ChatGPT

Google DeepMindによる「Prompt Repetition Improves LLM Accuracy」の研究詳細

論文の要旨と概要

Google DeepMindの研究者らは、**プロンプト（ユーザからの質問や指示）の「繰り返し」**という極めて単純なテクニックで、LLM（大規模言語モデル）の回答精度を大幅に向上できることを報告しました [1]。具体的には、**同一のプロンプト文をそのままコピー＆ペーストして2回連続でモデルに与える**だけで、多くの主要なモデル（例えばGoogleのGemini、OpenAIのGPT-4、AnthropicのClaude、DeepSeekなど）で性能が一貫して向上しました [1] [2]。この手法は**推論（理由づけ）を必要としない**タスクにおいて特に有効で、1回だけプロンプトを与えた場合よりも**統計的に有意な精度向上**が得られるケースが多かったといいます [3] [4]。驚くべきことに、この「プロンプト反復（Prompt Repetition）」は**モデルの応答生成速度や出力トークン数にほとんど影響を与えない**、いわば“無料の”最適化手法でもあります [5] [6]。以下では、この研究の詳細な内容と5つの論点について整理し考察します。

プロンプト反復の手法とメカニズム – なぜ精度が向上するのか

提案された手法自体は非常にシンプルです。**ユーザの入力クエリをそのまま2回繰り返してモデルに与える**だけで、モデルの内部では「質問を2度読む」ことになります [7] [8]。例えば、通常なら「`Q: 次の文章を要約してください: ...`」と1回だけ与えるところを、「`Q: 次の文章を要約してください: ... Q: 次の文章を要約してください: ...`」のように同じ指示文を二度連結して入力します（**QQメソッド**とも呼ばれています [9] [7]）。ポイントは**内容を言い換えたり追加の説明を入れたりせず、本当に全く同じプロンプトを繰り返すこと**です [10]。これだけで、モデルの**応答フォーマットや文体は変化せず**（例えばJSONなど構造化出力もそのまま維持される）、回答が冗長になることもありません [11] [12]。

では、なぜ**同じ入力を二度与えるだけでモデルの理解度や正答率が上がる**のでしょうか？ その鍵は、**Transformerアーキテクチャの「因果的（片方向）注意」の制約**にあります [13]。多くのLLMは**デコーダ専用（因果）言語モデル**として訓練されており、左から右への一方向の文脈でテキストを処理します [14]。すなわち、ある位置のトークンを処理する際、それより「未来」（右側）にあるトークンは参照できず、すでに読んだ「過去」（左側）のトークンのみ注意を向けられるという性質があります [14]。このため**情報の順序**がモデルの理解に大きな影響を与え、**入力文の前後関係によって予測性能が変わってしまう**のです [15] [16]。例えば、「<背景情報><質問>」という順序で与えるのと「<質問><背景情報>」という順序で与えるのでは、モデルの解答精度が変わりうることが知られています [15] [16]。特に**多肢選択問題**では、質問文と選択肢の提示順序を入れ替えるだけで正答率が変動するケースがあります [16]。これは、質問より先に選択肢リストを読んでしまうと、モデルは**何の問いに対する選択肢か分からないまま選択肢を処理する**ことになるためです。

プロンプト反復はこの制約を巧妙に突く方法と言えます [17]。一度目の入力でモデルは質問内容を最後まで「読み切り」ます。そして**二度目に同じ質問を読み始める時、既に一度目の質問全文を記憶した状態**になっています [18]。これにより、**2回目の質問内の各トークンは1回目の質問内の全トークンに注意（Attention）を向けることが可能**となります [18]。言い換えれば、**2回目の質問文はモデルにとって「全文脈を把握した上で読み直す」二周目**に相当し、結果として**双方向の注意機構で入力を理解したかのような効果**が得られるのです [18]。この二度読みのおかげで、**文中の曖昧さの解消**や**後半に出てくる情報の見落とし防止**が図られ、単一回の読み取りよりも正確に問いの意図をつかめるようになります [19] [20]。研究チームはこの効果を「**因果的ブラインドスポット（片方向読み取りの盲点）を埋める**」ものと表現しています [13] [18]。また、**入力を2**

