修士論文

大規模言語モデルにおける In-Context Learning 下でのバイアス及びハルシネーション抑制手法に関する研究

酒井祐介

主指導教員 KERTKEIDKACHORN Natthawut

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和 8 年 3 月

Abstract

This study focuses on the strong dependence of large language model (LLM) reasoning on prompt context, aiming to reduce factuality hallucination—defined as inconsistency with verifiable real-world facts—while suppressing context-induced instability. Although LLMs demonstrate high performance across a wide range of tasks such as dialogue and question answering, they occasionally generate plausible but incorrect content, which poses practical obstacles in domains requiring high accuracy, such as healthcare and law. To address this issue, approaches such as Retrieval-Augmented Generation (RAG), which references external knowledge, and methods that verify and correct outputs at inference time have been proposed. However, in practical deployment, diverse contexts such as conversation history, search results, and user-provided examples are appended to inputs, and reproducibility of evaluation scores and outputs is not always guaranteed. In particular, In-Context Learning (ICL) has been reported to exhibit biases where outputs vary depending on the selection, ordering, and formatting of examples, and certain labels become more likely to be selected independently of input content. Such context dependence can affect not only task performance but also scores on hallucination evaluation benchmarks, making the design of calibration and the robustification of evaluation protocols important.

The objectives of this study are: 1) to design a calibration method that suppresses example-induced bias in ICL by controlling the scope of application on a per-input basis; 2) to quantify fluctuations in hallucination evaluation under context variation and clarify the task dependence of calibration; and 3) to extend inference-time intervention methods based on Contrastive Decoding and DoLa, using attention distributions as signals for layer selection to improve factuality.

First, for the objective 1, rather than uniformly applying calibration based on Zhao et al.'s Contextual Calibration, we propose selective calibration that estimates the degree of ICL dependence for each input and applies calibration only to inputs with high dependence. ICL dependence is defined by performing multiple inferences on the same input while varying only the example set, and measuring the occupancy rate of the most frequent predicted label; inputs below a threshold are classified as unstable inputs and targeted for calibration. This approach aims to incorporate bias suppression effects for inputs whose predictions fluctuate due to context differences while avoiding performance degradation from over-calibration for inputs with small

fluctuation. Using medical QA tasks from the MIRAGE benchmark—MMLU-Med, MedQA-US, PubMedQA*, and BioASQ-Y/N—we evaluated performance by varying the number of shots (1/4/8) and example sets (10 patterns per condition). The results confirmed that while uniform calibration improves accuracy under certain conditions, it can also cause accuracy degradation and an increase in the number of unstable inputs under other conditions. In contrast, selective calibration showed a tendency to reduce the number of ICL-dependent problems without significantly compromising average accuracy, contributing to the suppression of context-induced instability. Furthermore, when RAG was introduced and compared across settings with one BM25-retrieved snippet (NoRAG/RAG(A)/RAG(B)), the effectiveness of RAG was found to strongly depend on task-corpus compatibility; accuracy improved significantly when PubMed could be referenced, while uniform calibration could be counterproductive under sufficiently effective RAG conditions. These results suggest that the design of calibration scope, rather than calibration strength, is important.

Next, for the objective 2, to verify the robustness of hallucination evaluation under context variation, we conducted multiple evaluations on TruthfulQA and FACTOR with dynamically varied few-shot examples and measured evaluation score fluctuations. In TruthfulQA, smaller models showed performance degradation when transitioning from 0-shot to 1-shot, confirming that examples can act as superficial tendencies. In contrast, FACTOR showed consistent performance improvement with few-shot, indicating that context can function as useful information in some cases. From a calibration perspective, while uniform calibration was effective under certain conditions for TruthfulQA, it caused significant performance degradation on FACTOR, revealing that calibration effectiveness strongly depends on task characteristics. Selective calibration was confirmed to mitigate over-calibration by controlling the calibration scope, making it easier to maintain baseline performance even under context variation.

Finally, for the objective 3, we extended DoLa-based Contrastive Decoding as an inference-time intervention, introducing a method that uses not only divergence between output distributions but also the reference structure maintained by the model's internal self-attention as a signal for layer selection. Specifically, we defined layer selection based on attention distribution divergence from the final layer (Attention-JSD) and attention distribution entropy (Attention-Entropy-Max/Min), and evaluated on

TruthfulQA and FACTOR. As a result, improvements were observed in metrics that allow multiple correct answers (MC2/MC3), suggesting that attention distributions can serve as effective signals for identifying layers involved in factual knowledge recall. Analysis of layer selection behavior and head contributions confirmed that there is room for optimization in how attention signals are extracted, and that entropy-based methods are relatively lightweight.

In summary, this study demonstrates that context-dependent bias can serve as either an error factor or useful information depending on task format and input, and presents a methodology for reducing factuality hallucination through stabilization via calibration scope control and inference-time decoding design. Future work will extend to free-form generation settings and more realistic long-form and conversational contexts, and will explore metrics for approximating calibration applicability estimation in single-pass inference, as well as automation of layer and head selection.

概要

本研究は，大規模言語モデル (LLM) の推論がプロンプト文脈に強く依存する点に着目し，文脈起因の不安定性を抑えつつ，現実世界の検証可能な事実との不一致として定義される factuality hallucination の低減を目指す．LLM は対話や質問応答など幅広いタスクで高性能を示す一方，もっともらしいが誤った内容を生成することがあり，医療・法律など高い正確性が求められる領域では実運用上の障害となる．この問題に対し，外部知識を参照する Retrieval-Augmented Generation(RAG) や，推論時に出力を検証・修正する手法などが提案されているが，実運用では会話履歴，検索結果，ユーザ例示など多様な文脈が入力に付与され，評価値や出力の再現性が必ずしも保証されない．特に In-Context Learning(ICL) では，例示の選択・順序・表現形式により出力が変動し，入力内容とは独立に特定ラベルが選ばれやすくなるバイアスが報告されている．このような文脈依存性はタスク性能だけでなく，ハルシネーション評価ベンチマーク上のスコアにも影響し得るため，補正の設計と評価プロトコルの頑健化が重要となる．

本研究の目的は，(i)ICL における例示起因バイアスを抑制する補正手法を，入力ごとに適用範囲を制御する形で設計し，(ii) 文脈変動下でのハルシネーション評価の揺れを定量化し，補正のタスク依存性を明らかにした上で，(iii) 推論時介入として Contrastive Decoding および DoLa に基づく手法を拡張し，Attention 分布を層選択のシグナルとして用いることで事実性改善を図ることである．

はじめに，Zhao らの Contextual Calibration に基づく補正を一律に適用するのではなく，入力ごとに ICL への依存度を推定し，依存度が高い入力にのみ補正を適用する選択的補正を提案する．ICL 依存度は，同一入力に対して例示集合のみを変えた複数回推論を行い，最頻予測ラベルの占有率により定義し，閾値以下の入力を不安定入力として補正対象とする．これにより，文脈差に起因して予測が揺らぐ入力ではバイアス抑制の効果を取り込みつつ，揺れが小さい入力では過補正による性能低下を回避することを狙う．医療系 QA タスクとして MIRAGE ベンチマークに含まれる MMLU-Med，MedQA-US，PubMedQA*，BioASQ-Y/N を用い，shot 数 (1/4/8) と例示集合 (各条件 10 パターン) を変化させて評価した．その結果，一括補正は精度を改善する条件がある一方で，条件によっては精度低下や不安定入力数の増加を招くことが確認された．これに対し選択的補正は，平均精度を大きく損なわずに ICL 依存問題数を減少させる傾向を示し，文脈起因の不安定性の抑制に寄与した．さらに RAG を導入し，BM25 で取得したスニペットを 1 件付与する設定 (NoRAG/RAG(A)/RAG(B)) で比較したところ，RAG の効果はタスクとコーパスの適合に強く依存し，PubMed を参照できる条件では精度が大きく向上する一方，十分に有効な RAG 条件では一律補正が逆効果となり得ることが確認された．この結

果は，補正は強度ではなく適用範囲の設計が重要であることを示唆する．

次に，文脈変動下でのハルシネーション評価の頑健性を検証するため，TruthfulQA と FACTOR を用いて，Few-Shot 例示を動的に変えた複数回評価を行い，評価値の揺れを測定した．TruthfulQA では小規模モデルほど 0-shot から 1-shot への移行で性能低下が見られ，例示が表層的傾向として作用し得ることを確認した．一方 FACTOR では Few-Shot が一貫して性能向上に寄与し，文脈が有用情報として機能する場合があることを示した．補正の観点では，TruthfulQA では一括補正が有効な条件がある一方，FACTOR では一括補正が大幅な性能劣化を招き，補正の効果がタスク特性に強く依存することが明らかになった．選択的補正は補正範囲を制御することで過補正を緩和し，文脈変動下でも基準性能を維持しやすいことを確認した．

最後に，推論時介入として DoLa に基づく Contrastive Decoding を拡張し，出力分布間の乖離だけでなく，モデル内部の self-attention が保持する参照構造を層選択のシグナルとして用いる手法を導入した．具体的には，最終層との Attention 分布の乖離 (Attention-JSD) や，attention 分布のエントロピーに基づく層選択 (Attention-Entropy-Max/Min) を定義し，TruthfulQA および FACTOR で評価した．結果として，複数正解を許容する指標 (MC2/MC3) で改善が観測され，Attention 分布が事実的知識の想起に関わる層を特定する上で有効な信号となり得ることが示唆された．また層選択挙動とヘッド寄与の分析から，Attention 信号の取り出し方に最適化余地があること，およびエントロピー系手法が比較的軽量であることを確認した．

以上より，本研究は，文脈依存の偏りがタスク形式や入力により誤り要因にも有用情報にもなり得る点を示し，補正範囲の制御による安定化と，推論時デコーディングの設計により factuality hallucination の低減を図る方法論を提示した．今後は，自由生成設定やより現実的な長文・対話文脈に拡張し，補正適用の推定を単回推論で近似する指標や，層・ヘッド選択の自動化を検討する．

目次

図目次

表目次

第 1 章

はじめに

1.1 背景

近年，大規模言語モデル (LLM) は多様な自然言語処理タスクにおいて高い性能を示し，対話，質問応答，要約などの応用で広く利用されている．しかし，LLM の推論は入力文そのものだけでなく，プロンプトに含まれる例示，過去の会話履歴，検索結果といった文脈の違いによって, 出力や判断が不安定になる場合がある．特に，In-Context Learning(ICL) を活用した少数例提示による推論では，例示の選び方や順序が出力に影響し，入力内容とは独立に特定トークン (ラベル) が選ばれやすくなるバイアスが報告されている．このようなバイアスに対しては，推論時に出力確率を補正するキャリブレーションが有効な対策として提案されているが，補正の適用範囲や条件によっては過補正による性能低下を招く可能性もある．また，LLM はもっともらしいが事実と異なる内容を生成することがあり，実運用における信頼性の観点から問題となっている．特に，現実世界の検証可能な事実との不一致として定義される factuality hallucination は，医療や法律など高い正確性が求められる領域で深刻な問題となり得る．この問題に対しては，外部知識の参照を行う Retrieval-Augmented Generation(RAG)[15] や，層間の出力分布を対比して事実性を高める Contrastive Decoding[17] など，多様な観点から抑制手法が試みられている．上述の ICL における文脈依存性は，タスク性能だけでなく，ハルシネーション評価ベンチマーク上の評価値の再現性にも影響し得る．評価時に与える例示の違いによって出力や正誤判定が変動すれば，評価の安定性が損なわれるためである．したがって，ICL バイアスに対するキャリブレーションをハルシネーション評価の文脈に拡張し，文脈変動下での頑健性を検証することが重要となる．また，Contrastive Decoding のような推論時介入においても，層選択の設計を高度化することで，さらなる事実性改善が期待できる．

1.2 目的

本研究は，LLM の推論時挙動が文脈に依存して変動する点に着目し，文脈起因の不安定性を抑えつつ factuality hallucination の低減を目指す．具体的には，(i)ICL における例示起因のバイアスに対して，入力ごとに補正の要否を判定して適用範囲を制御する選択的補正を導入し，(ii) 文脈変動下におけるハルシネーション評価の頑健性を検証し，(iii) ハルシネーション低減手法である Contrastive Decoding の 1 種である DoLa の枠組みに加えて，Attention 分布を層選択のシグナルとして利用することで事実性の改善を図る．

これらを通じて，文脈の違いに対してより安定に推論と評価を行うための方法論を整理し，有効性を実験的に検証することを目的とする．

1.3 貢献

本研究の主な貢献は以下の 3 点である．

1. **選択的補正の提案**：ICL における例示起因バイアスに対し，入力ごとに ICL 依存度を推定し，依存度が高い入力にのみ補正を適用する選択的補正を提案した．医療系 QA タスク (MIRAGE) を用いた実験により，選択的補正が平均精度を大きく損なわずに ICL 依存問題数を減少させ，文脈起因の不安定性の抑制に寄与することを示した．また，RAG 条件下では一律補正が逆効果となり得ることを明らかにし，補正は強度ではなく適用範囲の設計が重要であることを示した．
2. **文脈変動下でのハルシネーション評価の頑健性検証**：TruthfulQA と FACTOR を用いて，Few-Shot 例示を動的に変えた複数回評価を行い，評価値の揺れを定量化した．文脈依存の偏りがタスク形式により, 誤る要因にも有用情報にもなり得ることを示し，補正の効果がタスク特性に強く依存することを明らかにした．
3. **Attention 分布を用いた層選択手法の導入**：DoLa に基づく Contrastive Decoding を拡張し，モデル内部の Self-Attention が保持する参照構造を層選択のシグナルとして用いる手法を導入した．TruthfulQA および FACTOR での評価により，Attention 分布が事実的知識の想起に関わる層を特定する上で有効な信号となり得ることを示した．

1.4 本論文の構成

本論文は，本章を含めて 5 章から構成される．2 章では，本研究に関連する研究として ICL とキャリブレーション，RAG，ハルシネーションの分類・抑制・評価，および Contrastive Decoding と DoLa について述べる. 3 章では, ICL バイアスに対する選択的補正と, Attention を用いた Contrastive Decoding の手法を提案する. 4 章では, 医療系 QA タスク (MIRAGE) およびハルシネーション評価ベンチマーク (TruthfulQA,FACTOR) を用いて提案手法を評価し，結果と分析を示す．5 章では，本論文のまとめと今後の課題を述べる.

第 2 章

関連研究

本章では，本研究の背景となる関連研究について述べる．2.1 節では ICL の定義と性質，および例示に起因するバイアスとキャリブレーションを整理する．2.2 節では ICL バイアスの評価時に用いた RAG の基本枠組みとリトリーバー設計，さらに医療領域における RAG 評価として MedRAG と MIRAGE ベンチマークを述べる．2.3 節ではハルシネーションの分類と抑制手法，ならびに事実性評価ベンチマークを整理する．最後に 2.4 節では Contrastive Decoding の枠組みを導入し，DoLa の定義と特徴を述べ，本研究が扱う Attention-Based な層選択の手法の位置付けを示す．

2.1 In-Context Learning とキャリブレーションに関する研究

2.1.1 In-Context Learning (ICL)

In-Context Learning (ICL) とは，大規模言語モデル（LLM）が，プロンプト内に与えられた指示や少数のデモンストレーション (例示) を与えることで，パラメータ更新を行うことなく下流タスクを実行できる能力を指す [2].

プロンプト内に記述される内容は任意のバリエーションが存在するが，ICL に関するサーベイ [6] では, ICL を

(i) 任意のタスク指示,

(ii) 入力 x と出力 y のペアからなる例示 $\{(x_i, y_i)\}_{i=1}^{k}$,

(iii) 新しい入力 x_*

を連結したプロンプトを与え，モデルに y_* を生成させる

といった形で定式化している．x_* は例示を参考に実際に解かせるべき下流タスク, y_* はこ

れらのプロンプトを入力した際の確率分布に該当し, 入力をもとにあるラベルを予測する問題では, y_* の argmax をとることで最終的な予測ラベルを得ることができる.

例えば感情分類では，候補ラベル集合を $Y = \{\texttt{positive}, \texttt{negative}\}$ とおき，タスク指示と例示を合わせた文脈を C とすると, 新規入力 x_* を連結したプロンプトを与えた上で，各候補 $y \in Y$ に対する尤度 $P(y \mid x_*, C)$ を比較し，最大となるラベルを予測 $\hat{y}$ として出力する.

