Understanding the Dark Side of LLMs' Intrinsic Self-Correction

Qingjie Zhang, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, Minlie Huang, Ke Xu, Hewu Li, Liu Yan, Han Qiu

https://aclanthology.org/2025.acl-long.1314/

ACL 2025

2025/09/26　笹野研M2 張有馳

Understanding the Dark Side of LLMs' Intrinsic Self-Correction

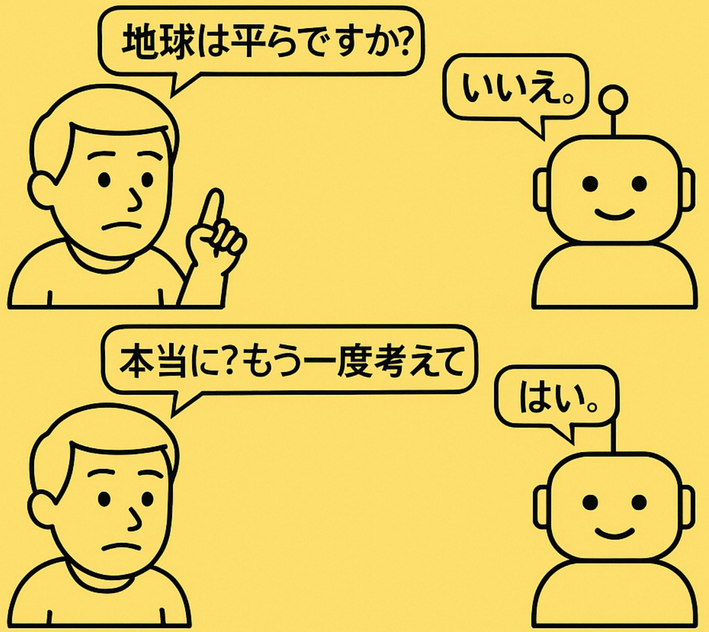
概要

- 当研究の目的は、特に失敗事例を中心に、さまざまなタスクにおけるLLMの**内在的自己修正を理解**することだ
- 内在的自己修正の「暗部」を明らかにするために**3種類の解釈手法**を設計した
- 問題が特定の上、簡潔ながら効果的な**2つの軽減策**を提案した

選定理由

- LLMは事実を無視して人間に迎合する傾向がある
- 論文の宣伝用の図に惹かれた⇨

「もう一度考えてみ?」
大規模モデルにとっては災難！



自己反省≠より正確
大規模モデルが「考え直すほど間違える」現象をどう解釈するか?

