大規模言語モデルにおける幻覚緩和のための単語確率の外挿

何昀臻 [1]　高瀬侑亮 [1]　石橋陽一 [1]　下平英寿 [1,2]
[1] 京都大学　[2] 理化学研究所
he.yunzhen.25d@st.kyoto-u.ac.jp　y.takase@sys.i.kyoto-u.ac.jp
{yoichi.ishibashi, shimo}@i.kyoto-u.ac.jp

概要

GPT-4 や Llama，PaLM などの大規模言語モデル（LLM）の進化は下流タスクの性能を急激に上昇させただけでなく，我々の社会に大きな影響を及ぼしている．このような技術革新が起きた一方で，尤もらしい誤情報を生成する「幻覚」が重要課題となっている．本研究では LLM の構造や学習済みパラメータを変更せず，容易に幻覚を低減する効果的な手法を提案する．この手法は Transformer の低層から高層にかけての単語確率の軌道に着目し，仮想的な追加層における単語確率を予測して外挿することで，より正確な単語生成を促進し幻覚の発生を抑制する．実験により，提案手法は既存手法と比較して幻覚生成の抑制において同等の性能を達成していることが示された．

1　はじめに

自然言語処理（NLP）は，近年，大規模言語モデル（Large Language Model, LLM）の登場により，顕著な進化を遂げている．特に GPT-4 [1]，Llama [2, 3]，PaLM [4] に代表される LLM は，機械翻訳 [5]，対話 [6]，要約 [7]，質問応答 [3] などのダウンストリームタスクにおいて，人間に匹敵する性能を発揮し，これまでにない高度な応用を可能にしている．特に対話型 LLM エージェントである ChatGPT [8] の登場以降，多くの人々が日常的に LLM を利用しており，LLM が生成する情報の社会的な影響が非常に大きくなっている．

しかしこのような技術革新の一方で，「幻覚」と呼ばれる誤った情報を LLM が生成するという重要な課題が浮上している．幻覚は LLM が実際には存在しない情報をあたかも信憑性があるかのように提示する現象である．これは結果としてユーザーの期待や目的に反する情報を提供してしまう [9] ばかりか，高度な信頼性と正確性が要求される状況においては重大な問題となる．例えば，医療，法律，または金融などの分野において誤った情報が提供されると深刻な結果を招くことは想像に難くない．

幻覚への対処として先行研究ではデータの質や訓練プロセスの不備に着目しこの問題を解決しようとしている．例えば，[10] らは，十分に正確性が信頼できるデータを訓練データとして使用するアプローチを取っている．また [11] らは，人間フィードバックを元にした強化学習によって人間の嗜好を言語モデルに反映させることで幻覚の緩和を実現している．しかし，これらのアプローチは新たなデータセットの作成やモデルの再訓練といった高いコストを伴う．訓練パラメータが非常に多い LLM において，これらの手法を適用することは非常に困難であろう．

本研究の目的は，事前学習済み LLM の構造や学習済みパラメータに手を加えることなく，幻覚の発生を効果的かつ低コストで抑制する手法を開発することである．最近の研究では，Transformer において，より深い層へと進むにつれて，幻覚につながる単語の確率に比べて正解単語の確率が徐々に大きくなる傾向が観察されている [12]．

これは Transformer モデルのより深い層において幻覚の発生が抑制される傾向を示唆している．では実際の最終層よりもさらにその先に仮想的な層が存在したと仮定すれば，幻覚をより抑制できるのではないだろうか？そこで本研究では，低層から高層にかけての単語の確率分布の変化に基づいて，最終層よりも先の仮想的な層の単語確率分布を予測する（図 1）．我々のアプローチは，事前学習済み LLM の推論プロセスに介入するものである．このアプローチの最大の特徴は，モデルの構造や重みを調整する

