復号手法が大規模言語モデルにおける不確実性推定に与える影響の調査

橋本 航　上垣外 英剛　渡辺 太郎
奈良先端科学技術大学院大学
{hashimoto.wataru.hq3, kamigaito.h, taro}@is.naist.jp

概要

GPT-3.5/4 や LLaMA などの大規模言語モデル（Large Language Models, LLM）は、自然言語処理における様々なタスクの性能を大きく発展させた．しかし，生成テキストに誤情報を含む「幻覚」をはじめとした品質の低い出力が依然として存在するため，特に医療や金融といった確実性が重要な領域での活用には課題が残されている．本研究では，生成テキストの品質向上を目指す復号手法が，質問応答と要約タスクにおいて大規模言語モデルの不確実性推定性能に与える影響を調査した．実験により，貪欲探索やビーム探索のようなシンプルな復号手法が，不確実性推定性能の観点で優れていることが判明した．

1　はじめに

近年、自然言語処理の様々なタスクは大規模言語モデル (Large Language Models, LLM) の登場により大きな発展を遂げている．特に GPT-4 [1] や LLaMA [2] に代表される LLM は，様々なダウンストリームタスクにおいて優れた性能を発揮している．さらに，AI エージェントとしてより複雑なタスクに LLM が組み込まれるなど，これまでにない高度な応用を可能にしている [3]．しかし，LLM は出力内容に誤情報が含まれる「幻覚」問題をはじめとし，品質の低い文章を依然として生成することがある．そのため，医療や金融，法律などの確実性が非常に重要な領域での積極的な活用を困難にしている．

このような幻覚などの問題を解決する手段の一つとして，テキスト復号手法が着目されている．復号手法は，LLM の出現以前よりテキスト生成モデルが様々なタスクを解くようになるための重要な役割を果たしている．また，近年復号手法が LLM の性能に大きな影響を与える可能性があることが示されており [4]，復号手法の選択は LLM の性能を引き出すためにも重要である．

一方で，LLM を確実性の高い領域で実応用するためには，生成品質だけでなく出力がどの程度不確実かを推定することも必要である．具体的には，どの程度出力が不確実かを数値化する不確実性スコア等で品質が低いと思われる出力を検知し，検知された場合は人手やより性能の高い LLM で修正したうえで出力を提供したい．多くの場合，復号手法により出力されるテキストやそれを構成するトークンのスコアが変わるため，復号手法は不確実性スコアの品質に影響を与える可能性が高い．このような不確実性推定性能の観点で復号手法が与える影響を明らかにすることは，LLM の出力をより信頼性が高いものにするためにも重要な問いである．

本研究では復号手法が不確実性推定性能がどのような影響を与えるかについて，不確実性推定手法と組み合わせて網羅的に調査する．実験では，質問応答タスクのデータセットである TriviaQA [5] と要約タスクのデータセットである XSum [6] を用いて，復号手法と不確実性推定手法を組み合わせた場合の選択的生成性能 [7] を評価した．その結果，貪欲探索やビーム探索のようなシンプルな復号手法が比較的優れた不確実性推定性能を示すことが判明した．

2　手法

本節では，実験で使用する復号手法と不確実性推定手法について説明する．

2.1　復号手法

貪欲探索 (Greedy Search, Greedy)　は，各ステップで最も確率の高いトークンを選択する．

ビーム探索 (Beam Search, BS)　は，各ステップで最も確率の高い k 個の系列を保持しながらトーク

ンを選択する [8]．ここで k はビーム幅と呼ばれるハイパーパラメータである．本研究では $k = 3$ とした．また，不確実性スコアの計算に用いるトークンの確信度には，各出力トークンで選択されたビームにおける確信度を使用した．

Contrastive Search (CS) は，意味の一貫性を損なうトークンにペナルティを課しながらトークンを選択する [9]．本実験では，ペナルティの係数を 0.6 に設定した．

Inference-Time Intervention (ITI) は，TruthfulQA [10] データセットで訓練した線形分類器により，事実性の高い注意ヘッドと，事実性を表現するベクトルを持つ層を取得し，復号時に介入を行うことで事実性の高い出力を得る [11]．本研究では，訓練済み線形分類器が付与されたモデルを使用した．[1)]

DoLa は，事実性の高い出力トークンのスコアが層の後半に伴い高くなるという分析に基づき，最終層とそれ以前の層のロジット差を対比することでより事実性が高められた出力トークン分布を取得する [12]．本実験では，Shi らの論文 [4] に従って対比対象に [0, 16) および [16, 32) からの偶数番号の層を選択し，それぞれ DoLa-low, DoLa-high として実験する．