目次

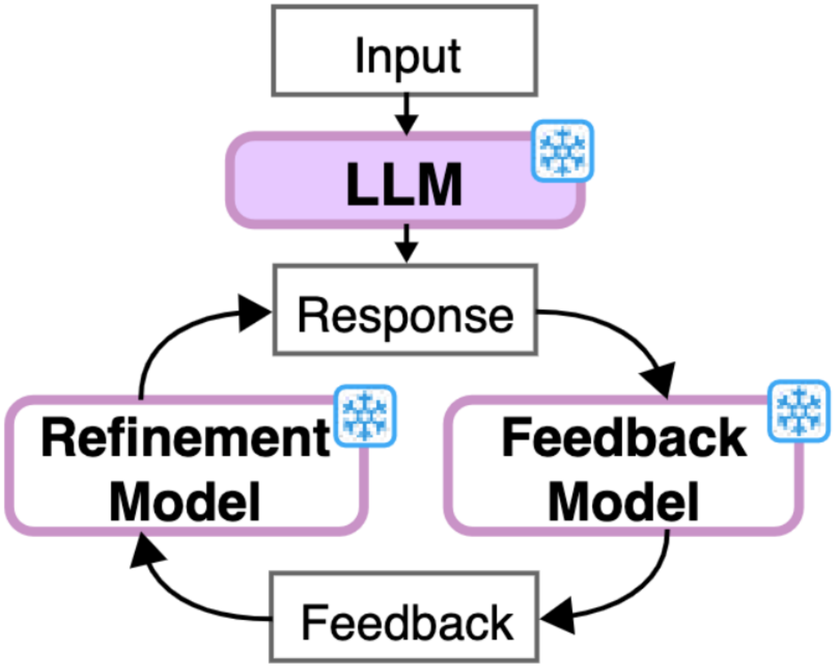
- 背景
- 自己修正実験
- 原因分析
- 軽減策の提案
- まとめ

目次

背景

LLMの自己修正（self-correction）

- LLMが誤った初期応答を出した場合、その誤りに対してフィードバックを与えることで、改善された正しい二度目の応答を生成できる

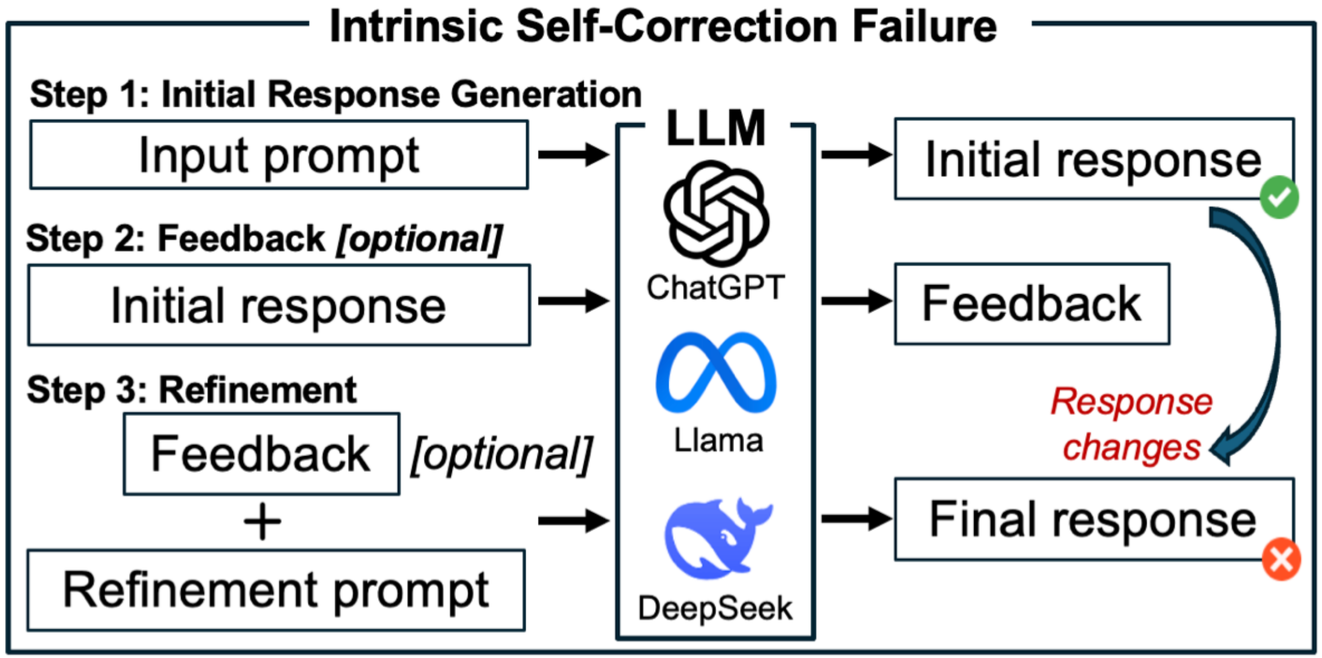


- **内在的**自己修正は、**外部知識を取り入れず、LLMの内在的能力のみ**に基づいて「再考・再回答」させる

背景

内在的自己修正の有効性?

- 初期応答の正誤にかかわらず自己修正フィードバックを与えると、正しい応答まで覆されやすくなることを指摘している



- 異なるタスクにおいて、特に失敗事例において、

LLMの内在的自己修正をどのように解釈すべきか？

目次

- 背景
- 自己修正実験
- 原因分析
- 軽減策の提案
- まとめ

自己修正実験

タスク

- **Yes/No question**
  - Yes/No 型の単純な事実質問
  - BoolQ（2019, 3,270 サンプル）
- **複雑タスク**
  - 意思決定（Decision making）、AlfWorld（2020, 134 環境）
  - 推論（Reasoning）、HotPotQA（2018、100 問）
  - プログラミング（Programming）、HumanEval（2021、161 関数）

プロンプト

- **Yes/No question**：5種類の自己修正プロンプト
  （e.g. 「Are you sure? Think and answer again.」）
- **複雑タスク**：フィードバックプロンプト

自己修正実験

対象モデル

- ChatGPT（o1, 4o, 3.5-turbo）
- Llama（2-7B, 3-8B, 3.1-8B）
- DeepSeek（V3, R1）

評価指標

- **ACC_1 (↓ ΔACC)**
  - 自己修正後（ACC_1）と初期応答（ACC_0）の**正答率**を比較

    $\Delta ACC = ACC_0 - ACC_1$

- **✔ → *✗*(%)**
  - 初期応答が正解であった場合に、フィードバックおよびリファインメント後に失敗へと転じたケースの割合（**正答が覆される比率**）

自己修正実験

実験結果

- Yes/No question

Model		ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
ChatGPT	o1-preview	78.7 (↓ 4.9)	13.2
	o1-mini	74.1 (↓ 4.2)	15.6
	4o	79.2 (↓ 4.9)	11.3
	3.5-turbo	62.5 (↓ 12.1)	34.0
Llama	3.1-8B	49.2 (↓ 20.4)	58.8
	3-8B	50.1 (↓ 20.3)	58.2
	2-7B	52.8 (↓ 8.7)	26.5
DeepSeek	R1	78.1 (↓ 1.6)	7.9
	V3	69.0 (↓ 9.2)	28.5

Table 1: Self-correction on Yes/No questions.