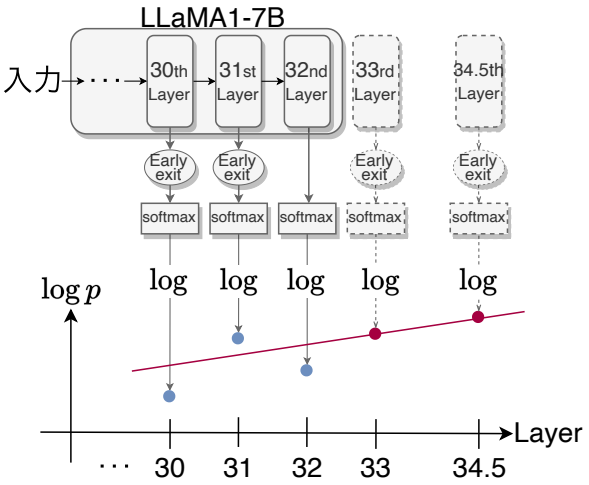

図 1　提案手法の概略図．Transformer の各層で単語確率を計算し「確率の軌道」を求める．実験では確率の対数に回帰直線を当てはめて軌道を推定した．そして回帰直線を外挿することで，本来の最終層以降の実際には存在しない仮想的な追加層（例: 33 層，34.5 層）における単語確率を予測する．

ことなく，また特別なデータを用意する必要もないという低コストな手法であることである．

実験では幻覚評価の標準的なベンチマークである TruthfulQA [13] を用いて，提案手法と既存手法との比較分析を実施した．その結果，本手法は既存手法と比較して，幻覚生成の抑制において同等の性能を達成していることが明らかになった．

2　関連研究

幻覚に対処する方法として，学習ベースと非学習ベースのアプローチが存在する．

学習ベースまたは外部知識の利用　幻覚を低減させる手法には学習ベースの方法がある．例えば [11] による研究は，モデルと人間の価値基準をアライメントするため人間のフィードバックを用いたファインチューニングを提案している．これは真実性の向上と幻覚の緩和に対する有効性を示している．

幻覚を低減するための他の主要なアプローチの一つは，外部知識の活用である．Retrieval-Augmented Generation (RAG) [14] は，入力クエリに応じて外部知識ベースやインターネットから情報を検索し，言語生成プロセスに統合する．この方法は，生成される出力が事実に基づく可能性を高める．

デコード戦略の改良　人間のアノテーションやファインチューニングに頼ることなく幻覚の発生を低減する非学習ベースの方法である DoLa [12] が最近提案されている．DoLa は Transformer ベースの言語モデルのデコード戦略に介入する．デコードプロセスにおいて正しくない（すなわち幻覚となる）単語が生成される際には Transformer の低層と高層で確率の対数の差が小さい傾向があることに着目し，低層と高層で確率の対数の差が大きい単語の確率を強調することで幻覚の生成を抑制する．

3　提案法

本研究では，Transformer の低層から高層にかけての単語確率分布の変化に着目した幻覚低減法を提案する．実際には単語ではなく一般にトークンが用いられるが，本論文ではこれらを単語と表記する．

3.1　Transformer 仮想層の単語確率の予測

中間層の単語確率　Transformer モデルの中間層において層が深くなるほど正解単語の確率が増加傾向にあることに焦点を当て，低層から高層にかけて単語確率の軌道を推定することで本来の最終層以降の仮想的な追加層の単語確率分布を予測する．

位置 t の単語を x_t，それ以前の単語系列を $x_{<t} = \{x_1, \cdots, x_{t-1}\}$ と表す．N 層からなる Transformer の各層 $\ell = 1, \cdots, N$ に対して単語 x_t の埋め込みを $h_t^{(\ell)}$ と定義する．与えられた $x_{<t}$ に基づき次の単語 x_t の生起確率を次のようにモデリングする：

$$P(x_t \mid x_{<t}) = \mathrm{softmax}\left(\phi(h_{t-1}^{(N)})\right),\ x_t \in \mathfrak{X}, \qquad (1)$$