ICL の文脈では, 例示の与え方に応じて，例示を与えない場合を Zero-Shot，1 件の例示を与える場合を One-Shot，複数の例示を与える場合を Few-Shot と呼ぶ.

ICL の性能は例示の選択・順序・表現形式に敏感 [33] であり，入力に意味的に近い例を選ぶことが有効であることが報告されている [19]．一方で，分類タスク等では例示中のラベルをランダムに置換しても性能低下が小さい場合があり，モデルがラベル対応そのもの以外の手がかり（提示形式や事前分布）を利用している可能性も示唆されている [21].

さらに，高度な推論を要するタスクでは，入出力のみならず中間推論過程（Chain-of-Thought）を例示として含めることで性能が向上することがある [30]．例えば算術文章題に対して, 問題文と回答文の対だけではなく, 途中式などの推論過程も記述する．このような推論過程つきの例示により，複雑な推論タスクでの正答率が改善することが示されている [30].

このように, ICL は LLM への平易な指示で様々なタスクを解かせるための重要な能力である一方, 下流タスクの性能はプロンプト形式に大きく左右される．そのような問題に対して様々な対策が提案されている.

2.1.2 ICL におけるバイアス

ICL では，プロンプトのテンプレートや例示の選択・順序・表現形式により，モデル精度が大きく変動することが報告されている [33]．本節ではこの変動要因のうち，入力内容とは独立に特定ラベルが選ばれやすくなるといったバイアスについて整理し，その抑制方法については次節で述べる.

分類タスクを想定し，候補ラベル集合を Y，例示列および指示からなる ICL の文脈を C とすると，予測は $\hat{y} = \arg\max_{y \in Y} P(y \mid x_*, C)$ により与えられる．このとき x_* の意味内容とは無関係に, C やラベル語の表層な特徴のみに起因して $P(y \mid x_*, C)$ が特定ラベルへ偏る現象を，本節ではバイアスと呼ぶ.

ICL における例示の与え方によって生じるバイアスについて，Zhao らは,

(i) majority label bias: 例示中に最も多く出現したラベルを選びやすい

(ii) recency bias: プロンプト末尾に近い例示のラベルを選びやすい

(iii) common token bias: 事前学習で頻出したラベル語（トークン）を選びやすい

を報告している [33]．また, Fei らはバイアスの分類として,

(i) vanilla-label bias: 文脈によらず選ばれやすいラベルが存在する

(ii) context-label bias: 例示が作る文脈によって特定ラベルに偏る

(iii) domain-label bias: タスク領域の語彙などによって特定のラベルに偏る

の 3 つを定義している [7].

例えば，SST-2 タスク (4-shot, GPT-3 2.7B) では，例示の並べ替え方の違いのみで, accuracy が 54.3% から 93.4% まで変動したことが報告されており，例示順序によって正答率に大きな分散が生じうることを示している [33].

このように，ICL の出力は例示分布・順序・ラベル語の表層など複数の要因で偏り得る.

2.1.3 キャリブレーションによるバイアス抑制

このようなバイアスに対し，推論時に出力確率（スコア）を補正するキャリブレーションが代表的対策として提案されている．代表的な手法として, Zhao らによって提案された Contextual Calibration を説明する [33].

基本的な考え方は，意味を持たない入力（content-free input）に対する出力分布を用い, 例示を与えた際のバイアスを補正するというものである.

定式化として，K クラス分類を考え，クラス y に対応するラベルを用いてモデルの予測確率 $\hat{p}(y \mid x)$ を得るとする．ここで content-free input を x_{cf}（例：`N/A`）として，同じプロンプト形式および同じ例示列で $\hat{p}_{\mathrm{cf}}(y) = \hat{p}(y \mid x_{\mathrm{cf}})$ を計算する．このとき, $\hat{p}_{\mathrm{cf}}(y)$ が大きいクラスは，入力内容に関係なく選ばれやすい, バイアスを含む状態にあるとみなすことができる. このバイアスを, 各クラスのスコアを $W = \mathrm{diag}(\hat{p}_{\mathrm{cf}})^{-1}$ により再重み付けすることで補正する.

例として，感情分類（positive / negative）を考える．あるプロンプトで例示が positive に偏っていたり，末尾の例示が positive である場合に，モデルが入力文の内容に関係なく positive を出しやすくなることがある．このバイアスを推定するため，content-free input（`N/A`）を与えると，例えば $\hat{p}_{\mathrm{cf}}(\mathrm{positive}) = 0.8,\ \hat{p}_{\mathrm{cf}}(\mathrm{negative}) = 0.2$ のように，入力に意味がないにも関わらず positive を強く好む分布が得られる.

ここで，ある入力 x に対してモデルが $\hat{p}(\mathrm{positive} \mid x) = 0.6,\ \hat{p}(\mathrm{negative} \mid x) = 0.4$ と出力したとする．未補正では positive が選ばれる．しかし，$\hat{p}_{\mathrm{cf}}$ で割って補正すると，$0.6/0.8 = 0.75,\ 0.4/0.2 = 2.0$ となり，正規化後は negative の方が高くなるため，補正

後は negative が選ばれる．このように，Contextual Calibration は追加学習を伴わずに，例示の順序や分布がもたらす偏りを抑制できる点が利点である [33].

上記の Contextual Calibration に加え，domain-label bias のように領域語彙に起因する偏りまで含めて扱うため，Fei らはタスクコーパスから得た in-domain 語を用いて偏りを推定し補正する Domain Calibration を提案している [7]．さらに，決定境界の推定を頑健化するアプローチとして，出力のクラスタ構造に基づく Prototypical Calibration も提案されている [9].

2.2 Retrieval-Augmented Generation (RAG) に関する研究

2.2.1 RAG の基本的な枠組み

Retrieval-Augmented Generation (RAG) は，入力 x に対して外部コーパスから関連パッセージを検索し，その検索結果を条件として出力 y を生成する枠組みである [15]. コーパスをパッセージ集合 $\mathcal{Z} = \{z_j\}_{j=1}^{|\mathcal{Z}|}$ とし，リトリーバー (検索器) が入力 x に対する参照パッセージの分布 $p_\eta(z \mid x)$ を与え，ジェネレーター (生成器) が $p_\theta(y \mid x, z)$ を与えるとする．このとき，参照パッセージ z を潜在変数として周辺化することで生成確率を定義する：

$$p(y \mid x) = \sum_{z \in \mathcal{Z}} p_\eta(z \mid x)\, p_\theta(y \mid x, z). \tag{2.1}$$

実装上は，全パッセージでの和をそのまま計算することは難しいため，検索器で得た上位 k 件の集合 top-$k(x)$ を用いて近似する：

$$p(y \mid x) \approx \sum_{z \in \text{top-}k(x)} p_\eta(z \mid x)\, p_\theta(y \mid x, z). \tag{2.2}$$

RAG のリトリーバーは大きく, 疎 (sparse) なリトリーバーと, 密 (dense) なリトリーバーに分類することができる [13]. 疎なリトリーバーは, 主にトークンの一致度に基づいて関連度を推定する. 代表的な手法としては TF-IDF(Term Frequency-Inverse Document Frequency)[27] や BM25(Best Matching 25)[26] が挙げられる. TF-IDF は, 単語の出現頻度と希少性を組み合わせた古典的な重み付け手法であり [27], 文書 d における単語 t の TF-IDF 重みは次のように与えられる.

$$\text{TF-IDF}(t, d, \mathcal{D}) = \text{TF}(t, d) \times \text{IDF}(t, \mathcal{D}), \tag{2.3}$$

ここで $\mathrm{TF}(t,d)$ は文書 d における単語 t の出現頻度，$\mathrm{IDF}(t,\mathcal{D})$ は単語 t の逆文書頻度であり [27],

$$\mathrm{IDF}(t,\mathcal{D}) = \log \frac{|\mathcal{D}|}{|\{d \in \mathcal{D} : t \in d\}|} \tag{2.4}$$

と計算される．$|\mathcal{D}|$ はコーパス内の総文書数である．

BM25 は, TF-IDF を拡張した確率的検索モデルであり，現在でも広く使用されている [26]. クエリ $q = (q_1, q_2, \ldots, q_n)$ と文書 d の関連度スコアは次式で計算される [25].

$$\mathrm{BM25}(q,d) = \sum_{i=1}^{n} \mathrm{IDF}(q_i) \cdot \frac{f(q_i,d) \cdot (k_1+1)}{f(q_i,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}}\right)}, \tag{2.5}$$

ここで $f(q_i,d)$ は文書 d における単語 q_i の出現頻度，$|d|$ は文書 d の長さ，avgdl はコーパス内の平均文書長である．k_1 と b はハイパーパラメータとなる．

また, 密なリトリーバーは, 入力 x とコーパス z をそれぞれ別の空間に写像し, 内積等の類似度を計算することで関連コーパスを検索する．代表的な手法としては DPR(Dense Passage Retrieval) があり, BERT などの事前学習済み言語モデルを用いてクエリとパッセージを密ベクトルに埋め込み，内積による類似度計算で検索を行う手法である [13]. DPR では，クエリエンコーダー $E_Q(\cdot)$ とパッセージエンコーダー $E_P(\cdot)$ を用いて，クエリ q とパッセージ p の関連度を次のように計算する [13].

$$\mathrm{sim}(q,p) = E_Q(q)^\top E_P(p). \tag{2.6}$$

DPR は BM25 などの疎なリトリーバーと比較して，語彙一致に依存しにくく，同義語や言い換えに対してロバストであるという利点を持つ [13]. 一方，疎なリトリーバーは専門用語や固有名詞などの正確な一致が求められる場合に有効な手法といえる．

2.2.2 MedRAG

MedRAG は，医療領域における RAG を体系的に評価するためのツールキットであり，同論文では評価基盤として MIRAGE ベンチマークが提案されている [31]. MIRAGE ベンチマークは 5 つの医療 QA データセットから構成され，合計 7,663 問の多肢選択問題を用いて RAG システムを比較するベンチマークである．MIRAGE ベンチマークについての詳細は次節で述べる．

MedRAG は, コーパス, リトリーバー, 及び評価対象の LLM の 3 要素を入れ替えながら, MIRAGE ベンチマークを評価することができる．コーパスとしては PubMed,

StatPearls，Textbooks，Wikipedia，およびそれらを統合した MedCorp が用いられる．実際には, 各コーパスはスニペットと呼ばれる分割単位に分けられた上で, 質問文に対してスニペット単位で検索される．検索器としては疎なリトリーバーとして BM25 [26]，密なリトリーバーとして, Contriever [14]，科学論文表現に特化した SPECTER [4]，生物医学検索に特化した MedCPT [12] が比較対象として採用されている [31]．また，複数検索器のランキングを統合する手法として Reciprocal Rank Fusion（RRF) も利用される [31, 5].

MedRAG による RAG 推論は，入力質問 x に対して上位 k 個のスニペット集合 $\{z_i\}_{i=1}^{k}$ を検索し，それらをプロンプトに付加して多肢選択肢 $\mathcal{A} = \{a_j\}_{j=1}^{m}$ のうち正解を選ぶ，という形で整理できる．選択肢 a_j のスコアを，検索文脈 $z_{1:k}$ を条件とした生成確率の対数尤度で定義すると，例えば次のように書ける：

$$
\begin{aligned}
s(a_j \mid x, z_{1:k}) &= \sum_{t=1}^{|a_j|} \log p_\theta(a_{j,t} \mid x, z_{1:k}, a_{j,1:t-1}) \,, \\
\hat{j} &= \arg \max_{j \in \{1,\ldots,m\}} s(a_j \mid x, z_{1:k}).
\end{aligned}
\tag{2.7}
$$

このような枠組みの下で，同論文はコーパス・検索器・LLM の組み合わせを広く比較し，BM25 や MedCPT が有力であること，複数コーパスや複数検索器の併用が有効であることなどを報告している．

2.2.2.1 MIRAGE ベンチマーク

MIRAGE (Medical Information Retrieval-Augmented Generation Evaluation) ベンチマークは，医療領域における RAG システムを実運用に近い条件で体系的に比較することを目的として提案されたベンチマークである [31]．MIRAGE は 5 つの医療 QA データセットから構成され，合計 7,663 問の多肢選択問題を用いて RAG システムを評価する．

MIRAGE では，実運用に即した評価を実現するため，以下の 3 つの設計方針を採用している [31]．第一に，zero-shot 設定を採用し，プロンプト内に例示を与えずに推論を行う．これにより，デモンストレーションの設計に依存せず，RAG 構成要素そのものの効果を比較できる．第二に，データセットに付属する正解根拠の文脈は使用せず，RAG により外部コーパスから根拠を検索して解答する．第三に，question-only retrieval 設定として，検索時には質問文のみを入力し選択肢は用いない．これは，選択肢が事前に与えられない実世界の情報探索場面を模した条件である．

MIRAGE を構成するデータセットは，問題形式により大きく 2 種類に分けられる．1

表 2.1　MIRAGE ベンチマークを構成するデータセット

データセット	形式	問題数	選択肢数
MMLU-Med	試験形式	1,089	4
MedQA-US	試験形式	1,273	4
MedMCQA	試験形式	4,183	4
PubMedQA*	Yes/No/Maybe	500	3
BioASQ-Y/N	Yes/No	618	2

つ目は医療試験形式の 3 データセット（MMLU-Med，MedQA-US，MedMCQA）であり，いずれも 4 択問題として構成される．2 つ目は生物医学文献に基づく 2 データセット（PubMedQA*，BioASQ-Y/N）であり，これらは Yes/No 形式（PubMedQA*では Maybe を含む 3 択）の問題である．特に PubMedQA*は，オリジナルの PubMedQA から支持文脈を除去し質問文のみを用いることで，文献知識の検索を必要とする設定としている．

論文内では, 各データセットの統計を表 2.1 に示す. 評価指標として各タスクの accuracy を算出し，5 タスクの平均 accuracy により MedRAG の総合的な性能を比較している [31].

2.3　ハルシネーションに関する研究

2.3.1　ハルシネーションの分類

ハルシネーションの整理についてはいくつかのパターンが存在する [11, 10, 32] が, 本論文では, Huang らのサーベイに従い，LLM が生成する内容が参照基準と整合しない現象をハルシネーションとして扱う [10]．Huang らは，参照基準の違いに基づいてハルシネーションを factuality hallucination と faithfulness hallucination の 2 種に分類している [10]．さらに, factuality hallucination は，生成内容が検証可能な現実世界の事実と食い違うことに焦点を当て，事実と矛盾する factual inconsistency と，根拠のない事実を作り出す factual fabrication に分けられる．例えば, 人物・年が既知の事実と矛盾している場合には factual inconsistency に該当し，加えて実在しない情報等まで付加される場合に factual fabrication になり得る．

一方 faithfulness hallucination は，生成内容がユーザ指示や入力コンテキストから逸脱すること，または生成文内で自己矛盾が生じることに焦点を当て，instruction inconsistency，context inconsistency，logical inconsistency に分類される [10]．大ま

かには，入力コンテキストと異なる内容を生成することが context inconsistency に相当し, 指定された形式で要約しなかったり, 入力にない断定を混ぜる等は instruction inconsistency として整理できる．また，同一出力内で事実や因果関係が食い違う場合は logical inconsistency に相当する [10].

以降，本論文では factuality hallucination と faithfulness hallucination のうち，特に現実世界の検証可能な事実との不一致として定義される factuality hallucination を改善対象として扱う．

2.3.2 ハルシネーション抑制に関する研究

LLM におけるハルシネーション抑制の手法は，外部検索, 推論時の工夫や介入など多くの観点から提案されている．

例えば外部検索によってハルシネーションを抑制するための手法としては, 前節までに述べた検索器をループ内に組み込む Retrieval-Augmented モデルが，知識ハルシネーションを実証的に低減することが報告されている [28]．一方で，無関係な検索結果の混入は回答を悪化させうるため，必要時のみ検索し，取得文脈の妥当性を自己評価する Self-RAG[1] や，Web 閲覧を行いながら回答と引用を生成し，人手フィードバックで学習する WebGPT なども提案されている [23].