回に増やしてもモデルの出力トークン数は増えないため、生成される回答のフォーマットや長さは元のプロンプトの場合と同じであり、既存システムにそのまま組み込める互換性も保たれます [21] [12]。

「推論専用モデル」で効果が限定的な理由 – チューニング済み対話モデルとの違い

本手法は万能ではなく、**モデルがすでに推論（reasoning）プロセスを積極的に用いる設定では効果が限定的**であることも示されています [22] [23]。論文では「Think step by step（一歩ずつ考えて答えてください）」のようにモデルにチェーン・オブ・ソート（逐次推論）を行わせる指示を与えた場合（**モデルに推論させるモード**）も検証されていますが、その場合は**28通り中わずか5通りでしか性能向上が見られなかった**と報告されています [22]。**大半のケースでプロンプト反復の有無による精度差は統計的に中立（有意差なし）となり、場合によってはごく僅かな改善か劣化**に留まったのです [23]。この理由について、著者らは**「モデルが推論を行う際には、そもそも自分で質問内容を繰り返す傾向がある」**と指摘しています [24]。事実、強化学習によって推論能力を強化した対話型モデル（例えばChatGPTのようなRLHFでチューニング済みのモデル）は、ユーザの質問に答える前に**質問や前提を自ら言い換えたり要約したりして確認する**ような振る舞いを訓練で学習していることがあります [25]。言い換えれば、**高性能な対話モデルは内部で既に「質問の再確認」というプロンプト反復に類する戦略を暗黙裡に実行している**可能性が高いのです [25] [26]。

このように**モデル自身が推論過程で質問内容を繰り返し吟味している場合**、わざわざ入力側で同じ質問を二度投げることは**効果が重複してしまい、大きな上乗せ効果をもたらしません** [27]。ChatGPTのようなチューニング済みの対話モデルではユーザの質問意図を汲み取るために**まず質問を理解・要約し直す挙動**が見られることがあり（回答の冒頭で「ご質問は○○ですね。」と確認するようなケースなど）、こうしたモデルではプロンプト反復の明示的適用による精度向上は小さいか全く見られない場合もあります [25]。一方、**追加の推論指示なしで直接応答させる（推論しない）設定**では、モデルは内部でそうした再確認プロセスを経ないため、プロンプト反復の恩恵が顕著に現れるというわけです [28] [4]。実際、論文でテストされた**7モデルすべて**において、推論なし条件ではプロンプト反復が**一度も負けることなく**性能を改善していました [29] [4]。特に**高性能なGPT-4のようなモデルではベースライン精度自体が既に高いため改善幅が小さく**、統計的に有意な差が出ない（=効果がないように見える）ケースもありましたが、これは上述のように当該モデルが高度に文脈を処理できることの裏返しとも言えます [22] [23]。総じて、**「推論しながら答える」高度にチューニングされた対話モデルほどこの手法の追加効果は限定的**であり、**「与えられた質問にすぐ答える」推論なしモードのモデル・タスクで最大の効果を発揮する**と考えられます [22] [28]。

プロンプト反復が効果を発揮したタスク – どのような課題で有効か

本研究では、7つの代表的ベンチマーク（＋カスタムテスト）でこの手法を検証しています [3]。具体的には、**科学試験問題**のARC-Challenge [30]、**常識や教科書知識QA**のOpenBookQA [30]、数学計算問題集のGSM8K [30]、**多分野知識テスト**のMMLU-Pro [30]、**数学競技問題**のMATH [30]、およびGoogle研究チームが設計した2つのカスタム課題（NameIndexとMiddleMatch）です [30]。結果として、**推論を促さないセッティングにおいて**は、ほぼすべてのタスクでプロンプト反復がベースライン（通常プロンプト）を上回りました [29]。その勝敗を統計的に検定したところ、**70通りのモデル×タスクの組み合わせ中47通りでプロンプト反復が有意に勝り、劣勢だった例は0件（残り23はほぼ同等）**という圧倒的な結果でした [29] [4]。以下に主なタスクでの効果と理由を見ていきます。