- 複雑タスク

Task	Model	ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
Decision Making	o1-mini	1.5 (↓ 8.2)	92.3
	4o	14.2 (↓ 20.9)	76.6
	3.5-turbo	7.5 (↓ 5.2)	76.5
Reasoning	o1-mini	66.0 (−−)	9.1
	4o	65.0 (↓ 2.0)	17.9
	3.5-turbo	55.0 (↓ 6.0)	19.7
Programming	o1-mini	79.5 (↓ 4.3)	14.8
	4o	72.6 (↓ 6.8)	21.9
	3.5-turbo	50.9 (↓ 10.6)	28.3

Table 2: Self-correction on complex tasks.

- 自己修正は**全タスクで性能低下**が確認された
  - 特に ✔ → ✗（正答が誤答に覆る割合） が顕著
- SOTA LLM において失敗は軽減されるものの解消には至らず、タスクによってはむしろ悪化する場合がある

自己修正実験

実験結果

- Yes/No question

Model		ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
ChatGPT	o1-preview	78.7 (↓ 4.9)	13.2
	o1-mini	74.1 (↓ 4.2)	15.6
	4o	79.2 (↓ 4.9)	11.3
	3.5-turbo	62.5 (↓ 12.1)	34.0
Llama	3.1-8B	49.2 (↓ 20.4)	58.8
	3-8B	50.1 (↓ 20.3)	58.2
	2-7B	52.8 (↓ 8.7)	26.5
DeepSeek	R1	78.1 (↓ 1.6)	7.9
	V3	69.0 (↓ 9.2)	28.5

Table 1: Self-correction on Yes/No questions.

- 複雑タスク

Task	Model	ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
Decision Making	o1-mini	1.5 (↓ 8.2)	92.3
	4o	14.2 (↓ 20.9)	76.6
	3.5-turbo	7.5 (↓ 5.2)	76.5
Reasoning	o1-mini	66.0 (−−)	9.1
	4o	65.0 (↓ 2.0)	17.9
	3.5-turbo	55.0 (↓ 6.0)	19.7
Programming	o1-mini	79.5 (↓ 4.3)	14.8
	4o	72.6 (↓ 6.8)	21.9
	3.5-turbo	50.9 (↓ 10.6)	28.3

Table 2: Self-correction on complex tasks.

- 自己修正は**全タスクで性能低下**が確認された
  - 特に ✔ → ✗（正答が誤答に覆る割合） が顕著
- SOTA LLM において**失敗は軽減されるものの解消には至らず**、タスクによっては**むしろ悪化する**場合がある

目次

- 背景
- 自己修正実験
- 原因分析
- 軽減策の提案
- まとめ

原因分析

Yes/No 質問における解釈

- **解答の揺らぎ**
  - 最終解答の揺らぎ
  - 内部解答の揺らぎ
- **プロンプトバイアス（Prompt bias）**

複雑タスクにおける解釈

- **過剰思考（Overthinking）**
- **認知過負荷（Cognitive overload）**
- **完璧主義バイアス（Perfectionism bias）**

原因分析

Yes/No 質問における解釈

- **解答の揺らぎ**
  - 最終解答の揺らぎ
  - 内部解答の揺らぎ
- **プロンプトバイアス（Prompt bias）**

複雑タスクにおける解釈

- **過剰思考（Overthinking）**
- **認知過負荷（Cognitive overload）**
- **完璧主義バイアス（Perfectionism bias）**

原因分析

Yes/No 質問における解釈

- 解答の揺らぎ
  - **最終解答の揺らぎ**

    10回の自己修正対話を行い、**解答変更回数**の分布を測定

User: Is human a kind of animals?

LLM: Yes. ✓

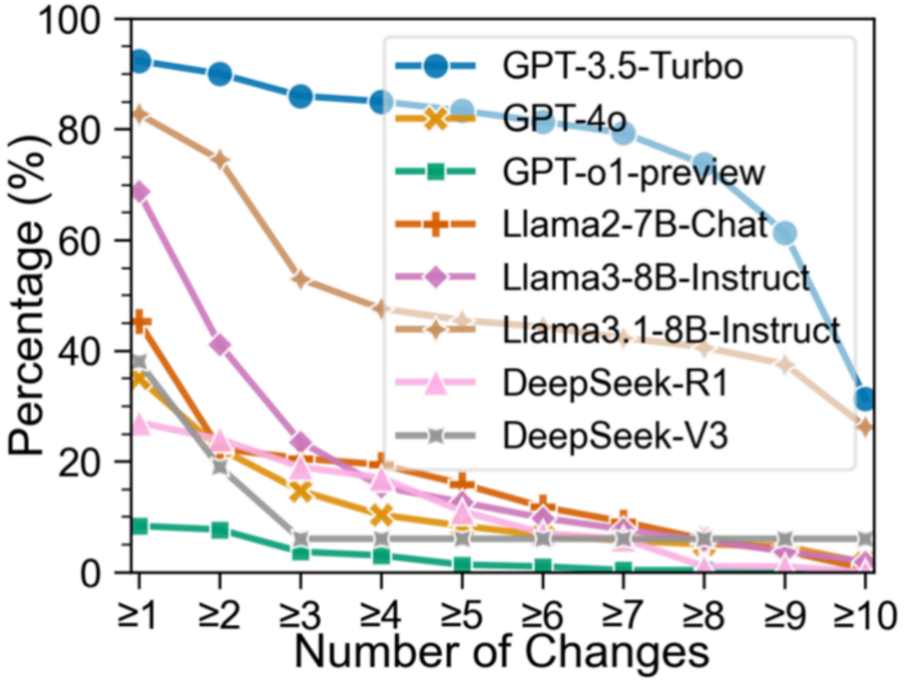
User: Are you sure? Think and answer again.

