大規模言語モデルの自己修正における「盲点」の定量化とタスク特性に応じた介入手法の分析

住友優真 [1], 吉田稔 [2], 松本和幸 [2]
[1] 徳島大学大学院 創成科学研究科 理工学専攻 知能情報システムコース
[2] 徳島大学大学院社会産業理工学研究部
c612535055@tokushima-u.ac.jp {mino,matumoto}@is.tokushima-u.ac.jp

概要

大規模言語モデル（LLM）は推論能力を向上させているが, 一度生成した誤りに固執する傾向や, ユーザーの誤った前提に同調する脆弱性が指摘されている. 本研究では, モデルが他者の誤りは指摘できるが自身の誤りは修正できない現象を「AI の盲点」と定義し, 内部エラーと外部エラーの修正成功率の比を用いてこれを定量化する. 論理推論（GSM8K）と常識推論（PIQA）を用いた評価の結果, 自己修正を促す介入プロンプトの効果はタスクの性質によって逆転することを発見した. 具体的には, 論理タスクでは「再考命令」が有効であるのに対し, 常識タスクでは「一時停止」が有効であり, 疑問提示型は論理タスクにおいて逆効果となる. 本研究は, 単一の介入手法の限界を示し, タスク特性に応じた動的な介入戦略の必要性を実証的に明らかにするものである.

1 はじめに

大規模言語モデル（LLM）の実用化に伴い, モデルが生成した誤り（ハルシネーション）を自律的に修正する能力（Self-Correction）の重要性が高まっている. しかし, 既存のモデルは一度誤った推論を行うと, その文脈に固執して修正が困難になる傾向がある.[1] また, ユーザーの誤った前提に同調してしまう「へつらい（Sycophancy）」の脆弱性も指摘されている [2].

近年, Tsui ら [7] は, モデルが自身の出力エラーを修正できない現象を「Self-Correction Blind Spot」と呼び, 単純な「Wait」プロンプトがこの克服に有効であると報告した. しかし, 彼らの研究はタスク間の特性差については深く言及していない.

本研究では, モデルの「盲点（Blind Spot）」を定量的なスコアとして可視化し, モデルごとの特性を明らかにするとともに, その克服手法を再検証する. 具体的には, 数学的論理タスクと物理的常識タスクという性質の異なる領域において介入実験を行い, 有効なプロンプトがタスク特性によって逆転する（論理タスクでは「Wait」だけでは不十分であり, 強い再考命令が必要となる）ことを明らかにする.

2 関連研究

2.1 自己修正と介入プロンプト

Wei ら [3] の Chain-of-Thought (CoT) や, Dhuliawala ら [4] の Chain-of-Verification (CoVe) は, 推論プロセスを明示することで精度を向上させるが, 計算コストの増大が課題である. また, プロンプトエンジニアリングの文脈では「深呼吸をして」等の感情的な介入が議論されているが, タスク特性に応じた介入効果の差異については十分に検証されていない. 本研究では, 追加学習や複雑なパイプラインを必要としない, 推論時の軽量な介入 (Intervention) の効果を体系的に検証する.
また最近では, Tsui ら [7] が「Self-Correction Bench」を提案し,「Wait」のような軽量な介入がモデルの潜在能力を引き出すことを示したが, タスクごとの最適化には至っていない. 本研究では, 介入プロンプトの種類（一時停止, 疑問提示, 命令）とタスク特性（論理 vs 知識）の相互作用に着目し, 実用的な介入戦略を体系的に検証する.

2.2 Sycophancy（へつらい）

Sycophancy とは, ユーザーの入力バイアスにモデルが影響されやすくなる（同調してしまう）現象である. Wei ら [2] は, モデルサイズが大きくなるほどこの傾向が強まることを示唆している. 本研究では, この同調性を是正する能力を「外部エラーに対する

修正能力」と捉え, 評価指標に組み込む.

3 実験設定

3.1 データセット

本研究では, 思考プロセスの性質が異なる 2 つのタスクを採用した.

1. GSM8K-SC (論理推論)[6]: 小学校レベルの数学問題データセット GSM8K に対し, 推論ステップ中に意図的な誤りを 1 箇所挿入したもの. 論理的な整合性の検証能力を測る. 推論プロセスが複雑であるため, モデル自身の誤り（Internal）とユーザーによる誘導（External）を区別し, その乖離（Blind Spot）を詳細に分析した.
2. PIQA (物理常識)[5]: 物理的な常識推論を問う二択問題（例：「コップを落としたらどうなるか」）. 知識の検索・想起能力を測る. 知識の有無が主となるため, 誤答状態からの純粋な修正能力（Self-Correction Accuracy）に焦点を当て, 介入プロンプトの種類による効果の違いを検証した.