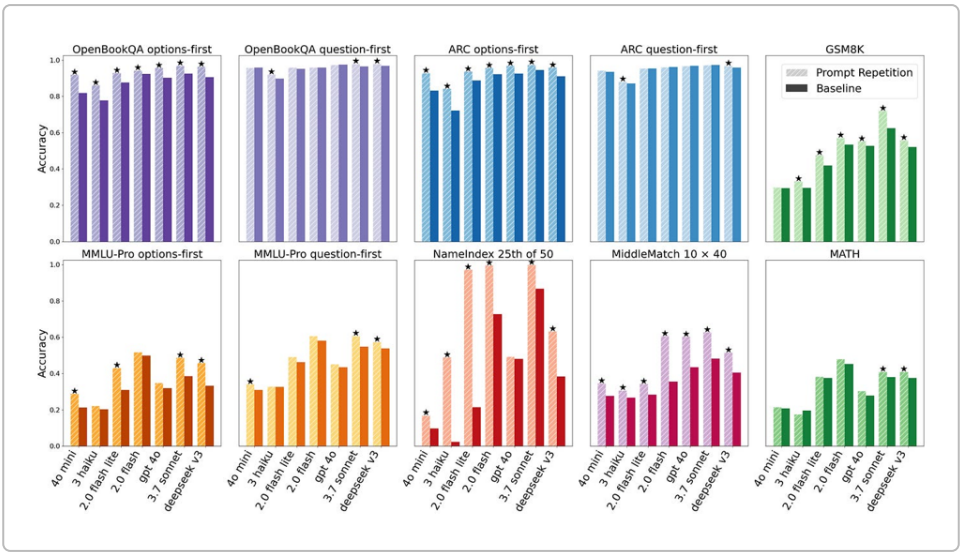

図1：各種ベンチマークにおけるプロンプト反復適用時（斜線ハッチのバー）と通常プロンプト時（無地のバー）の精度比較 [31] [29]。7つのモデル（横軸）についてプロンプト反復適用の有無で精度を測定。星印☆は統計的有意な差（McNemar検定による）でプロンプト反復の方が高精度だったことを示す [31]。70組のモデル・タスク中47組で有意な精度向上が確認され、劣化は0組だった [31] [29]。特にNameIndexやMiddleMatchなど入力内から特定情報を取り出す課題で大幅な改善が見られる。（出典：Yaniv Leviathanら, 2025）*

まず、**選択式問題**（多肢選択）への効果が顕著です。ARCやOpenBookQA、MMLUといった四択問題では、**質問文と選択肢の提示順**がモデルの解答精度に影響することが知られています [15] [16]。研究では各ベンチマークについて**「質問→選択肢」の標準順序**と、逆に**「選択肢→最後に質問」の順序**（質問文を後出しにする不利な形式）の両方でテストが行われました [32]。その結果、**質問後出し（選択肢先読み）**の形式ではプロンプト反復による大幅な精度向上が確認されています [33]。これは前述の通り、**初回読みでは選択肢を文脈なしに読む羽目になるところ、二度目の読み込み時には既に質問内容を把握して選択肢を再評価できる**ためです [17] [18]。実際、MMLUやARCでも**「選択肢→質問」の条件下ではプロンプト反復が安定して正答率を押し上げ** [33]、効果が大きく現れました（逆に**質問先出し**の条件では効果は控えめですが、それでも改善傾向は見られました [33]）。このことから、**プロンプト内で重要情報（質問）が後に出てくるタスク**において、繰り返し入力が**モデルの認知負荷を減らしパフォーマンスを改善する**と考えられます。