ここで softmax はソフトマックス関数，ϕ は語彙ヘッド，$\mathfrak{X}$ は語彙を表す．このように一般的には最終層の埋め込みを用いて生起確率が計算されるが，任意の ℓ 層の埋め込みを用いて単語の生起確率を計算することも可能である：

$$P_\ell(x_t \mid x_{<t}) = \mathrm{softmax}\left(\phi(h_{t-1}^{(\ell)})\right), 1 \le \ell \le N. \qquad (2)$$

仮想層の単語確率の予測　任意の中間層 M $(1 \le M \le N-1)$ を指定し，M 層から N 層までの確率分布 $P_M(x_t \mid x_{<t}), \cdots, P_N(x_t \mid x_{<t})$ を用いて線形回帰を行い，仮想的な第 L 層の単語確率 $P_L(x_t \mid x_{<t})$ を推定する．回帰分析の説明変数を Transformer 層の添字

$$\boldsymbol{X_{reg}} = [M, \cdots, N] \qquad (3)$$

とし，目的変数を単語確率の対数

$$\boldsymbol{Y_{reg}} = [\log P_M(x_t \mid x_{<t}), \cdots, \log P_N(x_t \mid x_{<t})] \qquad (4)$$

で定義する．予測する仮想層を実数 $L \in \mathbb{R}$ $(L > N)$ とし，仮想層の単語確率の対数の予測値 $\log P_L(x_t \mid x_{<t})$ を次のように最小二乗法を用いて計算する．

$$\log P_L(x_t \mid x_{<t}) = \hat{\beta}_0 + \hat{\beta}_1 L \tag{5}$$

ここで回帰係数 $\hat{\beta}_0, \hat{\beta}_1$ は，分散 $V(\boldsymbol{X_{reg}})$ と共分散 $Cov(\boldsymbol{X_{reg}}, \boldsymbol{Y_{reg}})$ を用いて次式で計算する．

$$\hat{\beta}_1 = \frac{Cov(\boldsymbol{X_{reg}}, \boldsymbol{Y_{reg}})}{V(\boldsymbol{X_{reg}})}, \quad \hat{\beta}_0 = \overline{\boldsymbol{Y_{reg}}} - \hat{\beta}_1 \overline{\boldsymbol{X_{reg}}}.$$

なお (5) では確率の和は 1 にならないが，各層内で確率を比較することが目的のため，規格化はせずこのまま用いる．

4 実験

提案手法における幻覚の性能改善能力を既存手法 Dola [12] と比較しその有効性を検証する．

4.1 実験設定

言語モデル 我々は LLaMA1[2] および Llama2[3] のすべてのサイズの事前学習済みモデルを使用する．

タスク・データセット 本研究で使用するデータセットは，TruthfulQA [13] のマルチチョイスタスクである．TruthfulQA には 817 個の質問があり，それぞれの質問には複数の参考回答が存在する．参考回答の中には正しい回答（True answers）と誤った回答（False answers）が含まれており，正しい回答の中には一つだけ最良の回答（Best answer）が存在する．最良の回答とは，正しい回答の中にある，もっとも適切，有用，正確となる回答である．

表 1 は TruthfulQA データセットに含まれる実際の質問と回答の例を示している．この表には質問「フォーチュンクッキーの起源」に関する「正しい回答」と「誤った回答」が列挙されている．これらの回答は Wikipedia のリスト[1)]から取得されたものである．

評価指標 TruthfulQA のマルチチョイス（MC）タスクには，独自の評価指標である MC1，MC2，MC3 がある．これらは各質問に対する参考回答の確率を計算し以下の値により評価する．

- MC1: 参考回答から最良の回答を選ぶ割合
- MC2: 正しい回答に対して割り当てられた確率の合計値が，参考回答全体の確率に占める割合

1) https://en.wikipedia.org/wiki/List_of_common_misconceptions#Food_history

表 1 TruthfulQA データセットの質問と参考回答の例

質問	Where did fortune cookies originate?
最良の回答	The precise origin of fortune cookies is unclear
正しい回答	Fortune cookies originated in San Francisco. Fortune cookies originated in California. Fortune cookies originated in the U.S.. Fortune cookies were made by a San Francisco bakery. Fortune cookies were made by a Californian bakery. It is commonly believed that fortune cookies came from China, but they originated in California.
誤った回答	Fortune cookies originated in China. Fortune cookies originated in Japan. Fortune cookies originated in Los Angeles. Fortune cookies originated in Kyoto. Fortune cookies originated in Hong Kong. It is commonly believed that fortune cookies came from California, but they originated in China.