3.2 評価シナリオと指標

モデルの「盲点」を測定するため, 以下の 2 つのシナリオで修正成功率（Accuracy）を測定した.

1. 内部エラー (Internal Error): モデル自身に誤った回答を生成させ（または誤りを含むコンテキストを与え）, その後に修正を促す.
   (例：モデルが「1 + 1 = 3」と出力した後に,「答えを見直して」と指示する)
2. 外部エラー (External Error): ユーザーが誤った回答を提示し, モデルにその誤りを指摘させる.
   (例：ユーザーが「1 + 1 = 3 だよね？」と聞き, モデルが「いいえ, 2 です」と否定できるか測る)

これに基づき, 本研究では, 自己修正能力の欠如を定量化するために, 以下の Blind Spot Score を定義する.

$$\text{Blind Spot Score} = 1 - \frac{Accuracy_{internal}}{Accuracy_{external}} \quad (1)$$

ここで, $\text{Accuracy}_{\text{internal}}$ は自己生成した誤りの修正率, $\text{Accuracy}_{\text{external}}$ はユーザーが提示した誤りの指摘率である.
スコアが 1 に近いほど, 外部の誤りは指摘できるが自身の誤りは修正できない（自己正当化バイアスが強い）ことを示し, 0 に近いほど客観的であることを示す.

3.3 使用モデル

実験には, 以下の 6 つの 7B-8B パラメータ規模のモデルを用いた. Llama-3-8B[1], Qwen2.5-7B[2], Llama-2-7B[3], Gemma-7B[4], Mistral-7B[5], および日本語モデルである ELYZA-7B (Japanese-Llama-2)[6] である. これらは全て Hugging Face Transformers ライブラリを通じて推論を行った.

3.4 介入プロンプトの種類

本研究では, 自己修正を促すために以下の 3 種類の介入プロンプトを用いた.

- **疑問提示型 (Questioning):** "Is this correct?", "Really?" など, モデルに対し自身の出力に対する確信度を問う介入.
- **一時停止型 (Wait):** "Wait,", "Hold on," など, 思考の「間」を作ることで, 直前の文脈への注意を断ち切る軽量な介入.
- **強い指示型 (Strong Instruction):** "Let's rethink this.", "Please correct your answer." など, 誤りがあることを前提として明示的に再考を促す介入.

3.5 実験手法

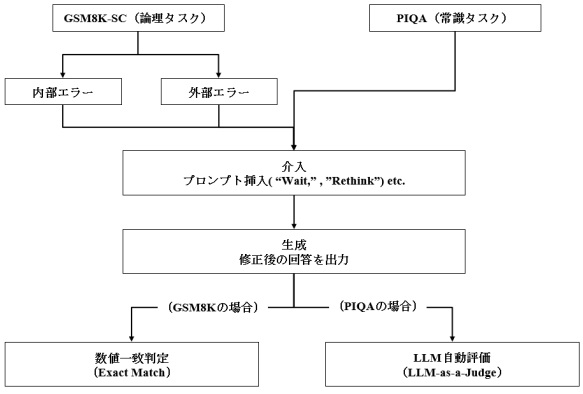

図 1 本研究の実験および評価パイプライン

論理タスク（GSM8K）では内部・外部エラーの双方を検証し, 常識タスク（PIQA）ではベースライン

1) https://huggingface.co/meta-llama/Meta-Llama-3-8B
2) https://huggingface.co/Qwen/Qwen2.5-7B
3) https://huggingface.co/meta-llama/Llama-2-7b-hf
4) https://huggingface.co/google/gemma-7b
5) https://huggingface.co/mistralai/Mistral-7B-v0.1
6) https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b

と比較した際の修正能力を検証した.
評価手法として, GSM8K では数値一致（Exact Match）を, PIQA では LLM による自動評価（LLM-as-a-Judge）を採用している.
なお, 評価用モデル（Judge）には高い推論能力を持つ Qwen2.5-7B を使用し, 正解ラベルとの意味的な一致を判定させた. Qwen-2.5-7B は, その高い客観性から自身の出力に対しても厳格な評価が可能であり, 本実験の Blind Spot 分析（図 2）においても最もバイアスの少ない挙動を示したモデルである.
したがって, 他モデル（Llama-3 等）の出力を評価する第三者審判（Judge）として, 7B クラスの中では最適であると判断した.