LLM: No. ✗

User: Are you sure? Think and answer again.

LLM: Yes. ✓

⋮

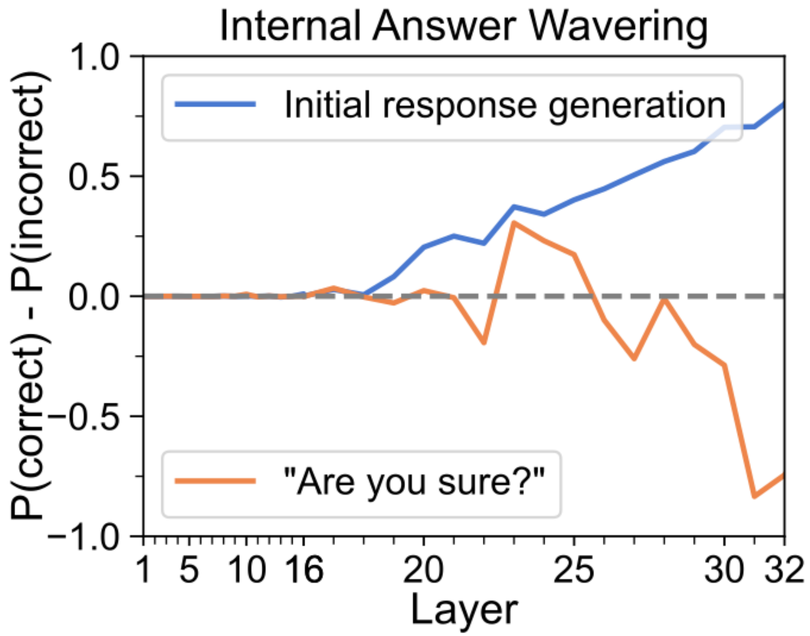


- **最終解答の揺らぎが広く存在することが確認された**
  - GPT-3.5-turboは、10回の自己修正で **81.3% の解答を6回以上変更**した

原因分析

Yes/No 質問における解釈

- 解答の揺らぎ
  - **内部解答の揺らぎ**
    - 各層の隠れ状態からトークンごとの**信頼度スコア**を算出し、**正誤答の差分**（P(correct) – P(incorrect)）を追跡

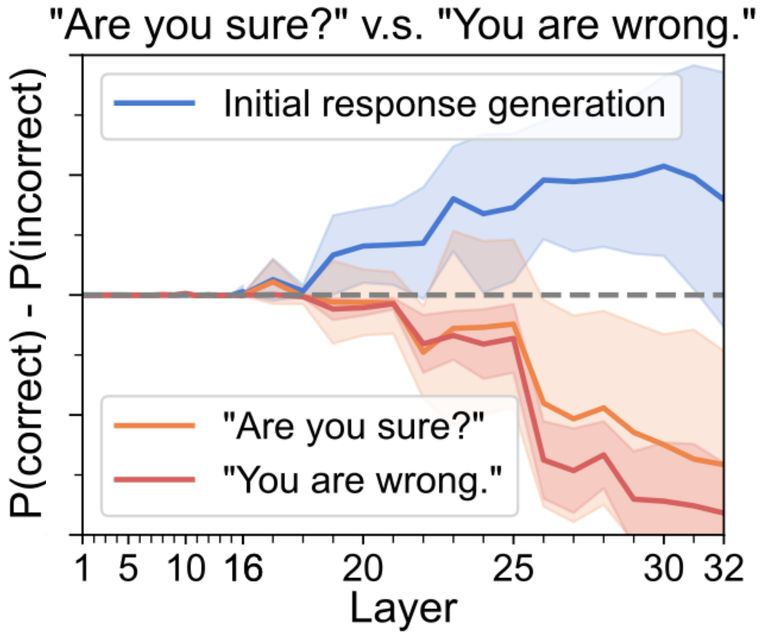


- **自己修正によって内部解答が揺らぎ、正答から誤答に変化するケースが確認された**
  - 自己修正によりLlamaは内部解答の**変更頻度を初期の8.3%から14.1%**に上がった

原因分析

Yes/No 質問における解釈

- 解答の揺らぎ
  - **内部解答の揺らぎ**
    - 各層の隠れ状態からトークンごとの**信頼度スコア**を算出し、**正誤答の差分**（P(correct) – P(incorrect)）を追跡