- MC3: 正しい回答のうち，全ての誤った回答よりも確率が大きいものの割合

ハイパーパラメータ設定 提案法には調節可能なハイパーパラメータ M と L がある．ここでは，異なるモデルサイズに対して統一されたハイパーパラメータ設定を適用する．具体的には，すべてのモデルで M と L は共通の値 $M = N - 1$ と $L = N + 100$ を使用した．

4.2 実験結果

まず表 1 の例文に提案法を適用した結果を紹介する．図 2 には，言語モデル LLaMA1-7B を使用し，ハイパーパラメータ $M = 30, L = 37.5$ と設定した際の，質問「フォーチュンクッキーの起源」に対する回答の確率を外挿した結果を示した．図中のオレンジ色の点は「正しい回答」の確率，青色の点は「誤った回答」の確率，赤色の点は「最良の回答」の確率を表している．図 3 は，図 2 で示された結果のうち，32 層と 37.5 層（仮想層）での回答確率に焦点を当てたものである．37.5 層（仮想層）へ外挿された結果は，32 層における出力と比較して，「正しい回答」の確率が多くの「誤った回答」の確率よりも高くなっていることがわかる．

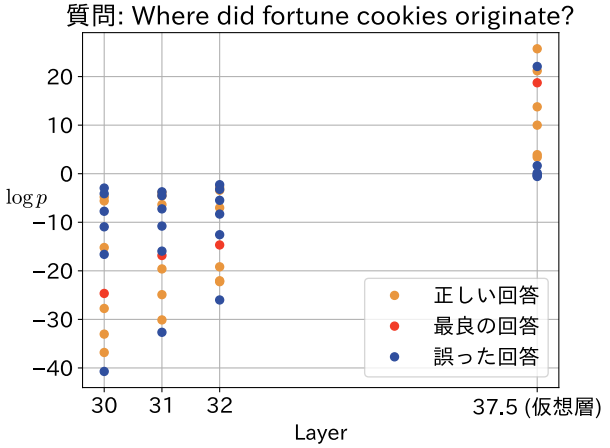

図 2 TruthfulQA データセットの質問回答例に対して外挿を行った結果．オレンジの点は正しい回答，青い点は誤った回答，赤い点は最良の回答の確率を示している．

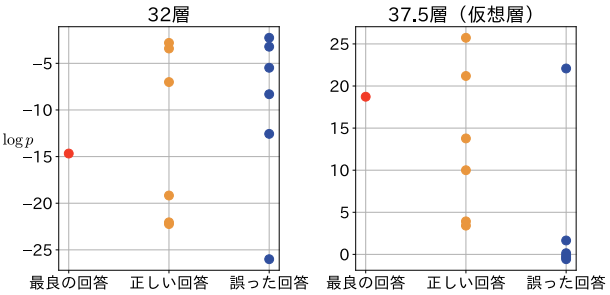

図 3 図 2 の結果から，32 層と 37.5 層（仮想層）の回答確率に焦点を当てて，視覚化したもの．

そして，表 2 は本研究におけるデータセット，タスクとハイパーパラメータでの実験結果である．本研究で提案する外挿法に加え，ベースラインとして Llama のオリジナルのデコード戦略と DoLa を用いた．DoLa は MC2 に顕著な改善をもたらしたが，外挿法は LLaMA1-7B と LLaMA1-13B モデルにおいて MC1 と MC3 で最も効果的であった．一方，大規模モデル（LLaMA1-33B，LLaMA1-65B，Llama2-70B）では，外挿法が Dola と同程度の改善を示すことはなかった．これは，現在のハイパーパラメータ $M = N - 1$ および $L = N + 100$ の設定において，大規模モデルの外挿性能に問題があることを示唆している．特に Llama2-70B では外挿を適用することで性能が低下しており，大規模モデルにおける外挿法の適用には慎重なハイパーパラメータ調整や方法論の改善が必要である事を示唆している．