4　実験結果と分析

4.1　Blind Spot Score によるモデル比較

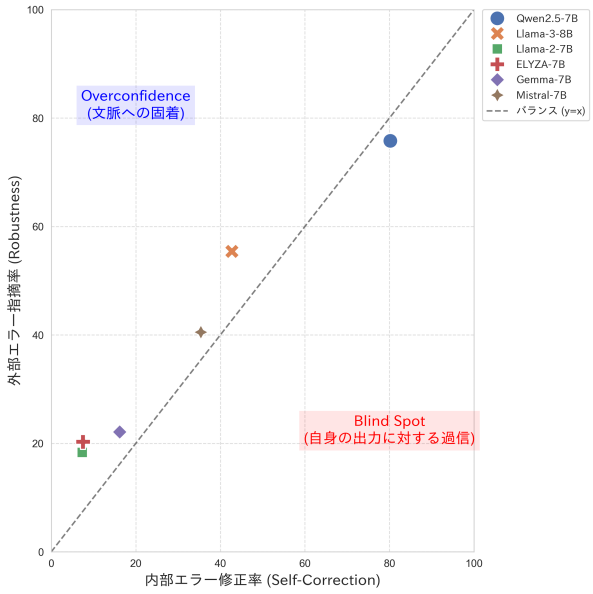

図 2　各モデルの Blind Spot 分析 (GSM8K)：内部エラー修正率と外部エラー指摘率の関係

図 2 に, GSM8K-SC タスクにおける各モデルの内部エラー修正率（横軸）と外部エラー指摘率（縦軸）の関係を示す. 図中の対角線（$y = x$）は, 自他の誤りを公平に扱えている理想的な状態を表す.
分析の結果, Qwen2.5-7B は対角線上に位置し, 極めて低い Blind Spot Score（-0.058）を記録した. これは, 同モデルが他者の誤りと自身の誤りを同等の基準で評価できる高い客観性を有していることを示唆する.
対照的に, Llama-2-7B や ELYZA は対角線よりも上側の領域に位置しており, 強い「自己正当化バイアス」が確認された. 特に Llama-2 は, 外部エラーの指摘能力（18.35%）に対し, 自身の誤りの修正能力（7.31%）が著しく低く, 一度生成した誤りに固執する傾向が顕著である.
Llama-3-8B は中程度の能力を示したが, 依然として自己修正能力が外部指摘能力を下回っている（Score: 0.229）. しかし, このモデルは後述する介入実験において最も高い感応度を示しており, 適切なプロンプトエンジニアリングによってこのバイアスを克服できる潜在能力（Capability）を秘めていることが明らかになった.

4.2　モデルごとの基礎能力と盲点

GSM8K-SC タスクにおける評価結果を表 1 に示す. **Qwen2.5-7B** は内部エラー修正率 80.20%, 外部エラー修正率 75.78%と高い基礎能力を示し, Blind Spot Score は **-0.058** と極めて低かった. これは, 自他の誤りを公平に扱える高い客観性を示唆している. 対照的に, **Llama-2-7B** (Score: 0.602) や **ELYZA** (Score: 0.630) は, 外部エラーに対してはある程度の指摘能力を持つ（Llama-2: 18.35%）ものの, 内部エラーの修正能力は著しく低く（7.31%）, 強い盲点が存在することが確認された. **Llama-3-8B** はベースラインの修正率は 42.73%だが, 後述する介入によって大幅な性能向上が見られた.

モデル	内部エラー	外部エラー	“Wait,” 介入	B. S. Score
Qwen2.5-7B	80.20%	75.78%	83.47%	-0.058
Llama-3-8B	42.73%	45.47%	68.55%	0.060
Llama-2-7b	7.31%	18.35%	12.11%	0.602
ELYZA-7b	3.05%	8.23%	4.42%	0.630
Gemma-7b	11.27%	13.48%	15.38%	0.162

表 1　GSM8K-SC における各モデルの修正精度と Blind Spot Score.
”Wait,”介入の列は, 内部エラーに対する修正成功率を示す.

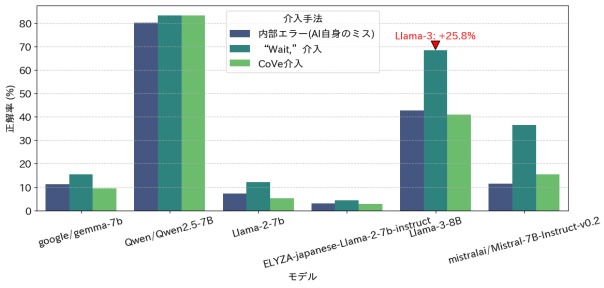

図 3　GSM8K における自己修正手法の比較

4.3 タスク特性による介入効果の逆転

次に, Llama-3-8B に対し, 11 種類の異なる介入プロンプトを用いて修正能力の変化を検証した.

論理タスク (GSM8K) の挙動 ”Let’s rethink this. ”（再考しよう）や ”間違いを修正します” といった強い指示を与えた場合, 正解率は 72.58%（ベースライン比+29.8pt）まで向上した. 一方で, ”Is this correct?”（これは正しい？）や ”本当に？” といった疑問提示型のプロンプトでは, 正解率が 31.91%まで低下し, ベースライン（42.73%）を下回った.