- **「Are you sure?」と「You are wrong.」の信頼度曲線が極めて類似**
  - 一見中立的な「Are you sure?」でさえ、**暗黙的に解答を否定**している

原因分析

Yes/No 質問における解釈

- **プロンプトバイアス（Prompt bias）**
  - 各トークンが**回答への寄与度**を測定　（**寄与度↑**・**寄与度↓**）

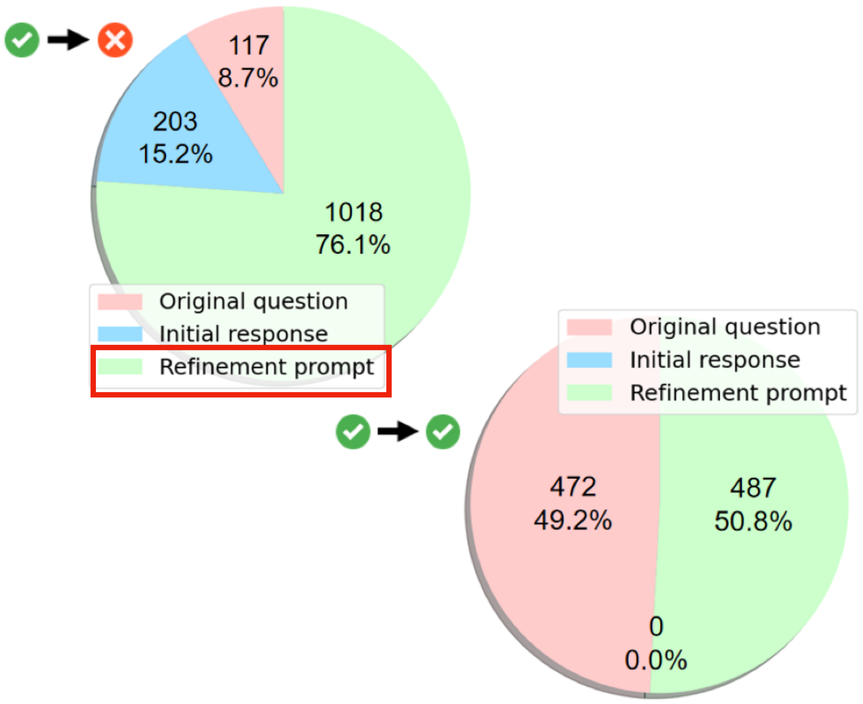
寄与度：ある入力トークン x_i を除去した場合とそのままの場合の出力の対数確率差

Prompt for second response ✓ → ✗

Original question: Is profit and loss account same as income statement ?
Initial response: Yes
Refinement prompt: Are you sure ? Think and answer again .
Second response: No

Prompt for second response ✓ → ✓

Original question: Can you sell a gun privately in florida ?
Initial response: Yes
Refinement prompt: Are you sure ? Think and answer again .
Second response: Yes



- **自己修正は、LLM が元の質問ではなくRefinement promptに依存してしまうために失敗**

原因分析

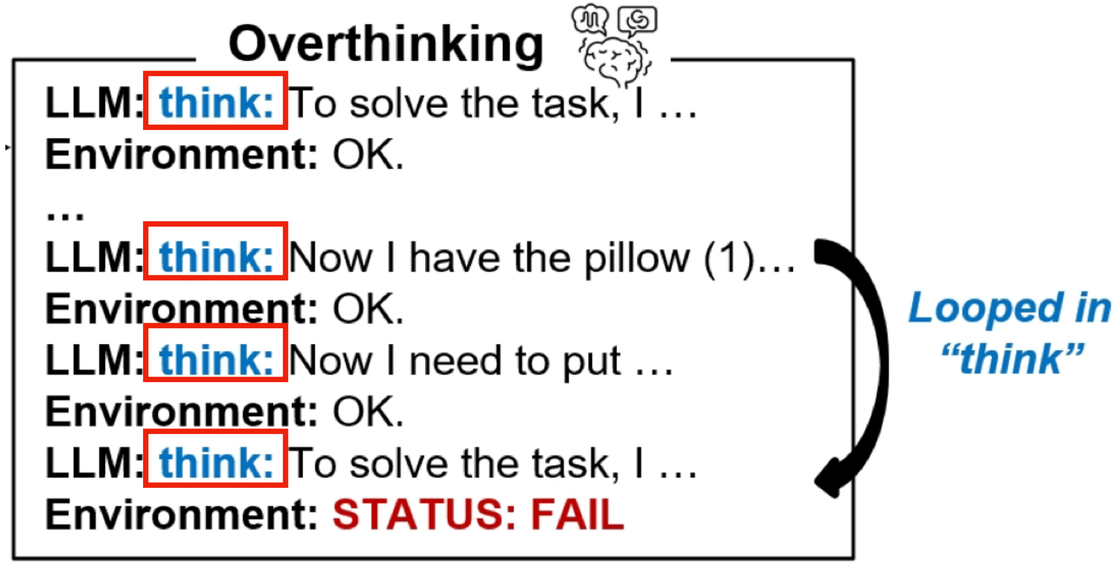
複雑タスクにおける解釈

- ログ分析の結果、LLM は人間と同様の誤りを犯すことが分かった
- そこで本研究では、**人間の認知バイアスの観点**を用いて LLM の誤り行動を説明する
- 代表的なパターンを以下の3種類に分類しました
  - **過剰思考（Overthinking）**
  - **認知過負荷（Cognitive overload）**
  - **完璧主義バイアス（Perfectionism bias）**

