評価は，学習済み分類器を各条件（ベースモデル，DPO，Few-shot）の評価データに適用し，その正解率を算出することで行った．