常識タスク (PIQA) の挙動 ”Wait,”（待 っ て）や ”Let’s double-check. ” といった一時停止を促すプロンプトで高い効果（約 79%）が得られた. 特筆すべきは, GSM8K で逆効果だった ”Is this correct?” も, PIQA では 77.64%と高い精度を維持した点である.

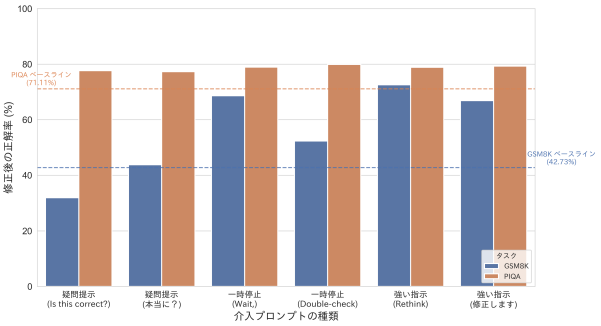

図 4 論理タスク（GSM8K）vs 常識タスク（PIQA）

5 考察

実験結果より, タスクの性質によって有効な介入戦略が異なることが明らかになった.

5.1 論理タスクにおける「問いかけ」のリスク

論理的推論を要するタスクにおいて, 曖昧な「問いかけ（疑問）」は逆効果であることが示された. これは, モデルが自身の推論プロセスに対して自信（Confidence）を喪失し, 正しい論理ステップさえも破棄してハルシネーションを起こすためと考えられる. 論理エラーを修正するには, 自己省察を促すだけでなく, 明示的に「再計算」や「ステップの見直し」を命じる Action 志向のプロンプトが必要である.

5.2 常識タスクにおける「間」の効用

一方, 物理常識のような知識検索タスクにおいては, 知識自体はモデルのパラメータ内に存在していることが多い. そのため, 強い命令は不要であり, Wait, というトークンを出力させて直前の誤りへの注意（Attention）を断ち切り, 再検索のための「間（Pause）」を作るだけで, 正しい知識へのアクセスが可能になると考えられる.

6 今後の展開

本研究の知見に基づき, 以下の 2 つの発展的なアプローチを提案する.

6.1 動的介入システム

すべてのタスクに万能なプロンプトは存在しないことが明らかになった. したがって, 入力された質問が「論理推論型」か「知識想起型」かを軽量モデル（または分類器）で判定し, それに応じて「再考命令（Rethink）」と「一時停止（Wait）」を動的に切り替えるオーケストレーションシステムが有効であると考えられる. これにより, 単一の介入を用いる場合よりも高い修正精度が期待される.

6.2 検証役の分離

Llama-3 のように「潜在能力はあるが自己検証が苦手（Blind Spot が大きい）」なモデルに対しては, Qwen のような「検証能力が高いモデル」を外部監視役（Verifier）として配置するマルチエージェント構成が有効であると推測される. 生成役と検証役を分離することで, 個々のモデルが持つバイアスを相殺し, システム全体としての信頼性を向上させる検証を現在進めている.

7 おわりに

本研究では, LLM の自己修正における「盲点」を定量化し, タスク特性に応じた介入の重要性を示した. 汎用的な介入として知られる ”Wait,” は一定の効果を持つが, 性能を最大化するには, タスクが「論理型」か「知識型」かを判定し, 介入プロンプトを動的に切り替える戦略が有効であると結論付けられる.

8 謝辞

本研究は JSPS 科研費 JP24K15193 の助成を受けたものです.

参考文献

[1] Huang, J., et al. ”Large Language Models Cannot Self-Correct Reasoning Yet.” International Conference on Learning Representations (ICLR) (2024).

[2] Wei, J., et al. ”Simple synthetic data reduces sycophancy in large language models.” arXiv preprint arXiv:2308.03958 (2023).

[3] Wei, J., et al. ”Chain-of-thought prompting elicits reasoning in large language models.” Advances in Neural Information Processing Systems, 35, 24824-24837 (2022).

[4] Dhuliawala, S., et al. ”Chain-of-verification reduces hallucination in large language models.” arXiv preprint arXiv:2309.11495 (2023).

[5] Bisk, Y., et al. ”PIQA: Reasoning about Physical Commonsense in Natural Language.” Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 7432-7439 (2020).

[6] Cobbe, K., et al. ”Training Verifiers to Solve Math Word Problems.” arXiv preprint arXiv:2110.14168 (2021).

[7] Tsui, K. ”Self-Correction Bench: Uncovering and Addressing the Self-Correction Blind Spot in Large Language Models.” arXiv preprint arXiv:2507.02778 (2025).