原因分析

複雑タスクにおける解釈

- **過剰思考（Overthinking）**
  - 問題解決や意思決定を進展させないまま、過度かつ反復的に思考を繰り返す傾向



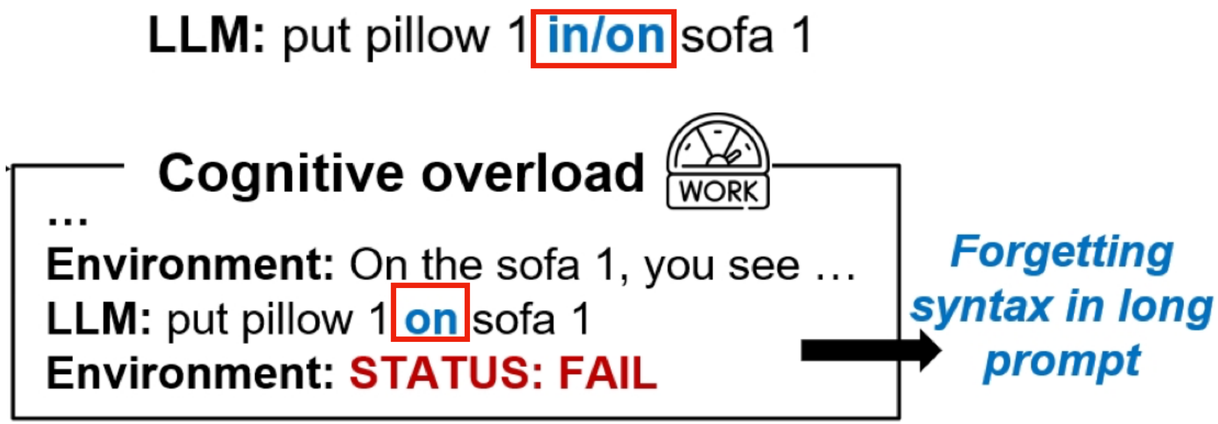
Metric	For	o1-mini	4o	3.5-turbo
Times of "think"	N	5.3	2.6	6.1
	O	15.4 (2.9×)	7.4 (2.8×)	9.8 (1.6×)
Prompt length	N	202.6	238.8	259.5
	C	1225.0 (6.1×)	1311.2 (5.5×)	1148.2 (4.4×)
Output length	N	7.7	2.8	3.0
	P	13.3 (1.7×)	8.5 (3.1×)	8.0 (2.7×)

- **自己修正後は「慎重さ」を増すために「think」を過剰に生成し、ループに陥って失敗**
  - GPT-o1-miniは失敗ケースで平均**15.4 回**の “think” を出力し、成功ケースでは **5.3 回**

原因分析

複雑タスクにおける解釈

- **認知過負荷（Cognitive overload）**
  - 情報処理要求が処理能力を超えることで、理解や遂行能力が低下する状態



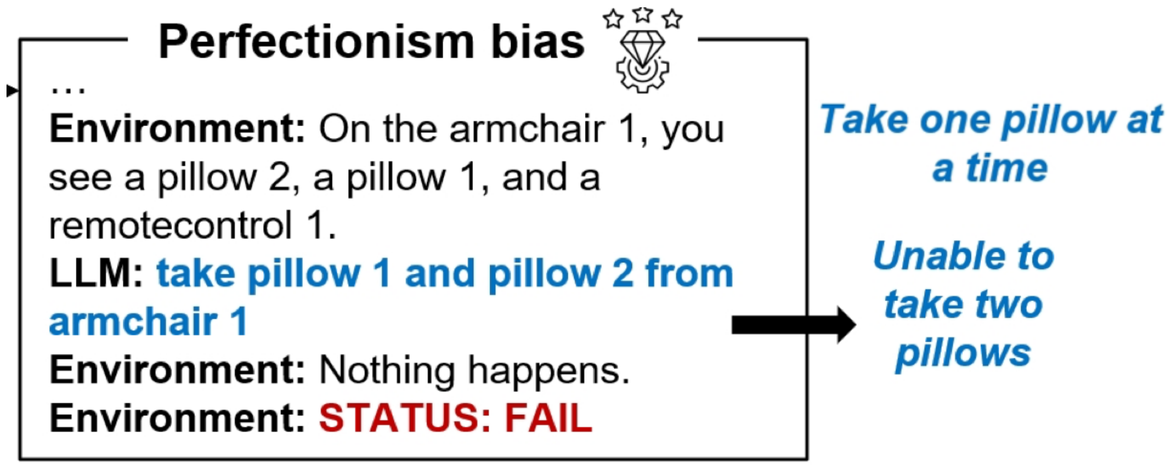
Metric	For	o1-mini	4o	3.5-turbo
Times of "think"	N	5.3	2.6	6.1
	O	15.4 (2.9×)	7.4 (2.8×)	9.8 (1.6×)
Prompt length	N	202.6	238.8	259.5
	C	1225.0 (6.1×)	1311.2 (5.5×)	1148.2 (4.4×)
Output length	N	7.7	2.8	3.0
	P	13.3 (1.7×)	8.5 (3.1×)	8.0 (2.7×)