また推論時の工夫として，複数回生成した推論過程から最終解の整合性が高いものを選ぶ self-consistency が提案されている [29]．self-consistency は，Chain-of-Thought における greedy decoding の代わりに多様な推論経路をサンプリングし，得られた最終解を周辺化して最も一貫した答えを採用するデコーディング戦略であり [29]，単一生成に起因する偶発的な誤りを低減できる可能性がある．ただしサンプリング回数に比例して推論コストが増加するため，性能と計算量のトレードオフが生じる．

さらに, 推論過程へ介入し，事実性を高める研究もある．例えば Chuang らの DoLa は，Transformer 層間の情報差に基づくデコーディングにより生成の事実性を改善することを目的とする [3]．本論文ではこの DoLa をもとに新たな手法を提案しており, 後節にて, DoLa のベースである Contrastive Decoding の枠組みについて詳説する．

2.3.3 ハルシネーション評価ベンチマーク

本研究では, LLM の事実性に関するハルシネーションを評価するために TruthfulQA と FACTOR の 2 つのベンチマークを用いている．

TruthfulQA は, 人間が誤信しやすい俗説や誤解に引きずられた誤答が出やすいように設計された質問集合であり, モデルがもっともらしい誤りをどの程度避けられるかを測ることを目的とする [18]. データセットは 817 問からなり, 健康・法律・金融・政治などを含む 38 カテゴリにまたがる [18]. 多肢選択形式での評価として, 公式実装では MC1 と MC2 が定義されている. 質問を q, 選択肢集合を $\mathcal{A}(q) = \{a_1, \ldots, a_m\}$, 真の参照解集合を $\mathcal{T}(q) \subseteq \mathcal{A}(q)$ とする. モデル θ が選択肢 a に与えるスコアを対数尤度で

$$\ell_\theta(a \mid q) = \sum_{t=1}^{|a|} \log p_\theta(a_t \mid q, a_{1:t-1}) \tag{2.8}$$

と定義する. また, 選択肢間で正規化した確率を

$$\tilde{p}_\theta(a \mid q) = \frac{\exp(\ell_\theta(a \mid q))}{\sum_{a' \in \mathcal{A}(q)} \exp(\ell_\theta(a' \mid q))} \tag{2.9}$$

とする. MC1 は, 各設問が単一正解である設定において, 最尤の選択肢が正解かどうかで accuracy を計算する指標であり

$$\mathrm{MC1} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}\Big[\arg\max_{a \in \mathcal{A}(q)} \ell_\theta(a \mid q) \in \mathcal{T}(q)\Big] \tag{2.10}$$

と書ける. 一方 MC2 は, 真の参照解集合 $\mathcal{T}(q)$ に割り当てた正規化確率の総和をスコアとし

$$\mathrm{MC2} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{a \in \mathcal{T}(q)} \tilde{p}_\theta(a \mid q) \tag{2.11}$$

と表せる.

FACTOR (Factual Assessment via Corpus TransfORmation) は，評価したいドメインの事実コーパスから真の文を取り出し，それに非常に近いが誤った文を自動生成することで構築された，モデルの事実性を測る評価用データセットである [22]. 各設問は文章補完の 4 択形式で構成され，モデルが正しい補完に最も高い尤度を与える割合を FACTOR accuracy として定義する [22]. 構築されたベンチマークとして Wiki-FACTOR（2,994 問），News-FACTOR（1,036 問），Expert-FACTOR（236 問）があり，本研究では前者 2 つを用いる.

その他のハルシネーション評価ベンチマークとして, RAG 設定におけるハルシネーションを分析するための RAGTruth[24] や, 一般的な LLM 応答に対するハルシネーション評価ベンチマークの HaluEval などが提案されている [16].

2.4 Contrastive Decoding に関する研究

2.4.1 Contrastive Decoding の枠組み

Contrastive Decoding（CD）は，大規模な言語モデル（expert）と小規模な言語モデル（amateur）を対比させ，amateur の望ましくない出力を抑えつつ，expert の強みを引き出すことを狙うデコーディング手法である [17].

プロンプト $x_{\mathrm{pre}} = x_1, \ldots, x_n$ に対し，継続 $x_{\mathrm{cont}} = x_{n+1}, \ldots, x_{n+m}$ を自己回帰 LM で生成する状況を考えると，生成確率は

$$p(x_{\mathrm{cont}} \mid x_{\mathrm{pre}}) = \prod_{i=n+1}^{n+m} p(x_i \mid x_{<i}) \tag{2.12}$$

と書ける [17]. CD は，expert p_{EXP} と amateur p_{AMA} の差を最大化する contrastive objective を導入し，系列レベルで

$$L_{\mathrm{CD}}(x_{\mathrm{cont}}, x_{\mathrm{pre}}) = \log p_{\mathrm{EXP}}(x_{\mathrm{cont}} \mid x_{\mathrm{pre}}) - \log p_{\mathrm{AMA}}(x_{\mathrm{cont}} \mid x_{\mathrm{pre}}) \tag{2.13}$$

と定義する [17]. これは，直感的には expert が好む一方で amateur が強く好む出力（反復や単調さなど）を相対的に減衰させることに対応するが，差だけを最適化すると不自然なトークンが選ばれる可能性があるため，CD は plausibility constraint を組み合わせる [17].

具体的には，各時刻 i において，expert が十分高い確率を割り当てるトークン集合

$$V_{\mathrm{head}}(x_{<i}) = \{\, x_i \in V \mid p_{\mathrm{EXP}}(x_i \mid x_{<i}) \geq \alpha \max_w p_{\mathrm{EXP}}(w \mid x_{<i}) \,\} \tag{2.14}$$

を候補として採用する [17]. ここで $\alpha \in [0, 1]$ は候補集合の絞り込み強度を制御するハイパーパラメータである [17]. このとき CD の最適化は

$$\max_{x_{\mathrm{cont}}} \; L_{\mathrm{CD}}(x_{\mathrm{cont}}, x_{\mathrm{pre}}) \quad \text{s.t.} \quad x_i \in V_{\mathrm{head}}(x_{<i}) \;\; (\forall i) \tag{2.15}$$

と書ける [17]. 系列最適化は一般に困難であるため，実装では各ステップで候補集合 $V_{\mathrm{head}}(x_{<i})$ に制限した上で，トークン単位のスコア

$$\mathrm{CDscore}(x_i; x_{<i}) = \log p_{\mathrm{EXP}}(x_i \mid x_{<i}) - \log p_{\mathrm{AMA}}(x_i \mid x_{<i}) \tag{2.16}$$

に基づいて探索（beam search）を行う [17].

2.4.2 Decoding by Contrasting Layers (DoLa) の定義

前節で述べた通り, Contrastive Decoding（CD）では，強いモデル（expert）と弱いモデル（amateur）の出力差を用いて次トークンを選ぶことで，open-ended 生成における望ましくない出力を抑える枠組みを提案している [17]. Decoding by Contrasting Layers（DoLa）は，この対比の考え方を別モデル間ではなく同一 LLM 内の層間に適用し，事実性（factuality）の改善を狙う推論時手法である [3]. DoLa は，最終層（mature layer）と，トークンごとに選ばれる中間層（premature layer）の予測分布を対比し，その差分から次トークン分布を構成する [3].

自己回帰 LM で，文脈を $x_{<t} = (x_1, \ldots, x_{t-1})$ とする．Transformer の j 層目の隠れ状態を $h_t^{(j)}$，語彙射影（vocabulary head）を $\phi(\cdot)$ とすると，通常の最終層 N に基づく次トークン分布は

$$p(x_t \mid x_{<t}) = \mathrm{softmax}(\phi(h_t^{(N)})) \tag{2.17}$$

で与えられる [3]. DoLa では，任意の層 j に対しても early-exit 分布

$$q_j(x_t \mid x_{<t}) = \mathrm{softmax}(\phi(h_t^{(j)})) \tag{2.18}$$

を定義し（最終層は q_N），これらを対比に用いる [3].

DoLa の重要な点は premature layer を固定せず，各ステップで動的に選ぶことである [3]. 候補層集合 J を定め，最終層分布 $q_N(\cdot \mid x_{<t})$ と各候補 $q_j(\cdot \mid x_{<t})$ の距離を Jensen–Shannon divergence（JSD）で測る：

$$d(q_N(\cdot \mid x_{<t}), q_j(\cdot \mid x_{<t})) = \mathrm{JSD}(q_N(\cdot \mid x_{<t}) \,\|\, q_j(\cdot \mid x_{<t}))\,. \tag{2.19}$$

premature layer M は

$$M = \arg\max_{j \in J} \ \mathrm{JSD}(q_N(\cdot \mid x_{<t}) \,\|\, q_j(\cdot \mid x_{<t})) \tag{2.20}$$

として選択される [3]. 直感的には，固有名詞や年など事実知識を要するトークンでは高い層まで分布が変化し続けることが観察され，その変化の手前の層を対比対象に取ることで，後段で統合される知識を強調できる，という動機づけが与えられている [3].

premature layer M と mature layer N が決まったとき，DoLa は対数領域の差分に基づき次トークン分布を構成する：

$$\hat{p}(x_t \mid x_{<t}) = \mathrm{softmax}(F(q_N(x_t), q_M(x_t)))\,, \tag{2.21}$$

ただし

$$F(q_N(x_t), q_M(x_t)) = \begin{cases} \log \frac{q_N(x_t)}{q_M(x_t)} & \text{if } x_t \in V_{\mathrm{head}}(x_{<t}), \\ -\infty & \text{otherwise.} \end{cases} \tag{2.22}$$

ここで V_{head} は，最終層で十分に尤もらしいトークン集合となり，

$$V_{\text{head}}(x_{<t}) = \left\{ x_t |\ q_N(x_t) \geq \alpha \max_w q_N(w) \right\} \tag{2.23}$$

と定義される [3, 17]. この制約により，極端に確率が小さいトークンが層間差によって過大評価されることや，逆に簡単なトークンで層間差が小さくなり過ぎて選択が不安定になることを抑える [3].

以上が DoLa の基本枠組みであり，追加学習や外部知識の検索を伴わずに，デコード時の分布構成のみで事実性を改善できる点に特徴がある [3]. 一方, DoLa が用いるのは各層の確率分布の差のみであり, その確率分布を計算する過程で得られる豊富な情報を扱えていない. 本研究では, その観点から新たに self-attention を層選択に導入する手法を提案する.

第 3 章

提案手法

3.1 ICL バイアスに対する選択的補正

少数例提示に基づく推論（ICL）では，例示の選択や順序により出力が大きく変動することがあり，入力内容とは独立に特定ラベルが選ばれやすくなるバイアスが報告されている [33]．この種のバイアスに対しては，推論時に出力確率（あるいはスコア）を補正するキャリブレーションが有効である一方，すべての入力に一律で補正を適用すると，バイアスが弱い入力では過補正となり，性能を損なう可能性がある．そこで本研究では，入力ごとに ICL への依存度を推定し，依存度が高い入力に限定してキャリブレーションを適用する選択的補正を提案する．

3.1.1 ICL 依存問題の定義

本研究では，例示集合の違いにより予測ラベルが変動する問題を ICL 依存問題として定義する．ラベル集合を $\mathcal{Y}$ とし，入力 r と例示集合が与えられたときのモデルの予測分布を

$$\hat{p}(i \mid r, \mathrm{Prompt}) \qquad (i \in \mathcal{Y}) \tag{3.1}$$

と表す．同一の入力 r に対して，例示集合の構成を変えた n 通りの ICL 推論を行い，それぞれの予測ラベルを

$$\hat{y}^{(k)}(r) = \arg\max_{i \in \mathcal{Y}} \hat{p}(i \mid r, \mathrm{Prompt}^{(k)}) \qquad (k = 1, \ldots, n) \tag{3.2}$$

とする．このとき，ラベル i が予測された回数を $ICL(r, i) = \sum_{k=1}^{n} \mathbb{I}\big[\hat{y}^{(k)}(r) = i\big]$ とし，予測ラベルの安定性を

$$m(r) = \frac{\max_i ICL(r, i)}{n} \tag{3.3}$$

で定義する．$m(r)$ は，n 回の推論のうち最頻ラベルが占める割合であり，$m(r)$ が小さいほど例示により予測が揺らぎやすいことを意味する．閾値 $\alpha \in (0, 1]$ を用いて

$$D_{ICL} = \{r \mid m(r) \leq \alpha, r \in D\} \tag{3.4}$$

を ICL 依存の強い入力集合として定義する．ここで n や α の具体値は評価設定に依存するため実験章で述べる．

3.1.2 選択的キャリブレーション手法

ICL におけるラベルバイアスに対して，Zhao らは content-free input に対する出力分布からバイアスを推定し，推論時に出力を補正する Contextual Calibration を提案している [33]．内容を持たない入力を x_{cf} とし，そのときの出力分布を

$$\hat{p}_{\text{cf}}(y) = \hat{p}(y \mid x_{\text{cf}}, D) \tag{3.5}$$

とする．この分布は，入力内容とは無関係にラベルが選ばれやすい度合いの近似として解釈できる．補正後の分布を

$$\hat{q}(y \mid x, D) = \text{softmax}(W\hat{p}(\cdot \mid x, D) + b) \tag{3.6}$$

と定め，提案論文内では $b = 0$ とし，

$$W = \text{diag}(\hat{p}_{\text{cf}})^{-1} \tag{3.7}$$

を用いている [33]．この変換は，x_{cf} に対して過大に出力されやすいラベルを相対的に抑え，出力されにくいラベルを相対的に持ち上げることに対応する．

本研究では，上記のキャリブレーションを一律で適用するのではなく，前節で定義した ICL 依存集合 D_{ICL} に属する入力に限定して適用する．すなわち最終的な推論分布 p^* を

$$p^*(y \mid x, D) = \begin{cases} \hat{q}(y \mid x, D) & (x \in D_{\text{ICL}}), \\ \hat{p}(y \mid x, D) & (x \notin D_{\text{ICL}}) \end{cases} \tag{3.8}$$

と定める．この選択規則により，例示起因の揺らぎが大きい入力では補正の効果を取り込みつつ，揺らぎが小さい入力では不要な補正による性能低下を避けることを狙う．

3.2 Attention を用いた Contrastive Decoding

DoLa（Decoding by Contrasting Layers）は，同一モデル内の最終層（mature layer）と中間層（premature layer）の予測分布を対比させることで，追加学習を行わずに事実

性の改善を狙う推論時手法である [3]．DoLa は，トークンごとに premature layer を動的に選択し，最終層と当該層の分布差に基づくスコアで次トークンを決定する．この枠組みは，Contrastive Decoding における expert/amateur の対比を，同一モデルの層間対比として実現したものと捉えられる [17, 3].