次に、**情報検索・抽出系のタスク**での効果です。カスタム課題のNameIndexは、50個の名前のリストを提示して「25番目の名前は何か？」と問うテストで、モデルの**記憶保持と正確な項目参照**能力を測るものです [34]。この課題において、**Gemini 2.0 Flash-Liteモデルの正答率は通常入力では21.33%と低迷していましたが、プロンプト反復を使うと97.33%にまで跳ね上がりました** [34]。**4.5倍以上**もの劇的な改善であり、モデルが**一度の読みでリスト中の位置を正確に数えられない**「うっかりミス」を、二度読むことで完全に解消できたことを示唆しています [35]。同様にMiddleMatchというカスタム課題（おそらく長い入力内の対応関係を見つけるテスト）でも、**各モデルで大幅な精度向上**が観察されました [36] [37]。このように**入力全体を見渡して特定要素を探す**タイプの問題では、**一度目の読みで曖昧だった手がかりを二度目で捕捉し、回答の正確さを高められる**ことが分かります [18] [34]。

一方で、**数学的推論や逐次的な論理思考を要するタスク**では、プロンプト反復だけでは根本的な課題解決力は向上しにくいことも確認できます。GSM8K（小学生〜中学生レベルの数学単語問題集）やMATH（数学コンペ問題）は本来**途中計算や論証**が必要なタスクであり、モデルに推論を促さず**いきなり答えさせる設定ではベースラインの性能自体が低く**なりがちです [38]。プロンプト反復を適用しても、問題の難解さゆえに**劇的なスコア向上は見られず**、多くの場合統計的にも有意差なし（あるいは僅かな改善）という結果でした [23] [22]。これは**問題そのものが一度読むだけで解ける性質ではない**ためで、真に性能を上げるにはチェーン・オブ・ソート等でモデルに論理展開させる必要があります。実際、著者らも**「この手法はステップバイステッ**

プの推論が不要なタスクに主眼を置いたもの」であり、複雑な推論課題に対しては万能ではないと述べています [39] [40]。総じて本手法の恩恵が大きいのは、**知識の検索・適用やシンプルな質問応答など、モデルが質問文と与えられた情報を正確に対応付けることが要求されるタスク**だと言えるでしょう [34] [23]。

実験に使用されたモデルと評価方法

研究では、**業界を代表する7種のLLMモデル**を選び、2025年2〜3月時点で各モデル提供元の公式APIを用いてテストを行いました [41] [42]。対象モデルは以下の通りです [2]。

- **Gemini 2.0 Flash / Flash Lite（Google）**: Googleが開発中の次世代LLM「Gemini」の軽量版モデル。Flash Liteはパラメータ数を抑えた小型モデル、Flashはその上位版と推測されます [2]。
- **GPT-4o / GPT-4o-mini（OpenAI）**: OpenAIのGPT-4モデル。論文中では「GPT-4o」と表記されていますが、これはGPT-4の標準モデル（おそらく2024年版のAPIモデル）と、その縮小版と思われるモデルです [2]。(注：「GPT-4o-mini」はGPT-4の小規模版かGPT-3.5に相当するモデルの可能性がありますが、詳細は公開されていません [41]。)
- **Claude 3 Haiku / Claude 3.7 Sonnet（Anthropic）**: Anthropic社のLLMで、Claude 2の後継となる**Claude 3**シリーズのモデルです [2]。HaikuやSonnetはモデルのバリエーション（例えば文体調整やサイズ違い）を示すコードネームと思われます。
- **DeepSeek V3（DeepSeek-AI）**: DeepSeek-AI社によるLLMモデル。Version 3という最新のモデルで、他の大手に比べ知名度は低いものの精度検証に加えられています [2]。