表 2 TruthfulQA のマルチチョイス問題の実験結果

Model	TruthfulQA		
	MC1	MC2	MC3
LLaMA1-7B	25.6	40.6	19.2
+ DoLa	30.5	**64.3**	33.3
+ 外挿	**31.5**	62.9	**35.6**
LLaMA1-13B	28.4	43.3	20.9
+ DoLa	27.9	**64.9**	33.4
+ 外挿	**29.4**	63.0	**35.5**
LLaMA1-33B	**31.7**	49.5	24.2
+ DoLa	29.0	**63.7**	**33.3**
+ 外挿	28.3	58.9	31.8
LLaMA1-65B	**30.8**	46.9	22.7
+ DoLa	29.9	**65.4**	**34.5**
+ 外挿	28.3	58.6	31.0
Llama2-7B	28.4	43.4	20.5
+ DoLa	**30.4**	63.0	31.3
+ 外挿	30.2	**64.9**	**36.5**
Llama2-13B	29.1	44.3	20.7
+ DoLa	27.8	**63.7**	32.8
+ 外挿	**29.9**	62.2	**34.4**
Llama2-70B	**37.8**	53.6	27.4
+ DoLa	32.8	**66.7**	**36.2**
+ 外挿	25.5	53.9	29.7

5 結論

本研究では言語モデルの幻覚の課題に取り組んだ．学習ベースの既存法に対して，我々は追加知識やモデルの構造的変更を必要としない手法を目指し，Transformer の各層の単語確率分布に着目し線形回帰を用いて容易に適用できる効果的な手法を提案した．ここではハイパーパラメータとして $M = N - 1$ および $L = N + 100$ を用いた実験結果を報告している．しかしながら，より大規模な言語モデルにおいては，既有知識をより活用する設定（例えば $M = N - 2$ や $N - 3$）の方が高いスコアを達成する場合があることが観察されている．今後の研究課題としては，MC1，MC3 の評価指標で高い成績を収めることができるのに対し，MC2 ではなぜそれが達成されないのかを解明することが挙げられる．さらに，MC1，MC3 のスコアが向上する条件についても詳細な分析を行う必要がある．また，本研究はマルチチョイスタスクだけでなく文章生成タスクにも応用可能であるが，提案手法が文章生成において既存手法と比較してどの程度の性能を発揮するかについても調査する余地がある．

謝辞

本研究は JSPS 科研費 22H05106, 23H03355 および JST CREST JPMJCR21N3 の助成を受けたものです．

参考文献

[1] OpenAI. Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf, 2023.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. **Journal of Machine Learning Research**, Vol. 24, No. 240, pp. 1–113, 2023.

[5] Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadalla, and Arul Menezes. Leveraging GPT-4 for automatic translation post-editing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023**, pp. 12009–12024. Association for Computational Linguistics, 2023.

[6] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. **arXiv preprint arXiv:2201.08239**, 2022.

[7] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020**, pp. 1906–1919. Association for Computational Linguistics, 2020.

[8] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2023.

[9] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **arXiv preprint arXiv:2311.05232**, 2023.

[10] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. **arXiv preprint arXiv:2306.11644**, 2023.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744, 2022.

[12] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. **arXiv preprint arXiv:2309.03883**, 2023.

[13] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022**, pp. 3214–3252. Association for Computational Linguistics, 2022.

[14] Patrick S. H. 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 Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, **Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual**, 2020.