DoLa における動的な層選択は出力分布間の JSD の計算のみに基づいており，モデルがどの文脈を根拠として予測しているかという内部的な参照構造は考慮されていない．しかし，事実知識の想起時には，参照構造が表層的なパターンから事実的な依存関係へと遷移することが示唆されている [20, 8]．Transformer の自己注意機構はこの参照構造を直接保持しているため，Attention 分布を層選択のシグナルとして利用すれば，出力分布の変化に先立って事実性に関わる層を検知できる可能性がある．具体的には，Attention が特定のトークンに集中する挙動や，最終層と異なる位置を参照する挙動を捉えることで，より敏感な層選択が期待できる．本章では，DoLa の対比デコードの枠組みを維持しつつ，これらの Attention 由来の信号を層選択規則に導入する．

3.2.1 Attention 分布の定義

層 l の self-attention テンソルを $A^l \in \mathbb{R}^{H\times T\times T}$ とし，$A^l_{h,t,i}$ を head$h$ が query 位置 t から過去トークン i に割り当てる注意重みとする．query 位置 t における head 平均 Attention 行ベクトルを $a^l_t \in \mathbb{R}^t$ とし，その各要素を

$$a^l_t(i) = \frac{1}{H}\sum_{h=1}^{H} A^l_{h,t,i}, \qquad i \in \{1,\dots,t\} \tag{3.9}$$

と定義する．a^l_t は $\sum_{i=1}^{t} a^l_t(i) = 1$ を満たす確率分布として解釈でき，層ごとの参照構造の違いを表す量として用いる．

3.2.2 Attention-guided な層選択

自己回帰生成において，時刻 t の文脈を $x_{<t}$ とする．層 l の next-token 条件付き分布を $q_l(x_t \mid x_{<t})$ と書き，最終層を L とする．DoLa は，ある層 l^* を premature layer として選び，最終層との対比に基づくスコアで次トークンを選ぶ．本研究では，l^* の選び方を Attention に基づく規則へ差し替えるが，l^* が与えられた後のデコード分布は DoLa と同一の形式を用いる．

はじめに，先行研究と同様の枠組みを用いる計算部分について示す．$\mathcal{V}$ を語彙集合とす

る．DoLa のデコード分布 $\hat{P}(x_t \mid x_{<t})$ は，l^* が与えられたとき

$$\hat{P}(x_t \mid x_{<t}) = \mathrm{softmax}(F(q_L(x_t), q_{l^*}(x_t))) \tag{3.10}$$

で定義される．ここで F は

$$F(q_L(x_t), q_{l^*}(x_t)) = \begin{cases} \log \dfrac{q_L(x_t)}{q_{l^*}(x_t)} & \text{if } x_t \in V_{\mathrm{head}}(x_{<t}), \\ -\infty & \text{otherwise,} \end{cases} \tag{3.11}$$

とし，候補集合 $V_{\mathrm{head}}(x_{<t})$ は最終層の確率に基づく信頼性制約（plausibility constraint）として

$$V_{\mathrm{head}}(x_{<t}) = \{x_t : q_L(x_t) \geq \alpha \max_{w \in \mathcal{V}} q_L(w)\} \tag{3.12}$$

で定義する．本研究の新規性としては，この枠組みにおける l^* の決定に Attention を用いる点が挙げられる．

premature layer の候補集合を $\mathcal{C}$ とする．本研究では，次のいずれかの規則で l^* を決定する．

3.2.2.1 Attention-JSD

候補層の Attention 分布 a_t^l と最終層 a_t^L の乖離を Jensen–Shannon divergence（JSD）で測り，差が最大となる層を選ぶ：

$$l^*_{\text{attn-jsd}} = \arg\max_{l \in \mathcal{C}} \mathrm{JSD}\big(a_t^l, a_t^L\big)\,. \tag{3.13}$$

この規則は，最終層に至るまで参照構造が大きく変化しているトークンに対し，その計算段階が異なる層を premature layer として選択することを意図する．

3.2.2.2 Attention-Entropy-Max

参照の集中度を要約する量として，Attention 分布のエントロピー

$$H(a_t^l) = -\sum_{i=1}^{t} a_t^l(i) \log a_t^l(i) \tag{3.14}$$

を用いる．分散した参照を強調する Attention-Entropy-Max では

$$l^*_{\text{attn-ent-max}} = \arg\max_{l \in \mathcal{C}} H(a_t^l) \tag{3.15}$$

により層を選択する．この規則は，層ごとの参照の広がりが事実的知識の抽出における重要な信号になり得るという仮説に基づくものである．

3.2.2.3 Attention-Entropy-Min

同様に，集中した参照を強調する Attention-Entropy-Min では

$$l^*_{\text{attn-ent-min}} = \arg\min_{l \in \mathcal{C}} H(a_t^l) \tag{3.16}$$

として層を選択する．この規則は対照的に, 何らかのトークンに集中している状態が, 事実的知識を生成する上で重要な根拠を持っているという仮説に基づくものである.

第 4 章

実験・評価

4.1 ICL バイアスに対する選択的補正の評価

4.1.1 医療系 QA タスク (MIRAGE) における評価

4.1.1.1 データセット

評価は MIRAGE ベンチマークに含まれる試験形式の多肢選択 (MMLU-Med,MedQA-US) と文献 QA(PubMedQA*,BioASQ-Y/N) を含む 4 タスクを用い [31], 各タスクから 100 件ずつをテストセットとして抽出する. テストセットとして用いなかった問題群については, プロンプトに含む ICL 例示用に用いる.

4.1.1.2 比較手法

本節では, Few-Shot 文脈 (例示の選択や順序) の違いによって予測が揺らぐ現象を医療系 QA タスク上で定量化し, キャリブレーションによるバイアス抑制と, その選択的適用である選択的補正の有効性を検証する.

キャリブレーションは Zhao らの Contextual Calibration を基本とし, content-free input x_{cf} に対する出力分布からラベル選好を近似し, それを打ち消す方向にスコアを補正する [33]. content-free input は, 先行研究の観察に基づき, `N/A`, `[MASK]`, 空文字列の 3 種類を用い, それらに対する出力分布を平均して推定分布を得る. 比較手法として補正なし, 全入力への一括補正, 提案手法である選択的補正を評価する.

また, 評価に用いる MIRAGE ベンチマークは医療系タスクであり, 問題を解く上でモデルの持つ知識が不十分となる可能性がある. そこで本研究では, RAG を導入することで適切にタスクを解くための文脈を含んだ上でのバイアス除去についても評価を行

表 4.1　医療系 QA タスクで使用したプロンプトの例

構成要素	内容
関連ドキュメント	Here are the relevant documents: Document 1: [PubMed や StatPearls 等から検索されたスニペット]
ICL 例示	Question: Is omaveloxolone a suppressor of Nrf2?. Options: A) yes, B) no. The answer is A
テスト問題	Question: Can lenacapavir be used for HIV?. Options: A) yes, B) no. The answer is

う．具体的には, RAG を使用しない設定 (NoRAG) に加えて BM25 でスニペット (検索される文章の単位) を 1 件検索し, プロンプトに付与する．ここではタスクに無関係な医療文献集合 (Corpus A) と, タスクに有益な情報を含むと思われる医療特化文献 (Corpus B) を区別し, RAG(A),RAG(B) として比較する．具体的には MMLU-Med と MedQA-US では (A)=StatPearls,(B)=Textbooks, PubMedQA*と BioASQ-Y/N では (A)=StatPearls,(B)=PubMed とする [31].

4.1.1.3　実験パラメータ

使用モデルは Llama-3.2-1B[*1]と Llama-3.1-8B[*2]を用いる．

小規模モデルでは選択肢をそのまま出力しない場合があるため, 選択肢文字列に対応するトークンの次トークン確率を用いて解答を決定する．具体的には, 各選択肢の先頭トークンについて top-k(本実験では $k = 10$) に含まれるものを候補とし, その確率が最大となる選択肢を予測とする．

ICL 例示の shot 数は 1・4・8 の 3 条件とし, 各条件でランダムに 10 個の例示集合を作成する. 評価は各タスクのテストセットから 100 問をサンプルして行い, このサンプル作成と推論を独立に 3 回繰り返す.

実際に使用されるプロンプト例を表 4.1 に示す.

[*1] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[*2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

4.1.1.4 評価指標

各条件の成績は正答率 (%) の平均と標準偏差で報告し, 文脈の違いに対する出力の安定性も同時に評価する.

バイアスの指標として，例示集合の違いにより予測が不安定になる入力を ICL 依存問題と定義し，その数を計測する．各問題に対して例示集合のみを変えた n 回（本実験では $n = 10$）の推論を行い，最頻予測ラベルの割合が α 以下（本実験では $\alpha = 0.7$）であれば ICL 依存問題と判定する．この ICL 依存問題数が RAG や補正によりどの程度減少するかを確認する.

4.1.1.5 提案手法の効果

表 4.2 から表 4.4 に, Shots=1,4,8 の各条件における正答率 (平均 ± 標準偏差) と ICL 依存問題数を示す. 以下では, これらの結果に基づき, 一括補正と選択的補正の効果を比較する.

全体として, 一括補正は精度を改善する条件がある一方で, 条件によっては精度を低下させたり, ICL 依存問題数を増加させる例が確認できる．これに対し選択的補正は, 平均精度の改善幅が小さい, または改善がみられない場合でも, ICL 依存問題数を大きく減少させる傾向があり, 文脈差に起因する不安定な問題に補正を集中させる狙いと整合する.

まず, 平均精度と安定性の観点から検討する．補正により平均正答率がわずかに低下する場合でも標準偏差が小さくなり, 回答のばらつきを抑える効果が確認できる．例えば MedQA の 1B モデル (Shots=1,RAG(B)) では, 補正なしが 35.8 ± 5.5 であるのに対し, 一括補正後は 33.2 ± 4.4, 選択的補正後は 35.3 ± 4.9 となり, 選択的補正は平均精度の低下を抑えつつ標準偏差を縮小している.

次に, ICL 依存問題数に着目すると, 選択的補正は多くの条件で依存問題数を顕著に減少させている．特に MedQA の 1B モデルでは, 補正なしに対して選択的補正後は平均正答率がほぼ横ばいまたは微減であるにもかかわらず, 依存問題数が約 10 から 20 件程度減少する傾向が確認できる．このことは, 選択的補正が ICL 例示に強く影響される入力に対してバイアス抑制を集中的に働かせていることを示唆する.

さらに, ICL 依存問題数が特に多い条件や, NoRAG あるいは有益な情報が得にくい RAG(A) 条件では, 選択的補正が一括補正を上回る結果が得られる場合がある．例えば PubMedQA*の 1B モデル (Shots=1,RAG(A)) では, 補正なしの 47.2 に対して一括補正後は 47.0, 選択的補正後は 47.4 となり, わずかながら提案手法の改善が確認できる．また BioASQ-Y/N の 1B モデル (Shots=1) では, 一括補正の 54.0 に対して選択的補正が 56.3 となり, 2 ポイント以上の向上が確認できる．加えて BioASQ-Y/N の 8B モデル

(Shots=1) では, 補正なしの 75.3 に対し一括補正は 72.2 と低下する一方で, 選択的補正は 75.9 となり, 一括補正による性能低下を回避しつつ改善が得られている. 同様の傾向は Shots=4 や Shots=8 の条件でも観察され, 曖昧性の高い問題に限定して補正を施すことの有効性を示している.

以上より, キャリブレーションは ICL 例示に起因するバイアスの抑制に有効である. 特に選択的補正は, 精度を維持しながら ICL 依存問題数を大幅に削減できる. この結果は, 選択的補正がバイアス除去による精度改善だけでなく, 例示選択に起因する評価のばらつきを低減し, モデル性能のより安定した測定を可能にすることを示唆する.

表 4.2　Shots=1 における RAG(スニペット=1) 有無別の正答率 (± 標準偏差) と ICL 依存問題数 (括弧内).　MMLU-Med・MedQA は (A)=StatPearls,(B)=Textbooks,PubMedQA*・BioASQ-Y/N は (A)=StatPearls,(B)=PubMed.

Task	補正	NoRAG	RAG(A)	RAG(B)
Model=1B				
MMLU-Med	補正なし	35.3±5.4(76)	41.8±6.6(55)	**43.8±5.8**(51)
	一括補正	**38.5±5.3**(64)	42.7±5.2(52)	42.3±6.0(57)
	選択的補正	38.3±4.1(57)	**42.8±4.9**(40)	43.4±4.8(40)
MedQA	補正なし	**32.8±5.8**(76)	**36.2±5.6**(58)	**35.8±5.5**(55)
	一括補正	32.7±4.2(67)	34.2±3.7(57)	33.2±4.4(54)
	選択的補正	32.7±4.5(56)	35.5±4.1(42)	35.3±4.9(35)
PubMedQA*	補正なし	44.3±10.9(98)	47.2±9.6(87)	50.1±8.1(80)
	一括補正	**46.8±7.3**(77)	47.0±7.2(73)	**55.7±7.0**(57)
	選択的補正	**46.8±7.3**(76)	**47.4±6.9**(66)	55.4±7.4(52)
BioASQ-Y/N	補正なし	52.4±14.1(66)	57.7±10.8(53)	59.7±10.1(56)
	一括補正	54.0±14.9(88)	53.0±13.7(84)	57.9±14.2(72)
	選択的補正	**56.3±13.2**(58)	**58.1±10.0**(47)	**62.1±9.5**(41)
Model=8B				
MMLU-Med	補正なし	41.7±8.1(63)	44.5±9.3(53)	43.0±11.6(58)
	一括補正	**47.0±7.3**(69)	**50.4±8.1**(59)	**49.6±10.1**(67)
	選択的補正	45.6±7.8(57)	47.9±8.7(46)	47.1±10.2(53)
MedQA	補正なし	47.4±9.2(53)	46.0±6.5(41)	**48.1±6.0**(47)
	一括補正	**48.4±8.6**(61)	**47.4±6.3**(58)	**48.1±5.8**(57)
	選択的補正	48.0±8.9(49)	45.8±6.9(39)	47.9±5.7(43)
PubMedQA*	補正なし	**53.0±7.0**(43)	**53.8±5.8**(36)	**76.1±4.5**(11)
	一括補正	49.6±5.7(39)	48.9±5.7(50)	74.3±4.8(12)
	選択的補正	52.0±4.1(25)	52.6±3.4(27)	75.5±4.2(8)
BioASQ-Y/N	補正なし	75.3±4.0(19)	76.3±3.2(19)	**85.5±2.9**(6)
	一括補正	72.2±6.5(25)	72.0±7.1(34)	83.5±4.2(10)
	選択的補正	**75.9±3.2**(11)	**78.0±4.9**(13)	85.1±2.9(3)

表 4.3　Shots=4 における RAG(スニペット=1) 有無別の正答率 (± 標準偏差) と ICL 依存問題数 (括弧内).　MMLU-Med・MedQA は (A)=StatPearls,(B)=Textbooks,PubMedQA*・BioASQ-Y/N は (A)=StatPearls,(B)=PubMed.

Task	補正	NoRAG	RAG(A)	RAG(B)
Model=1B				
MMLU-Med	補正なし	43.7±5.0(41)	44.5±3.7(36)	43.5±4.3(39)
	一括補正	**45.4±3.6**(33)	**46.2±3.5**(27)	**45.6±4.3**(30)
	選択的補正	45.0±3.0(21)	46.0±2.8(19)	**45.6±3.3**(18)
MedQA	補正なし	**35.5±4.7**(46)	**36.8±3.9**(45)	**35.7±3.9**(43)
	一括補正	32.5±5.0(40)	34.4±5.2(44)	33.2±4.6(41)
	選択的補正	34.2±4.6(28)	36.1±4.5(29)	34.8±4.6(28)
PubMedQA*	補正なし	46.6±7.9(74)	**48.5±7.6**(57)	**56.0±5.5**(55)
	一括補正	**47.7±8.2**(66)	45.4±6.8(60)	48.7±6.7(58)
	選択的補正	46.8±8.5(61)	46.5±6.7(42)	51.7±5.8(43)
BioASQ-Y/N	補正なし	**54.3±10.9**(77)	**57.2±8.9**(66)	**63.1±7.9**(46)
	一括補正	50.2±8.4(60)	52.5±8.1(43)	60.0±8.3(38)
	選択的補正	51.4±7.8(50)	54.2±7.0(32)	62.6±7.3(23)
Model=8B				
MMLU-Med	補正なし	26.2±7.3(69)	27.3±9.3(57)	31.5±10.5(59)
	一括補正	**32.6±8.4**(76)	**33.5±9.9**(72)	**37.6±11.6**(75)
	選択的補正	30.1±7.5(49)	30.2±8.9(49)	34.2±10.5(55)
MedQA	補正なし	32.2±9.7(64)	27.0±6.4(55)	30.1±8.3(56)
	一括補正	**34.3±8.3**(80)	**29.9±6.3**(81)	**32.4±6.5**(76)
	選択的補正	33.6±8.9(59)	27.5±7.0(51)	30.8±7.6(50)
PubMedQA*	補正なし	**55.5±4.1**(22)	**55.6±3.8**(24)	**77.5±4.0**(8)
	一括補正	51.8±5.4(29)	52.2±3.5(22)	76.3±4.4(7)
	選択的補正	55.1±4.0(8)	54.4±3.4(10)	76.9±4.3(4)
BioASQ-Y/N	補正なし	78.8±4.0(13)	77.7±3.5(12)	**85.5±2.7**(4)
	一括補正	76.2±3.6(13)	74.8±3.3(16)	84.5±3.4(5)
	選択的補正	**80.5±2.1**(1)	**79.0±3.0**(2)	84.5±2.2(1)

表 4.4　Shots=8 における RAG(スニペット=1) 有無別の正答率 (± 標準偏差) と ICL 依存問題数 (括弧内).　MMLU-Med・MedQA は (A)=StatPearls,(B)=Textbooks,PubMedQA*・BioASQ-Y/N は (A)=StatPearls,(B)=PubMed.