評価にあたって、各モデルには**統一のプロンプトフォーマット**で質問が与えられました。多肢選択問題については前述のように**質問文と選択肢の順序を2通り**（質問先/後）用意し、それぞれ**通常プロンプト**と**プロンプト反復（同じ入力を2回連結）**のケースで回答を生成させています [30] [32]。モデルには特に**思考過程を出力しないよう指示**し（推論なし設定）、**直接答えだけ返す**よう求めました [43]。また一部のタスクでは**「逐次推論せよ」という指示（Chain-of-Thought促進）**も併用し、その際のプロンプト反復効果も検証しています [23]。回答の正誤は各ベンチマークの標準に従って採点し、**プロンプト反復あり vs なしで精度に有意差があるか**を**McNemar検定**によって比較しています [44] [45]。この検定はペアごとの勝敗を分析するもので、先述の「47勝0敗」はこの統計的判定に基づくものです [31] [29]。

加えて**性能以外の要素**として、**各モデルの応答の長さ**（生成トークン数）や**応答までの遅延時間**も測定されました [46]。興味深いことに、**プロンプトを2回送ったからといってモデルの回答が冗長になったり長くなることはなく**、ほとんど全てのケースで**出力長は通常プロンプト時と変わりませんでした** [47]。**応答遅延**についても、プロンプト反復は**モデル内部の並列実行可能な「プレフィル段階」**（入力テキストを読み込むプロセス）に負荷を増やすだけで、**系列生成（デコード）段階には影響を与えない**ため、ユーザが感じる応答の**初速や全体の待ち時間はほぼ変化しない**ことが示されています [48] [6]。実験上も**大半のモデルで「1回読み」と「2回読み」で初回トークンが出てくるまでの時間差は検出されず**、結果として**精度だけ上がって速度とコストは据え置き**という理想的な状況が確認されました [6]。(ただしAnthropicのClaude系モデルのみ、入力が極端に長い場合にプレフィル処理がボトルネックとなりわずかに遅延が増加した例があったとのこと [49]。)

さらに著者らは**プロンプト反復の変種**についても付録で検証しています [50]。例えば**「Verbose（冗長）反復」**では、同じ質問を2回繰り返す代わりに**少し言い回しを変えて二度問いかける**方法、**「三回反復」**では**質問を3回連続で与える**方法などを試し、タスクによっては**それらが通常の2回反復より良い結果を出す場合もある**と報告しています [37] [51]。特にNameIndexやMiddleMatchのような極端に効果が高かったタスクでは、**3回反復が2回反復をさらに上回る精度**を示したとの指摘もあり [51]、この手法のさらなる最適化余地が示唆されています。対照実験として、**プロンプトをただ「.（ピリオド）」で水増しして長さだけ2倍にする**パディング（Padding）も試されましたが、こちらは**全く精度は向上せず**予想通りの結果となりました [52]。したがって**性能向上は入力長そのものではなく、内容の繰り返しによるものである**ことが明確に示されていま

す[52]。総合すると、評価は厳密かつ多角的に行われており、**プロンプト反復の有効性とその条件が定量的に裏付けられた**と言えます。

知見の意義と今後の応用可能性

この研究の示す知見は、実用面・理論面の双方で大きな意義を持ちます。まず実用的観点では、**極めて簡単な工夫でLLMの応答精度を底上げできる「無料のランチ」**である点が注目されます[5][6]。高度なモデルチューニングや追加の外部ツール統合を行わずとも、**プロンプトをコピー&ペーストするだけで最大数十%以上もの精度向上**が得られるのは驚くべき成果です[1][34]。企業のLLM導入においては、品質・速度・コストのトレードオフが常に課題ですが、本手法は**出力品質を向上させつつ推論コストやレイテンシは据え置き**という理想的な改善をもたらします[6][53]。実際、研究者らも「**闇雲に全シナリオで使える銀の弾丸ではないが、適切に見極めて使えば大きな恩恵が得られる戦略的手法**」と述べており[54]、例えばエンタープライズのLLM活用でも**対話システムの応答精度向上**や**質問応答ボットの正答率改善**にすぐ応用し得るテクニックです。特に**ユーザからの単発質問に直接答えるタイプのアプリケーション**では相性が良く、社内QAシステムやカスタマーサポートチャットボットの解答精度向上に寄与する可能性があります。