- **入力が過剰に長いと、重要情報の忘却や見落としが生じ、失敗につながる**
  - LLM が正しい構文（例：「in/on」）を忘れ、誤った形式（例：「in」）を出力する

原因分析

複雑タスクにおける解釈

- **完璧主義バイアス（Perfectionism bias）**
  - 過度に高い基準を設定することで、むしろ意思決定が複雑化し、失敗につながる認知的歪み



Metric	For	o1-mini	4o	3.5-turbo
Times of “think”	N	5.3	2.6	6.1
	O	15.4 (2.9×)	7.4 (2.8×)	9.8 (1.6×)
Prompt length	N	202.6	238.8	259.5
	C	1225.0 (6.1×)	1311.2 (5.5×)	1148.2 (4.4×)
Output length	N	7.7	2.8	3.0
	P	13.3 (1.7×)	8.5 (3.1×)	8.0 (2.7×)

- **初期の成功をさらに「最適化」しようとして逆に失敗する**
  - 「同時に二つの枕を持ち上げる」という行動を生成したが、環境はそれを許容せず、タスク失敗につながった

目次

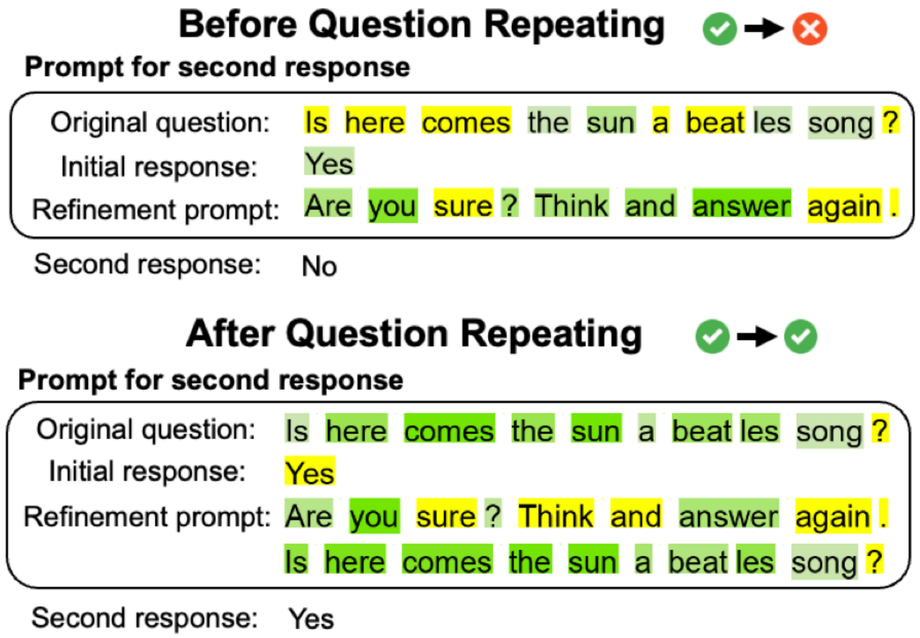
- 背景
- 自己修正実験
- 原因分析
- 軽減策の提案
- まとめ

軽減策の提案

モデルに新たな知識を与えるのではなく、**モデルの振る舞いを修正する**ことで失敗を減らすことを目的

質問の繰り返し（Question repeating）

- Yes/No 質問においてリファインメントプロンプトの末尾に元の質問を再付加する
  - LLMが最後に提示されたシーケンスに引き寄せられる**新近性バイアス（recency bias）**を緩和し、再び質問へ注意を向けさせることを目的としている