2.2 不確実性推定手法

本実験では、大規模言語モデルにおける不確実性推定のためのフレームワークである LM-Polygraph [13] に実装されている手法を用いる．具体的には，各トークンの確率分布を用いて不確実性スコアを出力する情報ベースの手法では，Maximum Sequence Probability および Mean Token Entropy [14] を，複数回の推論に基づくサンプリングベースの手法として Sentence Shifting Attention to Relevance [15] を用いた．また，訓練データの分布を近似した上で入力の尤度から不確実性スコアを求める密度ベースの手法では，Mahalanobis Distance [16] および Robust Density Estimation [17] を用いた．

Maximum Sequence Probability (MSP) では，$1 - P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ を不確実性スコアとして出力する．ここで，$\boldsymbol{x}$ は入力系列，$\boldsymbol{y}$ は出力系列，$\boldsymbol{\theta}$ はモデルパラメータである．

Mean Token Entropy (MTE) は，生成時における各トークンの確率分布のエントロピーを平均して不確実性スコアとする [14]．

Sentence Shifting Attention to Relevance (SentSAR) は，候補文を複数サンプリングした上で，他のサンプルに類似している文はより代表的であることを用いた重み付けで不確実性スコアを計算する [15]．1 つの入力に対して 10 件出力をサンプリングし，サンプル文間の類似度行列の計算には STS データセットで訓練済みの RoBERTa[2)] を用いた．

Mahalanobis Distance (MD) では，訓練データの表現集合に多変量ガウス分布を仮定したうえで、入力表現とのマハラノビス距離を不確実性スコアとする [16]．本研究では，文の表現として，decoder における最終層の隠れ表現を非 padding トークンに限定して平均したものを用いた．

Robust Density Estimation (RDE) では，カーネル PCA および最小共分散行列式 [18] を用いて，より外れ値に頑健になるように密度推定を行う [17]．訓練および推論ステップで使用する文の表現には MD と同様のものを用いた．

3 実験設定

言語モデル 我々はすべての実験において LLaMA 2 [19] 7B Chat モデルを用いた．[3)]

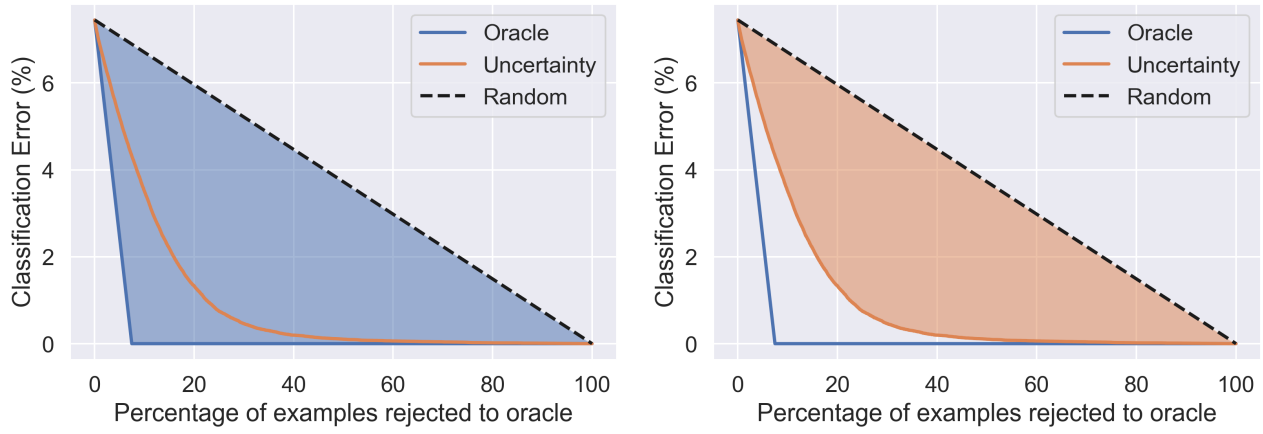
データセット 質問応答タスク向けに，文脈のない複雑な質問を含むデータセットである TriviaQA [5]，要約タスク向けに，抽象型要約のための大規模データセットである XSum [6] を用いた．各データセットの統計情報を付録 A の表 4 に示す．また，プロンプトにはすべて 0 ショットプロンプトを用いた．質問応答と要約に用いたプロンプトは付録 B に示す．

評価指標 本研究では，不確実性スコアに基づいて生成品質が低いテキストを拒否することが可能な場合の，適合率と再現率に関連する選択的生成性能を評価する．ここで，拒否されたモデル出力は，人手やより高度な LLM で処理することが可能であることを想定する．テキスト生成における不確実性推定に関する以前の研究 [13] に基づき，Prediction-Rejection Ratio (PRR) を評価指標と