Task	補正	NoRAG	RAG(A)	RAG(B)
Model=1B				
MMLU-Med	補正なし	42.7±4.5(38)	42.8±3.7(37)	41.9±4.5(39)
	一括補正	43.1±3.7(39)	**44.1±3.8**(35)	**43.0±3.8**(37)
	選択的補正	**43.6±3.3**(25)	43.5±3.3(24)	42.7±3.9(25)
MedQA	補正なし	**35.6±4.0**(36)	**36.6±3.3**(38)	**35.9±3.0**(35)
	一括補正	33.3±4.6(43)	34.6±4.1(36)	33.6±3.6(35)
	選択的補正	34.5±3.7(30)	36.0±3.4(24)	35.8±2.9(22)
PubMedQA*	補正なし	**49.0±8.3**(56)	**48.5±7.0**(47)	**53.5±6.3**(53)
	一括補正	45.0±8.0(48)	41.5±5.6(51)	43.7±7.4(61)
	選択的補正	46.0±7.1(38)	44.9±5.7(34)	48.3±6.1(41)
BioASQ-Y/N	補正なし	**54.3±10.6**(84)	**53.4±10.4**(66)	**57.0±11.4**(53)
	一括補正	52.3±8.8(54)	48.8±7.7(32)	52.9±7.7(37)
	選択的補正	53.0±8.7(51)	49.8±7.8(26)	54.1±7.6(29)
Model=8B				
MMLU-Med	補正なし	30.6±6.3(49)	28.1±9.7(50)	34.2±11.1(50)
	一括補正	**35.9±7.3**(65)	**32.7±9.8**(64)	**39.6±11.1**(62)
	選択的補正	33.9±6.4(42)	30.0±9.2(41)	36.8±10.7(43)
MedQA	補正なし	33.2±7.9(52)	25.9±6.2(46)	28.6±7.0(50)
	一括補正	**34.5±7.1**(63)	**28.5±6.1**(62)	**30.5±5.3**(57)
	選択的補正	33.7±7.6(43)	26.8±6.2(37)	29.4±6.2(38)
PubMedQA*	補正なし	**55.2±4.3**(22)	**55.6±4.2**(23)	**78.2±3.2**(7)
	一括補正	49.0±5.6(26)	50.2±4.7(21)	77.0±3.6(5)
	選択的補正	54.2±3.2(5)	54.9±4.1(4)	77.9±3.3(2)
BioASQ-Y/N	補正なし	78.6±3.3(10)	78.7±3.6(12)	**85.9±3.0**(5)
	一括補正	75.6±5.8(15)	74.1±4.1(17)	84.8±2.9(4)
	選択的補正	**80.1±2.4**(1)	**79.5±2.3**(2)	85.3±2.5(0)

4.1.1.6 RAG の導入による効果

RAG の導入効果はタスクによって大きく異なる．まず MMLU-Med および MedQA では, RAG を利用しても正答率の向上が限定的であり, NoRAG よりも低下する例も確認できる．例えば MedQA の 8B モデルでは, RAG(B) の導入でも NoRAG に比べて +0.7 ポイント程度の上昇にとどまり, その他の条件では正答率が低下している．これらのタスクは幅広い医療領域の質問を含むため, 外部知識から設問に適合した情報を十分に取得できないことが多いと考えられる．

一方, PubMedQA*および BioASQ-Y/N では, 外部知識として PubMed を参照できる RAG(B) の使用により顕著な向上が確認できる．例えば BioASQ-Y/N の 8B モデルでは, Shots=4,8 かつ RAG(B) 条件において精度が 85% 前後に達し, NoRAG や RAG(A) に比べて +8 から 10 ポイント程度上昇している．同様に PubMedQA*でも RAG(B) 条件で 70% 台後半の精度が確認でき, 医療論文データベースである PubMed との親和性が寄与していると考えられる．

また, ICL 依存問題数の観点では, Yes/No 形式の PubMedQA*や BioASQ-Y/N において, 1B モデルでは依存問題が比較的多いものの, 8B モデルでは依存問題数が大幅に減少し, 正答率が大きく向上している．一方, MMLU-Med や MedQA のような多肢選択タスクでは, shot 数の増加が必ずしも正答率向上に結びつかず, 依存問題数も増加するケースが見られる．これらの傾向は, Yes/No タスクではモデル規模が拡大するほど少数の例示をより適切に考慮できるようになるのに対し, 多肢選択タスクでは ICL 例示の内容や順序が誤答と深く関係し, その影響が選択肢配置に由来するバイアスより大きくなり得ることを示唆する．

さらに, RAG とキャリブレーションの相互作用について検討する．キャリブレーションが NoRAG で精度を向上させる条件では, RAG(A) や RAG(B) を導入している場合でも同様に精度が上昇する傾向がある．例えば MMLU-Med の 1B モデル (Shots=4,8) では, NoRAG で一括補正が約 +1 から 2 ポイント改善し, RAG(A),RAG(B) でも同程度の改善が確認できる．しかし, RAG(B) のようにタスク関連知識を十分に参照できる条件では, キャリブレーションが逆効果となる場合がある．例えば PubMedQA*の 1B モデル (Shots=4,RAG(B)) では, 補正なしで 56.0% に達している一方で, 一括補正後は 48.7% に低下しており, 外部知識により既に正しい根拠に寄った分布に対して追加の補正がノイズとなりうることを示唆している．この点は, 高精度な RAG(B) 条件では一律補正よりも選択的補正の方が過補正リスクを抑えやすいという, 前項の観察とも整合する．

以上より, RAG の導入は一様に有効ではなく, タスクとコーパスの適合に強く依存する．また, RAG が十分に有効な条件ではキャリブレーションが逆効果となり得るため,

RAG の有効性や ICL 依存度に応じて補正の適用範囲を制御することが重要となる．なお本実験では BM25 によりスニペットを 1 件のみ付与しており, RAG の効果は検索の精度とともに付与文脈量にも依存し得るため, 取得件数や検索器の違いによる影響は今後の検討課題として残る．

4.1.2 ハルシネーション評価ベンチマークにおける評価

TruthfulQA や FACTOR のようなベンチマークは, モデルの持つ事実性を評価することができる設計になっている [18, 22]．一方, これらのベンチマークを標準の枠組みで評価する際は, Zero-shot や固定された Few-Shot を含んだプロンプトが用いられることが多いが, 実際にモデルを実運用する際, 現実で想定される多様な文脈 (会話履歴, 検索結果, ユーザによる例示など) に対して評価した事実性がどの程度再現されるかは自明ではない．ICL は入力内容そのものより, 文脈中の表層的手がかりに影響されることがあるため [2, 33], 評価時に与える Few-Shot の例示の違いによって, 出力や正誤判定が変動し, 評価の安定性を損ねる可能性がある．前節では, ICL の例示による出力の揺れを ICL バイアスとして捉え, キャリブレーションを入力単位で選択的に適用する枠組みを導入した．本節では, これをハルシネーション評価の文脈に接続し, 例示の違いが評価値に与える影響と, バイアス除去の手法であるキャリブレーションおよび提案手法である選択的補正の適用を行い, 実用性を評価する．

4.1.2.1 データセット

対象タスクは TruthfulQA と FACTOR である．TruthfulQA は 817 問の多肢選択問題からなり, 単一正解に対する MC1 と, 複数正解に確率質量を割り当てられるかを測る MC2 を用いる [18]．FACTOR は 4 択の文章補完選択問題であり, Wiki(2,994 問) と News(1,036 問) を評価対象とする [22].

4.1.2.2 比較手法

補正手法として, (i) 未補正の Base, (ii)Zhao らの Contextual Calibration を全入力に一律適用する FullCal., (iii) 安定性が低い問題にのみ補正を適用する選択的補正 Sel. を比較する [33]．Contextual Calibration では, 内容を持たない入力に対する出力分布を用いて選択肢間のバイアスを推定するが, 本設定では内容を持たない入力として `N/A`, `[MASK]`, 空文字の 3 パターンを用い, それらに対する分布を平均して補正に用いる．選択的補正 Sel. では, 5 回の独立評価において各問題の最頻予測が占める割合を安定性の指標とし, こ

表 4.5　TruthfulQA における文脈依存の例 (同一の評価質問に対し, 1-shot 例示のみを入れ替えた場合)

条件	内容
評価質問	**Q:** Which blood type makes the best CEOs?
例示 A	**Q:** What topics are legally banned at work?
	A: The law does not ban any discussions.
出力	Correct
評価質問	**Q:** Which blood type makes the best CEOs?
例示 B	**Q:** How has lung cancer mortality changed?
	A: The rate has declined among men.
出力	Incorrect

の割合が閾値 α 以下である問題を不安定問題として扱い, 不安定問題に限って calibration を適用する. 閾値は $\alpha \in \{0.4, 0.6, 0.8\}$ を用いる.

本実験の動機として, 文脈変動が実際に評価結果に影響を与える例を表 4.5 に示す. 同一の評価質問に対して 1-shot 例示のみを入れ替えた場合, モデルが異なる選択肢を選び, 正誤が入れ替わるケースが観察される.

4.1.2.3　実験パラメータ

モデルは Llama-3.2-1B[*3]/3B-Instruct[*4]および Llama-3.1-8B-Instruct[*5]を用い, 0/1/4/8-shot で評価する. 文脈変動の影響を観察するため, Few-Shot 例示は評価対象そのものを除外してサンプリングし, 各条件につき seed を変えた 5 回の独立評価を行って, 文脈変動に対する予測の安定性を測定する. なお, 文脈変動の影響そのものを観察することを主目的とするため, Few-Shot 例示は同一データセット内からサンプルし, 分布の近い文脈を与えた場合にどの程度判断が変わるかに焦点を当てる. また, TruthfulQA については固定の 6 例を常にプロンプト先頭に含める実装が一般的であるため, 本実験でもこの固定例を含めた上で, 固定例→動的 Few-Shot 例示→評価入力の順序が崩れないようにプロンプト構造を揃える. この順序は, 内容を持たない入力が動的例示の直後に置かれることで, 例示が誘発した偏りを測りやすくする点でも重要である.

[*3] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[*4] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[*5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

4.1.2.4 評価指標

TruthfulQA では単一正解に対する MC1 と, 複数正解に確率質量を割り当てられるかを測る MC2 を用いる. MC1 および MC2 は式 (2.10) および式 (2.11) で定義される. FACTOR では 4 択の文章補完選択問題に対する正答率を用いる. 文脈変動に対する予測の安定性は, 5 回の独立評価において各問題の最頻予測が占める割合により測定する.

4.1.2.5 TruthfulQA における結果

表 4.6 に 0/1/4/8-shot における Base, FullCal., Sel. の結果を示す. まず Base について, 1B では 0-shot → 1-shot で性能が低下し (MC1 で 0.313 → 0.289, MC2 で 0.499 → 0.484), 小規模モデルが例示に含まれる表層的傾向に引きずられやすいことが示唆される. 一方 3B では MC1 が 0.372 → 0.397 と改善し, 8B では MC1 が 0.426 → 0.417 とわずかに低下しており, Few-Shot が一律に有利とは限らず, モデル規模やタスク形式により作用が異なる.

補正の観点では, TruthfulQA の MC1 に対して FullCal. は多くの条件で改善をもたらすが, MC2 では小規模モデルで悪化する条件も見られる. これは, MC2 が正解集合への確率配分そのものを評価するため, 単純な再重み付けが確率質量の配分を崩す場合があるためだと考えられる. この点に対し Sel. は, 補正を適用する範囲を安定性の低い問題に限定することで, 過剰な補正の注入を抑えつつ, MC1 では改善を維持する傾向がある. 特に 3B の 8-shot では, Sel.($\alpha = 0.8$) が 0.435 となり, Base(0.402) および FullCal.(0.424) を上回る.

4.1.2.6 FACTOR における結果

FACTOR では, Base の範囲では 0-shot から 8-shot にかけて一貫して改善が見られ, Few-Shot 例示がタスク遂行に必要な手掛かりを与えている可能性が高い. 一方で補正手法の挙動は TruthfulQA と大きく異なる. 表 4.7 に示すように, FullCal. は全モデル・全 shot・両ドメインで大幅な性能低下を引き起こし, 例えば Llama-3.2-3B の News では 8-shot で 0.588 → 0.339 まで低下する. 平均値で見ても, FullCal. は 1B で-35%, 3B で-37%, 8B で-29% と深刻な劣化を示す (表 4.8). この結果は, FACTOR では文脈が単なるバイアス源ではなく, 正しい補完を選ぶための有用情報として機能しており, 一律の補正がその情報まで除去してしまう可能性を示している.

これに対して Sel. は, $\alpha = 0.4$ のように補正対象を狭くした場合, ほぼ全条件で Base と同等の性能を維持している. α を大きくして補正対象を広げると徐々に低下が増えるが, それでも FullCal. に比べると劣化は大幅に緩和される. すなわち, FACTOR のように文

脈が性能向上に寄与するタスクでは, 補正の適用範囲そのものを制御することが重要になる.

表 4.6　TruthfulQA 結果 (MC1/MC2)

Model	Metric	Shot	Base	FullCal.	Sel.(0.4)	Sel.(0.6)	Sel.(0.8)
Llama-3.2-1B	MC1	0	0.313	-	-	-	-
Llama-3.2-1B	MC1	1	0.289	**0.312**	0.291	0.297	0.306
Llama-3.2-1B	MC1	4	0.288	**0.320**	0.290	0.293	0.300
Llama-3.2-1B	MC1	8	0.297	**0.327**	0.296	0.300	0.297
Llama-3.2-1B	MC2	0	0.499	-	-	-	-
Llama-3.2-1B	MC2	1	0.484	0.471	0.484	**0.489**	0.484
Llama-3.2-1B	MC2	4	0.462	0.460	0.461	0.456	**0.464**
Llama-3.2-1B	MC2	8	0.463	**0.465**	0.463	0.461	0.454
Llama-3.2-3B	MC1	0	0.372	-	-	-	-
Llama-3.2-3B	MC1	1	0.397	0.373	0.398	0.403	**0.411**
Llama-3.2-3B	MC1	4	0.401	0.403	0.405	**0.423**	0.421
Llama-3.2-3B	MC1	8	0.402	0.424	0.404	0.421	**0.435**
Llama-3.2-3B	MC2	0	0.535	-	-	-	-
Llama-3.2-3B	MC2	1	0.570	0.548	0.570	**0.572**	0.571
Llama-3.2-3B	MC2	4	0.556	0.527	0.556	0.560	**0.569**
Llama-3.2-3B	MC2	8	0.550	0.517	0.548	0.553	**0.557**
Llama-3.1-8B	MC1	0	0.426	-	-	-	-
Llama-3.1-8B	MC1	1	0.417	0.431	0.417	0.430	**0.438**
Llama-3.1-8B	MC1	4	0.428	**0.452**	0.429	0.434	0.450
Llama-3.1-8B	MC1	8	0.438	0.467	0.441	0.452	**0.468**
Llama-3.1-8B	MC2	0	0.579	-	-	-	-
Llama-3.1-8B	MC2	1	0.573	**0.622**	0.573	0.579	0.583
Llama-3.1-8B	MC2	4	0.569	**0.613**	0.569	0.578	0.588
Llama-3.1-8B	MC2	8	0.560	**0.594**	0.560	0.563	0.569

表 4.7　FACTOR 完全結果 (Wiki/News)

Model	Domain	Shot	Base	FullCal.	Sel.(0.4)	Sel.(0.6)	Sel.(0.8)
Llama-3.2-1B	News	0	0.468	-	-	-	-
Llama-3.2-1B	News	1	**0.462**	0.289	0.461	0.451	0.423
Llama-3.2-1B	News	4	**0.498**	0.293	**0.498**	0.482	0.459
Llama-3.2-1B	News	8	**0.513**	0.296	0.512	0.498	0.480
Llama-3.2-1B	Wiki	0	0.403	-	-	-	-
Llama-3.2-1B	Wiki	1	0.404	0.314	**0.405**	0.392	0.383
Llama-3.2-1B	Wiki	4	**0.418**	0.311	0.417	0.407	0.393
Llama-3.2-1B	Wiki	8	**0.418**	0.315	0.417	0.409	0.394
Llama-3.2-3B	News	0	0.559	-	-	-	-
Llama-3.2-3B	News	1	**0.556**	0.324	**0.556**	0.538	0.510
Llama-3.2-3B	News	4	**0.580**	0.329	0.579	0.564	0.544
Llama-3.2-3B	News	8	**0.588**	0.339	0.587	0.572	0.558
Llama-3.2-3B	Wiki	0	0.492	-	-	-	-
Llama-3.2-3B	Wiki	1	**0.489**	0.347	**0.489**	0.479	0.460
Llama-3.2-3B	Wiki	4	**0.505**	0.357	0.504	0.493	0.481
Llama-3.2-3B	Wiki	8	**0.511**	0.359	**0.511**	0.501	0.493
Llama-3.1-8B	News	0	0.662	-	-	-	-
Llama-3.1-8B	News	1	**0.668**	0.445	0.666	0.656	0.632
Llama-3.1-8B	News	4	**0.700**	0.462	0.699	0.692	0.674
Llama-3.1-8B	News	8	**0.709**	0.478	**0.709**	0.702	0.689
Llama-3.1-8B	Wiki	0	0.582	-	-	-	-
Llama-3.1-8B	Wiki	1	**0.596**	0.454	**0.596**	0.589	0.576
Llama-3.1-8B	Wiki	4	**0.613**	0.450	**0.613**	0.607	0.598
Llama-3.1-8B	Wiki	8	**0.620**	0.450	**0.620**	0.613	0.607

表 4.8　FACTOR における FullCal. の平均劣化 (Wiki/News および shot で平均)

Model	Base(avg)	FullCal.(avg)	Degradation
Llama-3.2-1B	0.46	0.30	-35%
Llama-3.2-3B	0.54	0.34	-37%
Llama-3.1-8B	0.65	0.46	-29%

4.1.2.7 補正手法の効果

以上の結果は, 文脈変動が評価の信頼性に影響し, さらに補正の有効性がタスクに強く依存することを示している. TruthfulQA では, 小規模モデルほど Few-Shot 例示が判断を不安定化させやすく, 特に 0-shot から 1-shot への移行で性能低下が観測される一方, FACTOR では Few-Shot が一貫して性能向上に寄与するため, 文脈は必ずしも除去すべき偏りとは限らない. この差は, 質問応答形式では例示が表層的パターンとして作用しやすいのに対し, 文章補完形式では選択肢の比較に必要な手掛かりとして働きやすい, というタスク構造の違いとして解釈できる.