次に理論的・研究的な観点では、**現在のデコーダ型LLMが抱える本質的制約（片方向文脈処理）を浮き彫りにし、その解決策の一端を示した**点が重要です[13][17]。Transformerモデルは強力ですが、**テキスト全体を一度に見渡せない**という弱点から、時に**重要な情報の取りこぼし**や**文脈の誤解**が生じます。本研究は、その問題に対し**モデル自身の内部構造を変えることなく、入力側の工夫だけで「二方向文脈」の利点を擬似的に与える**というアプローチを提案しました[17][18]。これは、将来的に**モデル設計や訓練の改良**にもインスピレーションを与えるでしょう。例えば、**モデルが自律的に入力を再読する機構**や、**まず入力全体を要約・反復してから回答を考える訓練**など、今回の知見を組み込んだ新たなアーキテクチャ・プロンプト手法が考案されるかもしれません。実際、既にRLHFで訓練されたモデルが**暗黙的にプロンプト反復戦略を身につけていた**という示唆は興味深く、**モデルが自己改善の中で見出した戦略を人間側が形式知化して活用する**好例と言えます[25]。

さらに、本手法は**他のプロンプト手法との併用や発展**も考えられます。例えば、チェーン・オブ・ソート（思考の連鎖）とは相性が良くないとされますが、**ツール使用や外部知識参照**と組み合わせる場合はどうか、**プロンプト反復＋少数ショット学習**で効果が増幅されるケースはないか、といった応用研究も期待されます。著者らも触れているように、**プロンプト反復のバリエーション（異なる言い回しでの反復や複数回反復）**は更なる改善ポテンシャルを秘めており[37]、今後も**「モデルに二度考えさせる」アイデアを発展させた研究**が進むでしょう。**人間が文章を読み返して理解を深める**ように、LLMに対しても追加の読解機会を与えるだけで知的性能が上がるという発見は、AIと人間の問題理解プロセスの類似性を感じさせる興味深い知見でもあります。以上のように、「プロンプトをただ2回繰り返すだけ」という魔法のように簡単な方法が示した効果は、実務的にも学術的にも大きなインパクトを与えており、**今後のLLM最適化や活用法における重要な一手**となる可能性があります[54][5]。

参考文献・出典：Prompt Repetition Improves Non-Reasoning LLMs[55][29]、VentureBeat[1][13]、リツアンSTC[28][22]、他。

---

[1][3][5][6][13][14][15][17][18][19][20][23][24][26][27][34][35][39][43][48][49][53][54] This new, dead simple prompt technique boosts accuracy on LLMs by up to 76% on non-reasoning tasks | VentureBeat
https://venturebeat.com/orchestration/this-new-dead-simple-prompt-technique-boosts-accuracy-on-llms-by-up-to-76-on

[2][4][7][8][9][10][16][22][28][38][40] 同じプロンプトを繰り返すだけで精度が上がる？「QQメソッド」とは | リツアンSTC
https://ritsuan.com/ai_news/24927/

11 12 Prompt Repetition: Google's 47/70 Benchmark Winner
https://prosperinai.substack.com/p/prompt-repetition-google-research

21 29 30 31 32 33 36 37 41 42 44 45 46 47 50 52 55 Prompt Repetition Improves Non-Reasoning LLMs
https://arxiv.org/html/2512.14982v1

25 Google's "Free Lunch" for LLMs: How Prompt Repetition Fixes Attention Bottlenecks with Zero Latency - DataSci Ocean
https://datasciocean.com/en/paper-intro/prompt-repetition/

51 Prompt Repetition Improves Non-Reasoning LLMs - a paper : r/LocalLLaMA
https://www.reddit.com/r/LocalLLaMA/comments/1qeuh0z/prompt_repetition_improves_nonreasoning_llms_a/