

プロンプトを2回繰り返すだけ！Google DeepMindが発見した最強の精度向上術「QQメソッド」

Google DeepMindが発表した「プロンプト反復」の仕組み、驚異的な効果、そして実用的な活用場面を視覚的に解説。

【課題】LLMが抱える「片方向読み」の限界



因果的注意 (Causal Attention)

因果的注意 (Causal Attention) の制約：
多くのLLMはテキストを左から右へ順番に処理するため、後に出てくる情報を知らない状態で前の文を読み進める。

Q:要約してください[内容]

- Option 1
- Option 2
- 要約

+

Q:要約してください[内容]

- Option 1
- Option 2
- Option 3

✓ 正答率UP

✗ 正答率DOWN

情報の順序による精度の変動：質問より先に選択肢を提示するなど、情報の配置順序が変わるだけでモデルの正答率が値下する不安定さ。



「因果的ブラインドスポット」の発生：
文中の曖昧さや後半の重要情報を見落とすことで、情報の検索や抽出といったタスクでミスが発生しやすい。

【解決策】驚くほど単純な「プロンプト反復」

Q:要約してください[内容]

+

Q:要約してください[内容]

同一プロンプトを2回コピーして繋げるだけ



二度読みによる「双方向的」な理解：2回目の読み込み時には、1回目の内容がメモリにあるため、全トークンに対して注意を向けられるようになり、情報の盲点が解消。



「無料」の最適化（コスト・速度への影響なし）：
出力トークン数や応答の初速にほとんど影響を与えず、既存のシステムにそのまま組み込める高い互換性。

【検証結果】70通りのテストで47勝0敗の圧倒的成果

70

47勝

(有意な向上)

23件

(ほぼ同等)

0敗

(低下なし)

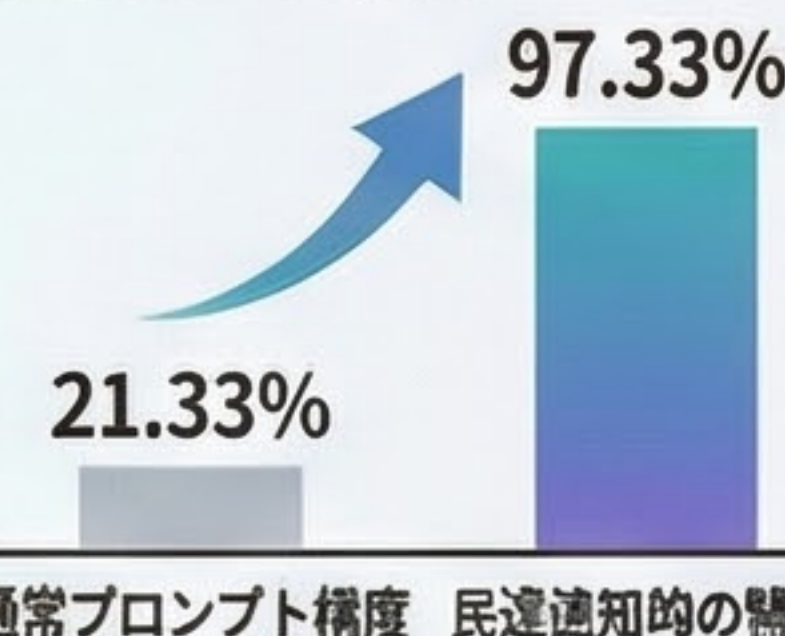
70モデル×タスクで精度が低下したケースは「ゼロ」：
47件で統計的に有意な向上が見られ、残りの23件もほぼ同等。改悪が一件もなかったことが手法の安定性を示す。

特定情報の抽出タスクで
精度が4.5倍に



Gemini 2.0 Flash-Liteを用いた
テストでは、正答率が21.33%から
97.33%へと劇的に跳ね上がった。

主要な全モデルで一貫した効果
Gemini、GPT-4o、Claude 3、
DeepSeekなど、開発元を問わず主要
なLLMで性能向上が確認されている。



実践ガイド：いつ使うべきか？



最も効果的なタスク：
非推論系

- ・ 情報の抽出
- ・ 長いリストからの検索
- ・ 選択証問題
- ・ 要約

モデルが情報と質問を正理に紐付け
る必要があるタスクに最適。



効果が限定的なケース：
高度な推論

- ・ 数学の証明
- ・ 被想な調理思考
(Chain-of-Thought)

モデルが自ら再理推を行うため、
過剰の効果は小さい。



3回以上の反復も可能性あり：一部の複雑に難しい抽出タスクでは、2回よりも2回繰り返すことでさらに精度が向上する余地も示唆されている。