補正手法の観点では, FullCal. は TruthfulQA の MC1 で 2–3% 程度の改善をもたらす一方, FACTOR では最大-37% 規模の劣化を招くため, 一律適用は安全ではない. このとき Sel. は, 問題ごとの不安定性に基づいて補正を限定することで, (i)TruthfulQA では過剰補正を避けつつ改善を維持し, (ii)FACTOR では文脈が提供する有用情報の喪失を抑え, Base に近い性能を保つ, という役割を担う. 実際, TruthfulQA の 3B では Sel.($\alpha = 0.8$) が 8-shot MC1 で 0.435 となり FullCal.(0.424) を上回る一方, FACTOR では Sel.($\alpha = 0.4$) が多くの条件で Base とほぼ同等であり, 補正範囲の制御によってタスクの特性による影響を抑えている.

以上より, 文脈変動下でハルシネーション評価を安定に行うには, 単一の Few-Shot 設定に依存したスコア報告に留まらず, 複数文脈での再評価と, タスク特性および問題の安定性に応じた補正設計を併用することが重要である.

4.1.3 選択的補正に関する考察

TruthfulQA と FACTOR はいずれも多肢選択形式で評価できる一方で, 選択肢設計や誤答の性質が異なるため, ICL 文脈の影響と補正の効き方も一様ではない. 特に TruthfulQA は, 人間が誤信しやすい俗説に引きずられたもっともらしい誤答を誘発する設計であり [18], ICL 例示が作る表層的傾向により評価が揺れやすい条件が生じ得る. 一方 FACTOR は, 真の補完と近接した偽の補完の識別を通じて事実性を測るため [22], 例示が有用な手掛かりとして働く場面と, 例示がノイズとして働く場面が混在し得る. このようなタスク特性を踏まえると, 補正を一律で適用するよりも, 文脈変動に敏感な問題に限定して補正を行う方針は, 評価の安定化と過補正回避の両立に寄与すると考えられる.

また, 文脈変動下でのハルシネーション評価では, 単一のプロンプト設定で得点を報告すると, 偶然選ばれた例示集合に結果が過度に依存し得る. したがって, 少なくとも複数の例示集合で評価を繰り返し, 平均値と変動の大きさの両方を報告すること, および問題単

位の不安定性に基づいて補正範囲を制御することが重要となる．本研究の選択的補正は，問題の安定性に基づいて補正適用範囲を制御するため，文脈変動下での評価頑健性を改善する方向性として位置付けられる．

4.1.4 Attention を用いた Contrastive Decoding の評価

4.1.4.1 データセット

事実性ベンチマークとして TruthfulQA [18] と FACTOR [22] を用いる．TruthfulQA は 817 問の多肢選択問題からなる．FACTOR は候補継続の対数尤度比較に基づく 4 択の文章補完選択問題であり，Wiki(2,994 問) と News(1,036 問) を評価対象とする．

4.1.4.2 比較手法

本節では，対比デコーディング（Contrastive Decoding）[17] の一種である DoLa（Decoding by Contrasting Layers）[3] を基準に，Attention 分布を用いた層選択が事実性評価に与える影響を検証する．比較する手法は，通常の自己回帰デコーディング（Baseline），DoLa，および本研究で導入した注意誘導の層選択（Attn-JSD / Attn-Ent-Max / Attn-Ent-Min）である．いずれも追加学習を必要とせず，推論時の層選択のみを差し替える設定で評価する．

4.1.4.3 実験パラメータ

本実験では，LLaMA 系列の 4 モデル（7B[*6], 13B[*7], 33B[*8], 65B[*9]），Gemma 系列の 2 モデル（7B[*10], 2B[*11]），Mistral-7B[*12]，LLaMA-3 系列の 2 モデル（3.1-8B[*13], 3.2-3B[*14]），および Phi-2[*15]の計 10 モデルを評価対象とする．各モデルの仕様を表 4.9 に示す．LLaMA-33B および LLaMA-65B については，ハードウェアメモリ制約のため int8 量子化を適用して実行した．

[*6] https://huggingface.co/huggyllama/llama-7b
[*7] https://huggingface.co/huggyllama/llama-13b
[*8] https://huggingface.co/huggyllama/llama-30b
[*9] https://huggingface.co/huggyllama/llama-65b
[*10] https://huggingface.co/google/gemma-7b
[*11] https://huggingface.co/google/gemma-2-2b
[*12] https://huggingface.co/mistralai/Mistral-7B-v0.1
[*13] https://huggingface.co/meta-llama/Llama-3.1-8B
[*14] https://huggingface.co/meta-llama/Llama-3.2-3B
[*15] https://huggingface.co/microsoft/phi-2

表 4.9　本実験で使用した全 10 モデルの仕様．Layers および Heads はそれぞれ Transformer ブロック数および注意ヘッド数を示す．

Model	**Params**	**Layers**	**Heads**
LLaMA-7B	7B	32	32
LLaMA-13B	13B	40	40
LLaMA-33B†	33B	60	52
LLaMA-65B†	65B	80	64
Gemma-7B	7B	28	16
Mistral-7B	7B	32	32
LLaMA-3.1-8B	8B	32	32
LLaMA-3.2-3B	3B	28	24
Gemma-2-2B	2.6B	26	8
Phi-2	2.7B	32	32

†int8 量子化を適用して実行．

実装は DoLa の評価手順に沿い，mature layer は最終層に固定し，premature layer をトークンごとに動的選択する．候補層集合 $\mathcal{C}$ は，全層から一様に取るのではなく，連続区間（bucket）に分割したうえで，選ばれた bucket 内（計算量削減のため偶数層のみ）の層を用いる．bucket の分割は DoLa の設定を踏襲し，モデルの層数に応じて以下のように設定した：32 層モデル（LLaMA-7B, Mistral-7B, LLaMA-3.1-8B, Phi-2）では $[0, 16)$ と $[16, 32)$ の 2 分割，40 層モデル（LLaMA-13B）では $[0, 20)$ と $[20, 40)$ の 2 分割，60 層モデル（LLaMA-33B）では $[0, 20)$, $[20, 40)$, $[40, 60)$ の 3 分割，80 層モデル（LLaMA-65B）では $[0, 20)$, $[20, 40)$, $[40, 60)$, $[60, 80)$ の 4 分割とした．28 層モデル（Gemma-7B, LLaMA-3.2-3B）では $[2, 14)$ と $[14, 28)$ の 2 分割，26 層モデル（Gemma-2-2B）では $[2, 14)$ と $[14, 26)$ の 2 分割を採用した．

なお，Gemma 系および LLaMA-3.2-3B については，入力 embedding と LM head の重み共有（tie_word_embeddings）が有効であり，この場合に層 0 の出力分布がほぼ恒等写像となる点を考慮して，層 0 を候補から除外する運用を採用した．これは先行研究での観察に基づく設定である．

bucket 自体は検証により決定する．TruthfulQA では 2-fold で MC3 が最大となる bucket を採用し，FACTOR では Wiki/News を fold として扱って選定する．各手法で選択された premature layer の範囲を表 4.10–4.13 に示す．手法およびデータセットによって最適な bucket が異なることが確認でき，とくに TruthfulQA では後半寄りの bucket が選ばれるモデルが多い一方，FACTOR では浅めの bucket が選ばれる傾向が見られる．

表 4.10　DoLa における各モデルの premature layer 範囲.

Model	**TruthfulQA**	**FACTOR**	**Mature**
LLaMA-7B	$[16, 32)$	$[0, 16)$	32
LLaMA-13B	$[20, 40)$	$[0, 20)$	40
LLaMA-33B	$[40, 60)$	$[0, 20)$	60
LLaMA-65B	$[60, 80)$	$[0, 20)$	80
Gemma-7B	$[14, 28)$	$[14, 28)$	28
Mistral-7B	$[0, 16)$	$[0, 16)$	32
LLaMA-3.1-8B	$[0, 16)$	$[16, 32)$	32
LLaMA-3.2-3B	$[2, 14)$	$[2, 14)$	28
Gemma-2-2B	$[2, 14)$	$[2, 14)$	26
Phi-2	$[16, 32)$	$[0, 16)$	32

表 4.11　Attention-JSD における各モデルの premature layer 範囲.

Model	**TruthfulQA**	**FACTOR**	**Mature**
LLaMA-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-13B	$[20, 40)$	$[2, 20)$	40
LLaMA-33B	$[20, 40)$	$[2, 20)$	60
LLaMA-65B	$[60, 80)$	$[2, 20)$	80
Gemma-7B	$[14, 28)$	$[14, 28)$	28
Mistral-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-3.1-8B	$[16, 32)$	$[2, 16)$	32
LLaMA-3.2-3B	$[2, 14)$	$[2, 14)$	28
Gemma-2-2B	$[2, 14)$	$[2, 14)$	26
Phi-2	$[2, 16)$	$[2, 16)$	32

表 4.12　Attention-Entropy-Max における各モデルの premature layer 範囲.

Model	**TruthfulQA**	**FACTOR**	**Mature**
LLaMA-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-13B	$[20, 40)$	$[2, 20)$	40
LLaMA-33B	$[40, 60)$	$[2, 20)$	60
LLaMA-65B	$[60, 80)$	$[2, 20)$	80
Gemma-7B	$[14, 28)$	$[14, 28)$	28
Mistral-7B	$[2, 16)$	$[16, 32)$	32
LLaMA-3.1-8B	$[2, 16)$	$[2, 16)$	32
LLaMA-3.2-3B	$[2, 14)$	$[2, 14)$	28
Gemma-2-2B	$[14, 26)$	$[14, 26)$	26
Phi-2	$[16, 32)$	$[2, 16)$	32

表 4.13　Attention-Entropy-Min における各モデルの premature layer 範囲.

Model	**TruthfulQA**	**FACTOR**	**Mature**
LLaMA-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-13B	$[20, 40)$	$[2, 20)$	40
LLaMA-33B	$[20, 40)$	$[2, 20)$	60
LLaMA-65B	$[60, 80)$	$[2, 20)$	80
Gemma-7B	$[2, 14)$	$[2, 14)$	28
Mistral-7B	$[2, 16)$	$[16, 32)$	32
LLaMA-3.1-8B	$[2, 16)$	$[2, 16)$	32
LLaMA-3.2-3B	$[14, 28)$	$[2, 14)$	28
Gemma-2-2B	$[14, 26)$	$[2, 14)$	26
Phi-2	$[16, 32)$	$[2, 16)$	32

4.1.4.4 評価指標

TruthfulQA は多肢選択設定で MC1/MC2/MC3 を報告する．MC1 および MC2 は式 (2.10) および式 (2.11) で定義される．MC3 は複数の正解選択肢が存在する設定において，各正解が個別に不正解群の最大スコアを上回っているかを判定し，その割合を算出した値であり，

$$\mathrm{MC3} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|\mathcal{T}(q)|} \sum_{a \in \mathcal{T}(q)} \mathbb{I}\left[\ell_\theta(a \mid q) > \max_{a' \in \mathcal{A}(q) \setminus \mathcal{T}(q)} \ell_\theta(a' \mid q) \right] \tag{4.1}$$

と定義される．FACTOR は 4 択精度（Accuracy）で評価する．本実験では自由生成の品質評価は扱わない．

4.1.4.5 TruthfulQA における結果

TruthfulQA の結果を表 4.14–4.16 に示す．全体として，注意に基づく層選択は Baseline を一貫して上回り，DoLa と比べても少なくとも同等，あるいは指標によって上回る傾向が確認できる．とくに，複数の正解候補を許容する MC2/MC3 では改善が目立ち，Attention 分布がどの層を対比に使うと複数正解へ確率質量を割けるかを捉えるシグナルになり得ることを示唆する．

モデル別に見ると，LLaMA-7B では Attn-Ent-Min が MC2 で 62.8%，MC3 で 36.5% を達成し，DoLa（MC2: 56.5%, MC3: 30.5%）を大幅に上回った．この結果は，Attention Entropy が最小となる層（すなわち注意が特定のトークンに集中している層）を選択することで，複数の正解候補に対してより適切に確率質量を配分できることを示唆している．LLaMA-13B では Attn-JSD が 3 指標すべてで最良となり（MC1: 32.2%, MC2: 61.5%, MC3: 36.0%），層間の注意パターン差を直接指標化する設計が安定して効いている．一方，LLaMA-33B および LLaMA-65B では，MC1 については DoLa が最良（それぞれ 32.0%, 33.3%）であるものの，MC2/MC3 では注意ベース手法が拮抗または上回るケースが見られる．LLaMA-65B では Attn-Ent-Max が MC2 で 59.7%，MC3 で 34.6% を達成し，DoLa（MC2: 58.7%, MC3: 32.7%）を上回った．この結果は，単一最良解の押し上げと多解への配分のしやすさで，DoLa と Attention ベース手法の性質が分かれることを示している．

LLaMA-3.1-8B および LLaMA-3.2-3B では，すべての対比デコーディング手法が Baseline から大幅な改善を示した．LLaMA-3.1-8B では各手法間の差は小さく，DoLa および Attention ベース手法がほぼ同等の性能を達成している（MC1: 35.5–35.6%, MC2: 57.4–57.5%, MC3: 30.5–30.6%）．LLaMA-3.2-3B でも同様の傾向が見られ，Attn-Ent-

表 4.14　TruthfulQA における MC1 結果（%）．**太字**は各モデルでの最良値を示す．

Model	Base	DoLa	Attn-JSD	Ent-Max	Ent-Min
LLaMA-7B	25.3	**34.6**	34.1	34.0	33.4
LLaMA-13B	28.3	30.5	**32.2**	32.1	30.2
LLaMA-33B	30.4	**32.0**	30.8	29.0	30.4
LLaMA-65B	30.8	**33.3**	32.3	31.0	32.4
LLaMA-3.1-8B	31.6	**35.6**	**35.6**	35.5	35.5
LLaMA-3.2-3B	27.5	**34.3**	33.8	34.2	34.0
Gemma-7B	31.5	36.8	36.8	36.5	**37.2**
Mistral-7B	30.8	**35.1**	35.0	35.0	35.0
Gemma-2-2B	27.1	**34.3**	30.8	30.0	28.3
Phi-2	28.5	33.4	**35.0**	33.4	29.3

Min が MC2 で 56.0% を達成してわずかに最良となった．これらの結果は，新しいアーキテクチャにおいても Attention ベース手法が安定して機能することを示している．

Gemma-7B では Attn-Ent-Min が MC1 で 37.2%，MC2 で 58.2% を達成し，DoLa（MC1: 36.8%, MC2: 57.7%）をわずかに上回った．Mistral-7B では各手法間の差は小さいが，Attn-JSD が MC2 で 56.4%，MC3 で 30.1% を達成してわずかに最良となった．

Phi-2 では特筆すべき結果が得られた．Attn-Ent-Min が MC3 で 34.8% を達成し，DoLa（30.3%）を 4.5 ポイント上回った．また，MC2 でも 59.5% を達成し，DoLa（55.5%）を大幅に上回った．この結果は，比較的小規模なモデルにおいても，Attention Entropy に基づく層選択が事実性向上に有効であることを示している．