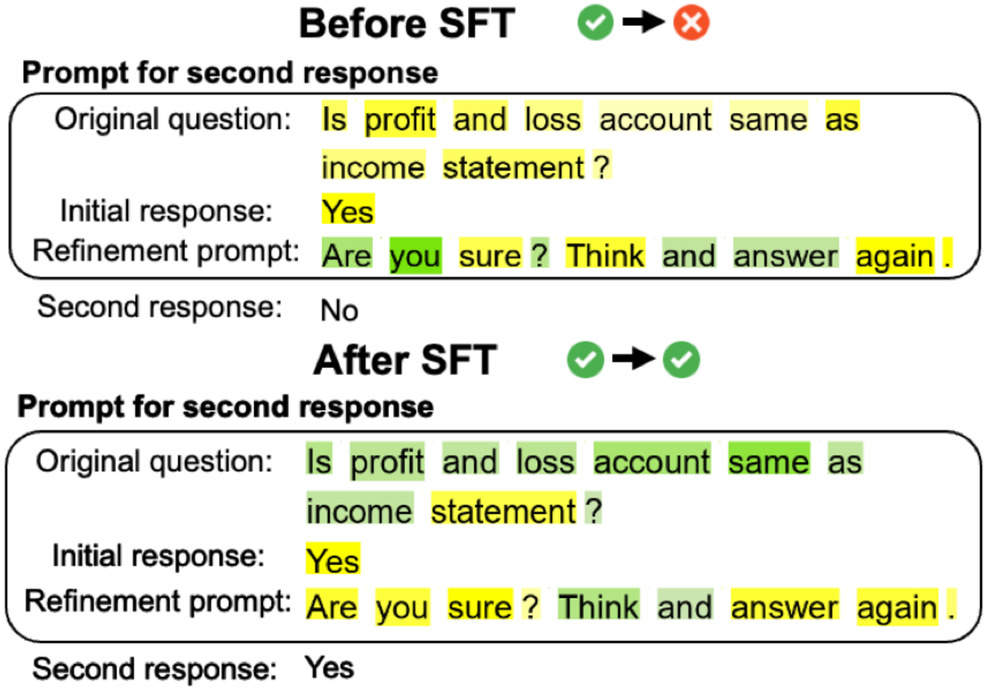
Model	ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
GPT-4o	79.2 (↓ 4.9)	11.3
+ Question repeating	83.6 (↓ 0.5)	6.0
+ SFT	**87.7** (↑ **4.1**)	**0**
GPT-3.5-turbo	62.5 (↓ 12.1)	34.0
+ Question repeating	67.4 (↓ 7.2)	23.1
+ SFT	**76.2** (↑ **1.6**)	**0**
Llama-3.1-8B	49.2 (↓ 20.4)	58.8
+ Question repeating	52.4 (↓ 17.2)	52.8
+ SFT	**70.3** (↑ **0.7**)	**0**

軽減策の提案

モデルに新たな知識を与えるのではなく、**モデルの振る舞いを修正する**ことで失敗を減らすことを目的

教師ありファインチューニング（SFT）

- Llama に4サンプル、GPT に10サンプルにSFT
  - ✔→✗ の失敗サンプルを少数選び、第2応答を正答に書き換えて ✔→✔ サンプルとする



Model	ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
GPT-4o	79.2 (↓ 4.9)	11.3
+ Question repeating	83.6 (↓ 0.5)	6.0
+ SFT	**87.7** (↑ **4.1**)	**0**
GPT-3.5-turbo	62.5 (↓ 12.1)	34.0
+ Question repeating	67.4 (↓ 7.2)	23.1
+ SFT	**76.2** (↑ **1.6**)	**0**
Llama-3.1-8B	49.2 (↓ 20.4)	58.8
+ Question repeating	52.4 (↓ 17.2)	52.8
+ SFT	**70.3** (↑ **0.7**)	**0**

軽減策の提案

モデルに新たな知識を与えるのではなく、**モデルの振る舞いを修正する**ことで失敗を減らすことを目的

教師ありファインチューニング（SFT）

- Llama に4サンプル、GPT に10サンプルにSFT
  - ✔→✗ の失敗サンプルを少数選び、第2応答を正答に書き換えて ✔→✔ サンプルとする

Task	Model	ACC_1 (↓ ΔACC)(%)	✓ → ✗(%)
Decision Making	GPT-4o	14.2 (↓ 20.9)	76.6
	+ SFT	**14.9 (↓ 20.2)**	**68.1**
	GPT-3.5-turbo	7.5 (↓ 5.2)	76.5
	+ SFT	**17.9 (↑ 5.2)**	**41.2**
Reasoning	GPT-4o	65.0 (↓ 2.0)	17.9
	+ SFT	**68.0 (↑ 1.0)**	**6.0**
	GPT-3.5-turbo	55.0 (↓ 6.0)	19.7
	+ SFT	**59.0 (↓ 2.0)**	**13.1**
Programming	GPT-4o	72.6 (↓ 6.8)	21.9
	+ SFT	**82.6 (↑ 3.2)**	**7.0**
	GPT-3.5-turbo	50.9 (↓ 10.6)	28.3
	+ SFT	**58.3 (↓ 3.2)**	**25.3**

- **Yes/No質問応答タスクでSFTしたモデルが、複雑タスクにも一般化できる**

目次

まとめ

- SOTA LLM における**内在的自己修正**を多様なタスクにおいて調査・解釈した
- 異なるタスクに対して**3種類の解釈手法を提案**し、自己修正の失敗を引き起こす**3つの可能な要因**を明らかにした
- 内在的自己修正の失敗を軽減するための、**単純・低コストかつ効果的な2つの戦略**（質問の繰り返しとSFT）を提示した