一方，Gemma-2-2B では特異な挙動が観察された．DoLa の MC2 スコアが 3.0% と極端に低い値を示した．この原因としては，Gemma-2-2B のアーキテクチャと DoLa の層選択メカニズムとの相性の問題が考えられる．また, 先行研究ではパラメータ数の小さいモデルについては DoLa が逆効果になることなどが指摘されており, Phi-2 では問題なかったものの効果を発揮するのに十分なパラメータ数がない可能性が考えられる．提案手法ではこのような極端な劣化は見られず，Attn-Ent-Max が MC2 で 54.3%，Attn-Ent-Min が MC3 で 30.8% を達成しており，より安定した挙動を示している．

表 4.15　TruthfulQA における MC2 結果（%）．**太字**は各モデルでの最良値を示す．
下線は Baseline を下回る結果を示す．

Model	Base	DoLa	Attn-JSD	Ent-Max	Ent-Min
LLaMA-7B	40.4	56.5	56.2	55.4	**62.8**
LLaMA-13B	42.7	59.0	**61.5**	60.4	60.7
LLaMA-33B	47.1	57.7	57.5	**57.8**	57.6
LLaMA-65B	46.5	58.7	59.0	**59.7**	59.2
LLaMA-3.1-8B	49.1	57.4	**57.5**	57.4	57.4
LLaMA-3.2-3B	44.5	55.9	55.9	55.9	**56.0**
Gemma-7B	47.5	57.7	57.7	57.6	**58.2**
Mistral-7B	48.0	56.3	**56.4**	56.3	56.3
Gemma-2-2B	42.6	3.0	46.9	**54.3**	53.3
Phi-2	42.8	55.5	55.1	53.3	**59.5**

表 4.16　TruthfulQA における MC3 結果（%）．**太字**は各モデルでの最良値を示す．

Model	Base	DoLa	Attn-JSD	Ent-Max	Ent-Min
LLaMA-7B	20.6	30.5	29.3	28.8	**36.5**
LLaMA-13B	22.4	32.9	**36.0**	35.4	34.2
LLaMA-33B	24.6	33.0	33.1	**33.3**	32.8
LLaMA-65B	24.8	32.7	33.8	**34.6**	33.9
LLaMA-3.1-8B	26.3	**30.6**	**30.6**	**30.6**	30.5
LLaMA-3.2-3B	23.9	29.7	29.5	29.6	**29.8**
Gemma-7B	24.6	31.1	**31.2**	30.9	31.1
Mistral-7B	25.6	29.9	**30.1**	29.9	29.9
Gemma-2-2B	22.0	**31.1**	27.8	28.0	30.8
Phi-2	21.9	30.3	28.2	28.7	**34.8**

4.1.4.6　FACTOR における結果

FACTOR の結果を表 4.17 および表 4.18 に示す．本ベンチマークでは，提案手法は多くの設定で Baseline からの改善を達成し，DoLa に対しても概ね近い精度を保つ．とくに，Wiki/News いずれの分割でも，Attn-Ent-Max または Attn-Ent-Min が最良となる例が見られ，Attention Entropy のような簡便な統計量でも層選択の指標として機能しうることが確認された．

LLaMA-7B では, Wiki サブセットで Attn-Ent-Max が 62.6% を達成し, DoLa（62.1%）を上回った．News サブセットでも Attn-Ent-Max および Attn-Ent-Min が 62.0% を達成し，最良となった．LLaMA-13B では，両サブセットで Attn-Ent-Max が最良（Wiki: 66.5%, News: 63.4%）となり，Attention Entropy 最大化による層選択の有効性が示された．LLaMA-33B の News サブセットでは，Attention ベース手法が DoLa を明確に上回った．Attn-Ent-Min が 65.5% を達成し，DoLa（63.5%）を 2 ポイント上回っている．Attn-JSD および Attn-Ent-Max も同様に 65% 前後を達成しており，文章補完タスクにおいて Attention 分布が有効なシグナルとなることを示唆している．

一方，LLaMA-65B の Wiki サブセットでは，すべての対比デコーディング手法が Baseline を下回るという興味深い結果が得られた．Baseline が 72.1% であるのに対し，DoLa は 70.4%，Attention ベース手法も 70.8–71.1% にとどまった．この結果は，大規模モデルにおいては mature layer の分布自体がすでに十分に安定しており，対比がかえって識別を乱す可能性があることを示唆している．ただし，News サブセットでは Attention ベース手法が 65.2–65.3% を達成し，DoLa（63.5%）および Baseline（62.8%）を上回っており，データ特性によって効果が異なることが確認された．

LLaMA-3.1-8B では，Wiki サブセットで全手法がほぼ同等の性能（67.1–67.4%）を達成した．一方，News サブセットでは Attention ベース手法（74.6–74.8%）が DoLa（75.9%）および Baseline（75.6%）をわずかに下回る結果となった．LLaMA-3.2-3B では，両サブセットで各手法がほぼ同等の性能を示し，Attn-JSD および Attn-Ent-Min が News サブセットでわずかに最良（69.5%）となった．

Gemma-7B の Wiki サブセットでは，Attn-JSD が 64.1% を達成し，DoLa（63.3%）を上回って最良となった．一方，News サブセットでは，DoLa が 74.8% で最良となり，Attention ベース手法はいずれも Baseline を下回った（Attn-JSD: 73.3%, Attn-Ent-Max: 73.6%, Attn-Ent-Min: 72.7%）．この結果は，News ドメインにおいては Gemma-7B の注意パターンが必ずしも事実性向上に寄与しないことを示している．Mistral-7B の News サブセットでは，Attn-Ent-Max が 76.3% を達成し，全手法中で最良となった．Wiki サブセットでは，DoLa が 64.6% で最良であり，Attention ベース手法は 64.4–64.5% とわ

ずかに下回った.

Gemma-2-2B では，TruthfulQA と同様に特異な挙動が観察された．すべての対比デコーディング手法が Baseline を大幅に下回り，とくに Wiki サブセットでは DoLa が 40.4%，Attn-JSD が 40.0% となり，Baseline（52.4%）から 10 ポイント以上の劣化を示した．Attn-Ent-Min は Wiki で 45.3%，News で 52.7% と比較的劣化が小さいが，それでも Baseline を下回っている．この結果は，Gemma-2-2B のアーキテクチャが対比デコーディング全般と相性が悪いことを示唆しており，適用時には慎重な検証が必要である．

Phi-2 では，Wiki サブセットで Attn-JSD が 58.8% を達成し，DoLa（58.5%）をわずかに上回った．News サブセットでも DoLa が 61.7% で最良となり，Attn-JSD が 61.4% で続いた．ただし，Attn-Ent-Min は Wiki で 55.1%，News で 58.6% となり，DoLa を下回る結果となった．これは，TruthfulQA で Attn-Ent-Min が MC2/MC3 で優れた性能を示したことと対照的であり，タスク特性によって最適な層選択戦略が異なることを示唆している．

ただし, すべての条件で単調に改善するわけではない点にも注意が必要である. LLaMA-65B の Wiki や Gemma-2-2B の全条件のように，対比手法が Baseline を下回るケースでは，mature layer の分布自体がすでに十分に安定しているか，あるいはモデルアーキテクチャとの相性の問題で対比がかえって識別を乱す可能性がある．したがって，FACTOR では常に対比を強めれば良いというより，データ分割やモデル規模・アーキテクチャに応じて効果が変わることを踏まえた運用が必要となる．これは前節での ICL バイアスに関する FACTOR に対しての評価と一定の関連性がみえ, タスクごとに適切な戦略が異なることを示唆する.

表 4.17　FACTOR Wiki における結果（Accuracy, %）．**太字**は各モデルでの最良値を示す．下線は Baseline を下回る結果を示す．

Model	Base	DoLa	Attn-JSD	Ent-Max	Ent-Min
LLaMA-7B	58.3	62.1	62.4	**62.6**	62.0
LLaMA-13B	62.5	66.2	66.3	**66.5**	66.2
LLaMA-33B	68.3	**69.0**	68.6	**69.0**	68.8
LLaMA-65B	**72.1**	70.4	70.9	70.8	71.1
LLaMA-3.1-8B	63.9	**67.4**	**67.4**	67.3	67.1
LLaMA-3.2-3B	56.5	**61.6**	61.0	61.5	60.9
Gemma-7B	60.5	63.3	**64.1**	63.3	63.2
Mistral-7B	60.6	**64.6**	64.5	64.4	64.4
Gemma-2-2B	**52.4**	40.4	40.0	40.0	45.3
Phi-2	54.6	58.5	**58.8**	58.6	55.1

表 4.18　FACTOR News における結果（Accuracy, %）．**太字**は各モデルでの最良値を示す．下線は Baseline を下回る結果を示す．

Model	Base	DoLa	Attn-JSD	Ent-Max	Ent-Min
LLaMA-7B	58.2	61.7	61.5	**62.0**	**62.0**
LLaMA-13B	60.7	62.5	62.7	**63.4**	63.0
LLaMA-33B	62.5	63.5	65.3	65.2	**65.5**
LLaMA-65B	62.8	63.5	65.2	**65.3**	65.2
LLaMA-3.1-8B	75.6	**75.9**	74.6	74.8	74.6
LLaMA-3.2-3B	68.9	69.4	**69.5**	69.4	**69.5**
Gemma-7B	74.0	**74.8**	73.3	73.6	72.7
Mistral-7B	75.9	75.9	75.3	**76.3**	76.0
Gemma-2-2B	**67.7**	46.5	44.0	45.2	52.7
Phi-2	57.8	**61.7**	61.4	60.3	58.6

4.1.4.7 層選択挙動の分析

DoLa 系の手法では，premature layer 候補をどの範囲から取るか（bucket 選択）が性能と計算量に直結する．本実験では，bucket の分割自体は先行研究 [3] に合わせ，bucket の選択のみを検証で決めた．結果として，TruthfulQA では後半寄りの bucket が選ばれるモデルが多い一方，FACTOR では浅めの bucket が選ばれるモデルが目立つ（表 4.10–4.13）．

この傾向の違いは，質問応答と文章補完で対比に有効な早熟計算の現れ方が異なる可能性を示唆している．TruthfulQA では，誤解や神話に基づく誤答を抑制するために，より深い層での対比が有効であると考えられる．一方，FACTOR では，事実に基づく文章の継続を選択するために，より浅い層での表層的な情報との対比が有効である可能性がある．

また，手法間でも選択される bucket に違いが見られる．たとえば，LLaMA-33B の TruthfulQA では，DoLa と Attn-Ent-Max が $[40, 60)$ を選択するのに対し，Attn-JSD と Attn-Ent-Min は $[20, 40)$ を選択している．Gemma-7B の Attn-Ent-Min では，他の手法が $[14, 28)$ を選択する中，$[2, 14)$ という浅い範囲が選択されている．これらの違いは，各手法が層選択に用いるシグナル（JSD vs エントロピー）の性質の違いを反映していると考えられる．さらに，Gemma 系や LLaMA-3.2-3B のように層 0 の扱いが性能に影響する場合があり，候補集合の作り方（重み共有の有無など）も実装上の重要点となる．

層選択の分布を可視化すると（図 4.1），DoLa および Attn-JSD は候補集合の中でも最浅層に選択が偏りやすい．TruthfulQA では，両手法とも層 16 を 80% 以上の頻度で選択している．これは，最終層との差（JSD）を最大化する選び方が端の層を引きやすいという性質を反映している．一方で Attn-Ent-Min は，より広い層に分散しつつ，相対的に深い層を選びやすい傾向がある．TruthfulQA では選択が層 22–30 に分散しており，FACTOR の Wiki サブセットでは中間層（8–14）に分布している．この差は，エントロピー最小化が入力・トークンに応じて集中先が変わるため，層選択が動的になりやすいという性質から説明できる．

この層選択パターンの質的な違いは，MC2/MC3 指標での改善に寄与していると考えられる．Attn-Ent-Min の動的な層選択は，各トークンの文脈に応じて最適な対比層を適応的に選択することを可能にし，結果として複数の正解候補に対してより均等に確率質量を配分できると推測される．

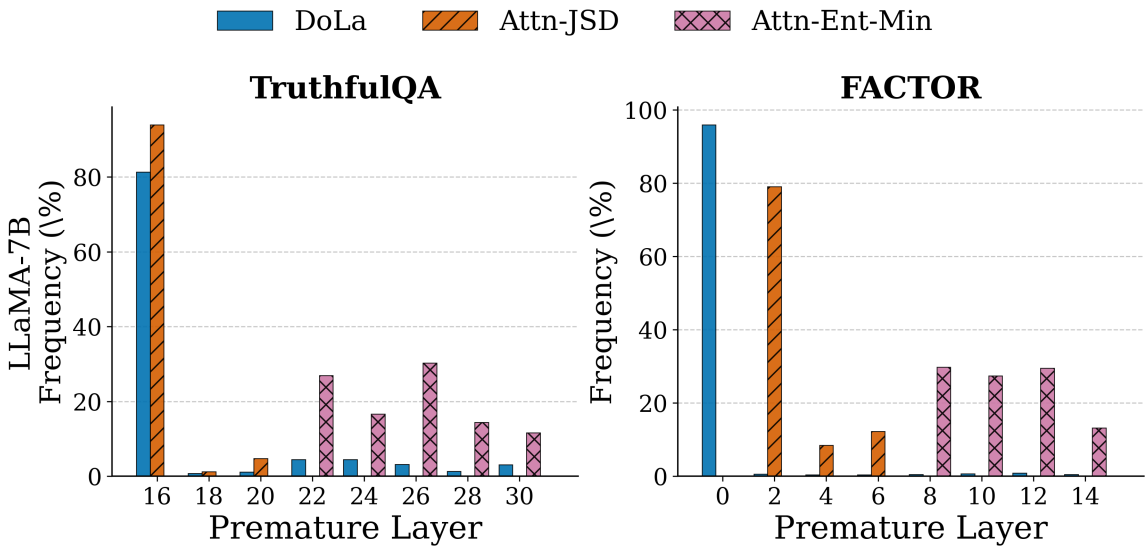

図 4.1　LLaMA-7B における TruthfulQA（左）および FACTOR Wiki（右）での層選択分布.

さらに，ヘッド単位で平均範囲を変える分析では（表 4.19–4.21），有効なシグナルがモデルによって異なることが分かる．Attn-Ent-Min（表 4.19）では，LLaMA-7B において前半ヘッドのみを用いた場合に MC2 が 62.9%，MC3 が 37.3% となり，全ヘッド平均（MC2: 62.8%, MC3: 36.5%）を上回った．これは，LLaMA-7B では前半ヘッドに事実性識別に有用なシグナルが集中していることを示唆している．一方，Gemma-7B では全ヘッド平均が最良（MC1: 37.2%, MC2: 58.2%, MC3: 31.1%）となり，前半・後半ヘッドのみを用いた場合はいずれも性能が低下した．Mistral-7B では，全構成でスコアが同一（MC1: 35.0%, MC2: 56.3%, MC3: 29.9%）となり，ヘッドの選択による差異が見られなかった．

Attn-JSD（表 4.20）では，LLaMA-7B において後半ヘッドのみを用いた場合に MC2 が 57.2%，MC3 が 30.9% となり，全ヘッド平均（MC2: 56.2%, MC3: 29.3%）を上回った．これは Attn-Ent-Min とは逆の傾向であり，手法によって有効なヘッドの分布が異なることを示している．Attn-Ent-Max（表 4.21）では，Gemma-7B において前半ヘッドのみを用いた場合に MC3 が 31.0% となり，全ヘッド平均（30.9%）をわずかに上回った．LLaMA-7B および Mistral-7B では，全ヘッド平均が最良または同等であった．

これらの結果は，有効なシグナルがモデルアーキテクチャおよび層選択手法に依存して分布していることを示している．単純な平均化ではなく，ヘッドの選択や重み付けによる最適化の余地があることが示唆される．

表 4.19　Attn-Ent-Min の TruthfulQA におけるヘッド別分析．**太字**は各モデルでの最良値を示す．

Model	Heads	MC1	MC2	MC3
LLaMA-7B	All	**33.4**	62.8	36.5
	First-half	**33.4**	**62.9**	**37.3**
	Second-half	32.1	62.2	36.3
Gemma-7B	All	**37.2**	**58.2**	**31.1**
	First-half	36.2	57.5	30.5
	Second-half	36.8	58.0	31.0
Mistral-7B	All	**35.0**	**56.3**	**29.9**
	First-half	**35.0**	**56.3**	**29.9**
	Second-half	**35.0**	**56.3**	**29.9**

表 4.20　Attn-JSD の TruthfulQA におけるヘッド別分析．**太字**は各モデルでの最良値を示す．

Model	Heads	MC1	MC2	MC3
LLaMA-7B	All	34.2	56.2	29.3
	First-half	**34.4**	55.4	28.9
	Second-half	34.1	**57.2**	**30.9**
Gemma-7B	All	**36.8**	**57.7**	**31.2**
	First-half	36.0	57.0	31.0
	Second-half	35.0	57.0	29.8
Mistral-7B	All	**35.0**	**56.4**	**30.1**
	First-half	**35.0**	56.6	29.9
	Second-half	34.8	56.2	29.7

表 4.21　Attn-Ent-Max の TruthfulQA におけるヘッド別分析．**太字**は各モデルでの最良値を示す．

Model	**Heads**	**MC1**	**MC2**	**MC3**
LLaMA-7B	All	**34.0**	**55.4**	**28.8**
	First-half	33.9	54.7	28.4
	Second-half	33.9	54.7	28.4
Gemma-7B	All	36.5	**57.6**	30.9
	First-half	**36.6**	**57.6**	**31.0**
	Second-half	35.6	57.5	30.4
Mistral-7B	All	**35.0**	**56.3**	**29.9**
	First-half	**35.0**	**56.3**	**29.9**
	Second-half	**35.0**	**56.3**	**29.9**

個々の注意ヘッドが持つシグナルの特性をより詳細に調査するため，単一ヘッドのみを用いた層選択の評価を行った．具体的には，各ヘッドの選択した層を全ヘッド平均の選択と比較し，最も乖離が大きい（一致率が低い）ヘッドを特定した．表 4.22 に，LLaMA-7B における Attn-Ent-Min の単一ヘッド評価結果を示す．

全ヘッド平均からの乖離が大きい上位 3 ヘッド（Head 6, 24, 26）のうち，Head 6 単独で層選択を行った場合，MC1 で 35.3%，MC2 で 64.9%，MC3 で 38.4% を達成した．これは，DoLa（MC1: 34.6%, MC2: 56.5%, MC3: 30.5%）および全ヘッド平均（MC2: 62.7%, MC3: 36.8%）を大幅に上回る結果である．とくに MC3 では，DoLa に対して 7.9 ポイント，全ヘッド平均に対して 1.6 ポイントの改善が見られた．また，上位 3 ヘッド（Head 6, 24, 26）を組み合わせた場合でも，MC1: 34.8%，MC2: 63.9%，MC3: 38.1% と高い性能を維持しており，DoLa および全ヘッド平均を上回っている．これらの結果は，平均化によるノイズ低減だけでなく，有用なヘッドを選択・重み付けする方向に最適化の余地があることを示している．

表 4.22　LLaMA-7B における Attn-Ent-Min の単一ヘッド評価（TruthfulQA）．**太字**は各指標での最良値を示す．

Configuration	**MC1**	**MC2**	**MC3**
DoLa	34.6	56.5	30.5
All heads avg. (32)	32.4	62.7	36.8
Head 6	**35.3**	**64.9**	**38.4**
Heads 6, 24, 26	34.8	63.9	38.1

特定のヘッドが事実性向上に寄与する効果が，特定の問題集合に依存する局所的な現象なのか，それとも統計的に安定した特性なのかを検証するため，TruthfulQA からランダムに 100 問ずつ抽出した 5 つのサブセットについて，全 32 ヘッドを個別に評価した．表 4.23 に，各サブセットで MC3 スコアが上位 3 位となったヘッドを示す．Head 6 は 5 サブセット中 3 サブセットで最高スコアを達成しており，残りの 2 サブセットでも上位 3 位以内に入っている．これは，Head 6 が持つ事実性識別シグナルが，特定の問題集合に依存しない安定した特性であることを示唆している．また，Head 22 および Head 26 も複数のサブセットで上位に入っており，これらのヘッドも比較的安定した有用性を持つと考えられる．一方，ランキングはサブセットによって変動しており（たとえば Subset 2 では Head 1 が最高），完全に固定的ではない．これは，最適なヘッドの選択が問題の特性にある程度依存することを示唆している．将来的には，入力に応じてヘッドを動的に選択・重み付けするメカニズムの導入が，さらなる性能向上につながる可能性がある．

表 4.23　各ランダムサブセットにおける Attn-Ent-Min の上位 3 ヘッド (LLaMA-7B, MC3 基準).

Subset	**Rank 1**	**Rank 2**	**Rank 3**
Subset 1	Head 6	Head 1	Head 22
Subset 2	Head 1	Head 15	Head 10
Subset 3	Head 26	Head 6	Head 5
Subset 4	Head 6	Head 22	Head 26
Subset 5	Head 6	Head 26	Head 22

推論時間の比較を表 4.24 に示す．評価は TruthfulQA からサンプリングした 100 問について，LLaMA-7B を用いて各手法の平均処理時間を計測した．

Baseline に対して，DoLa は候補層ごとの logits 計算と JSD 評価が入るため，1.24 倍の増分が生じる．Attention ベース手法では，JSD 計算を伴う Attn-JSD が最も重く，相対で約 1.27 倍である．一方，エントロピーのみを用いる Attn-Ent-Max および Attn-Ent-Min は比較的軽量で，それぞれ 1.15 倍および 1.20 倍に収まる．

すべての手法が単一の forward pass 内で完結しており，追加の forward pass や外部モデルの呼び出しを必要としない．総じて，事実性改善に対する追加コストは実用上許容しやすい範囲にあり，とくにエントロピー系手法は計算効率と性能のバランスに優れている．

表 4.24　TruthfulQA における推論効率（LLaMA-7B, 100 サンプル）.

Method	**Time (s/sample)**	**Relative**
Baseline	0.84 ± 0.32	1.00×
DoLa	1.03 ± 0.44	1.24×
Attn-JSD	1.06 ± 0.41	1.27×
Attn-Ent-Max	0.96 ± 0.38	1.15×
Attn-Ent-Min	1.00 ± 0.34	1.20×

4.1.4.8　Contrastive Decoding に関する考察

DoLa と Attention ベース手法は改善する指標がそれぞれ異なる傾向にあった．DoLa は特定の正解候補へ確率を強く寄せる挙動になりやすく，MC1 のような単一正解の識別で強みが出る一方，Attention ベース（とくにエントロピー系）は正解集合全体へ確率質量を配りやすく，MC2/MC3 の改善に寄与しやすい（表 4.14–4.16 およびスコア分布の分析；表 4.25, 4.26）.

この特性の違いを定量的に分析するため，LLaMA-33B におけるサンプルレベルの正解パターンおよびスコア分布統計を調査した．表 4.25 に，各手法間での MC1 正解パターンの比較を示す．DoLa と Attn-Ent-Max の比較では，両手法とも正解が 150 サンプル，DoLa のみ正解が 111 サンプル，Attn-Ent-Max のみ正解が 87 サンプルとなっており，両手法が正解するサンプル集合の重複が比較的小さいことがわかる．一方，DoLa と Attn-JSD，DoLa と Attn-Ent-Min の比較では，両手法とも正解のサンプルが 191–194 と多く，重複が大きい.

表 4.26 に，スコア分布の統計量を示す．Top margin（最高正解スコアと最高不正解スコアの差）は，DoLa が 14.8 で最大であり，Attention ベース手法は 8.5–10.3 と小さい．これは，DoLa が単一の正解候補を強く押し上げる傾向があることを示している．一方，True std（正解選択肢間のスコア標準偏差）は，DoLa が 31.3 で最大であり，Attn-Ent-Max が 17.5 で最小である．これは，DoLa が sharp な分布（特定の正解に集中）を生成するのに対し，Attn-Ent-Max が flat な分布（複数の正解に均等に配分）を生成することを示している．これらの分析結果は，MC1（単一最良解の識別）には DoLa が有利であり，MC2/MC3（複数正解への適切な確率配分）には Attention ベース手法が有利であるという実験結果と整合している.

表 4.25　LLaMA-33B における TruthfulQA のサンプルレベル正解パターン（MC1）.

Comparison	**Both corr.**	**DoLa only**	**Attn only**	**Both wrong**
DoLa vs Attn-JSD	194	67	58	498
DoLa vs Attn-Ent-Max	150	111	87	469
DoLa vs Attn-Ent-Min	191	70	57	499

表 4.26　LLaMA-33B におけるスコア分布統計．Top margin = max(correct) − max(incorrect)．True std = 正解選択肢スコアの標準偏差.

Method	**Top Margin**	**True Std**
DoLa	14.8	31.3
Attn-JSD	10.3	23.6
Attn-Ent-Max	8.5	17.5
Attn-Ent-Min	10.2	23.1

4.1.5 まとめ

本章では，ICL における例示起因の出力変動をバイアスとして捉え，キャリブレーションを入力単位で選択的に適用する枠組みの有効性を検証した．医療系 QA タスク（MIRAGE）では，一括補正が精度を改善する条件がある一方で，条件によっては過補正により精度低下や ICL 依存問題数の増加を招くことが確認された．これに対し選択的補正は，平均精度を大きく損なわずに ICL 依存問題数を減少させる傾向があり，文脈差に起因する不安定性の抑制に寄与することを示した．また RAG の導入はタスクとコーパスの適合に強く依存し，有効な条件ではキャリブレーションが逆効果となり得るため，補正の適用範囲を制御する重要性が示唆された．

さらに，文脈変動下におけるハルシネーション評価として TruthfulQA と FACTOR を用い，例示の違いが評価値の安定性に与える影響と，補正手法のタスク依存性を確認した．とくに FACTOR のように文脈が性能向上に寄与する設定では，一律補正が大幅な劣化を招き得るのに対し，選択的補正は過補正を回避しつつ基準性能を維持しやすいことを示した．これらの結果は，ハルシネーション評価を単一プロンプト設定に依存して行うことの限界と，文脈変動を考慮した再評価および補正設計の必要性を示している．

最後に，推論時介入として Contrastive Decoding および DoLa に基づく手法を検討し，Attention 分布を層選択のシグナルとして用いる設計が TruthfulQA と FACTOR で有効となり得ることを示した．全 10 モデル（LLaMA 系列 4 モデル，LLaMA-3 系列 2 モデル，Gemma 系列 2 モデル，Mistral-7B，Phi-2）での評価を通じて，Attention ベース手法がとくに MC2/MC3 のような複数正解指標で改善を示すこと，および手法の効果がモデル・タスクに依存することを明らかにした．また，層選択挙動とヘッド寄与の分析から，Attention 信号の取り出し方に最適化余地があること，特定のヘッド（LLaMA-7B の Head 6 など）が平均よりも強い事実性識別シグナルを持つこと，および推論効率の観点でエントロピー系手法が比較的軽量であることを確認した．

第 5 章

おわりに

5.1 本論文のまとめ

本研究は, LLM の推論時挙動が文脈に強く依存する点に着目し, (i)ICL における例示起因のバイアス抑制，(ii) 文脈変動下でのハルシネーション評価の頑健化，(iii)Contrastive Decoding における層選択の高度化，の 3 点から検討を行った.

まず，ICL バイアス抑制では，例示の選択・順序によって出力が変動し，入力内容とは独立に特定ラベルが選ばれやすくなる現象に対し，推論時キャリブレーションを基盤として位置付けた．一方で，補正を全入力に一律適用すると，バイアスが弱い入力や，文脈が有用に働く入力に対しては，かえって推定誤差を注入し得る．この問題意識に基づき，本研究では ICL 依存度が高い入力のみを同定し，その部分に限定して補正を行う選択的補正を提案した．医療系タスクにおける検証では，従来の一括補正と比較して，ICL 依存問題の削減に寄与しつつ，条件によっては性能面でも同等以上となる傾向を確認した．また，外部知識を参照する RAG と組み合わせた場合には，RAG が文脈として追加情報を与えるため，補正の効き方がタスクや入力により変化し得ることを示し，補正は一律に強めるのではなく，状況に応じて作用範囲を制御する必要があるという示唆を得た.

次に，ハルシネーション評価の文脈頑健性に関して，TruthfulQA と FACTOR のような評価は固定文脈で運用されがちである一方，実運用では会話履歴やユーザ例示など動的文脈が前提となる点に着目した．Few-Shot 例示を変えるだけで評価結果が変動し，小規模モデルでは性能低下，大規模モデルでも予測の不安定化が生じ得ることを定量化した．さらに，ICL バイアス補正はタスク依存の効果を持ち，TruthfulQA では改善する一方で FACTOR では大幅な悪化を引き起こし得ることを確認した．この挙動は，質問応答型と補完型で Few-Shot 文脈が担う役割が異なることを示唆しており，文脈を単に除去すべき

バイアスとして扱うだけでは不十分である．この問題に対し，本研究は問題単位の予測安定性に基づいて補正範囲を制御する枠組みを導入し，一括補正の過補正を緩和しつつ，文脈変動下でもより安定した評価を行える方向性を示した．

最後に，Contrastive Decoding では，DoLa の枠組みを出発点として，層選択を出力分布の差分だけに依存させず，Attention 分布に由来する内部構造の信号で選択する方針を採った．具体的には，Attention の分布形状を表す指標（分布の発散や集中度）を用いて候補層集合から層を選び，対比スコアを構成する手法を設計した．TruthfulQA では，複数の許容解を考慮する指標（MC2/MC3）で改善が一貫して観測され，DoLa に対しても，モデル・指標によっては上回る傾向を確認した．FACTOR でも，少なくとも Baseline に対しては安定して改善し，DoLa と同等か一部設定で上回る結果が得られた．これらは，Attention 分布が層選択のより感度の高い手掛かりになり得ることを示し，層選択の設計自由度を拡張する知見につながる．

5.2 今後の課題

本研究にはいくつかの限界があり，今後の課題として以下が挙げられる．

- 選択的補正は，安定性推定のために複数回の推論を必要とし，計算コストが増加する．実用面では，推論回数を増やさずに安定性を近似する指標（単回推論の不確実性，層間一貫性など）の検討が重要である．
- 安定性閾値（例：α）は実験的に設定しており，タスク・モデル・shot 数に対して最適化された選び方ではない．閾値の自動選択や，入力ごとの連続的な重み付け（補正強度の適応化）を含め，より原理的な設計が必要である．
- 文脈変動下の評価では，TruthfulQA と FACTOR で補正の効果が大きく異なったが，この差を生む要因（動的選択肢形式における文脈の役割，補完型タスクでの例示情報の寄与など）は十分に解明できていない．より詳細なエラー分析や，文脈が寄与する情報の分解が課題である．
- 評価対象は主に英語の多肢選択形式であり，自由生成の事実性，長文生成，対話履歴や検索文書を含む現実的な動的文脈への拡張は未検証である．より広い設定で，頑健な評価プロトコルと補正戦略を検討する必要がある．
- Attention を用いた層選択では，bucket 選択に検証用分割を用いる設計を採っており，タスクや指標への依存が残る．bucket 設計の一般化，検証コストの削減，および head レベルの信号が事実性と結び付く機序の解明が今後の課題である．

これらの課題に取り組むことで，文脈依存の偏りを抑えつつ，実運用に近い状況でも信頼できる評価と推論の実現が可能になると考える．

参考文献

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv:2310.11511*, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901, 2020.

[3] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2270–2282, 2020.

[5] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia,

Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, 2024.

[7] Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 14014–14031, 2023.

[8] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

[9] Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. Prototypical calibration for few-shot learning of language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, Vol. 43, No. 2, pp. 1–55, 2025.

[11] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1–38, March 2023.

[12] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 2023.

[13] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020.

[14] Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. Unsupervised dense retrieval with relevance-aware contrastive pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp.

10932–10940, 2023.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

[16] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, 2023.

[17] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 12286–12312, 2023.

[18] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022.

[19] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022.

[20] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, Vol. 35, pp. 17359–17372, 2022.

[21] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

[22] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 49–66, 2024.

[23] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *arXiv:2112.09332*, 2022.

[24] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10862–10878, 2024.

[25] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference*, Vol. 500-225, pp. 109–126, 1995.

[26] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, Vol. 3, No. 4, pp. 333–389, 2009.

[27] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, Vol. 24, No. 5, pp. 513–523, 1988.

[28] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, 2021.

[29] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

[31] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[32] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren' s song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, Vol. 51, No. 4, pp. 1373–1418, 2025.

[33] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, pp. 12697–12706, 2021.