1) https://huggingface.co/likenneth/honest_llama2_chat_7B

2) https://huggingface.co/cross-encoder/stsb-roberta-large

3) https://huggingface.co/meta-llama/Llama-2-7b-chat-hf



(a) PRC_{orc}　　(b) PRC_{uns}

図 1　Prediction-Rejection 曲線の例 [20]．分類タスクの例では，PRC_{orc} は拒否された例の割合が誤分類の数に等しい部分でエラー率が 0% になるまで下がる．本研究においては生成品質を縦軸にとった上で PRR を計算する．

して用いる．評価のために，テストデータセット $\mathcal{D} = (\boldsymbol{x}_i, \boldsymbol{y}_i)$ を考え，LLM の出力を $f(\boldsymbol{x}_i)$, LLM の出力から得られた不確実性スコアを $\mathcal{U}(\boldsymbol{x}_i)$ とする．Prediction-Rejection 曲線は，不確実性 $\mathcal{U}(\boldsymbol{x}_i) < a$ である出力の生成品質 $\mathbb{Q}(f(\boldsymbol{x}_i), \boldsymbol{y}_i)$ の平均が，拒否するかどうかを定める閾値 a にどのように依存するかを示す．

PRR はオラクル (テキスト生成品質) とランダムスコアの Prediction-Rejection 曲線下の面積 PRC_{orc} と，得られた不確実性スコアとランダムスコアの Prediction Rejection 曲線下の面積 PRC_{uns} の比率から計算できる (図 1 を参照):

$$PRR = \frac{PRC_{\text{uns}}}{PRC_{\text{orc}}} \tag{1}$$

PRR の値が高いほど，選択的生成のための不確実性スコアの品質が優れていることを意味する．[4] ここで，$\mathbb{Q}$ は問題やタスクごとに定義することができ，本研究では質問応答タスク向けに RougeL [21], [5] 要約タスク向けに RougeL, BARTScore [22] および FactKB [23] を用いる（生成品質に関連する評価指標の詳細は付録 C を参照）．以降, RougeL, BARTScore, および FactKB を用いて計算された PRR はそれぞれ PRR-RougeL, PRR-BARTScore, および PRR-FactKB と表記する．また，生成品質を示すそれぞれの指標，および不確実性推定のための指標である PRR は，便宜上 100 倍して提示する．

4)　誤分類よりも正確に予測されたデータの方が不確実性スコアが高いようなケースでは，負の値をとる．

5)　TriviaQA データセットでは正解データに複数の回答が含まれているため，それぞれの回答との RougeL を算出し，最も大きい値を利用する．

4　結果

本節では，質問応答タスクおよび要約タスクにおける，不確実性推定手法に復号手法を組み合わせた場合の不確実性推定性能の結果を提示する．

4.1　質問応答

表 1 に，TriviaQA データセットを用いた質問応答タスクでそれぞれの復号手法と 2.2 節で提示した不確実性推定手法を組み合わせて出力の不確実性スコアを得た場合の PRR-RougeL を示す．また，あわせて各復号手法を用いたときの RougeL も示す．

生成品質および不確実性推定性能の双方で，ビーム探索が優れた性能を示している．また，生成品質の点では他手法に劣後するものの貪欲探索も優れた不確実性推定性能を示し，これらの手法はシンプルであるにも関わらず強力なベースラインであることがわかる．一方で，Contrastive Search や DoLa-low は生成品質では貪欲探索を上回っているものの，不確実性推定性能については劣後しており，生成品質の高さが必ずしも不確実性推定性能に直結するとは限らないことも確認された．

不確実性推定手法間で性能を比較すると，密度ベースの手法は他手法と比較し性能が劣後していることがわかる．これは，LLM におけるトークンのスコアが，事前学習やファインチューニングなどの事後学習に用いたデータの情報を十分に反映しているのに対し，密度ベースの手法では利用している情報が訓練データにおける密度のみを用いており，事前学習や事後学習に用いたデータの密度情報を十分

表 1 質問応答タスクにおける各復号手法の RougeL と、不確実性推定手法を組み合わせた場合の PRR-RougeL. 太字は復号手法間でスコアが最も高いことを示し，下線はスコアが 2 番目に高いことを示す.

復号手法	RougeL	PRR-RougeL				
		MSP	MTE	SentSAR	MD	RDE
Greedy	11.36	62.97	49.13	**64.12**	**46.56**	**44.04**
BS	**12.16**	**63.82**	**51.58**	63.52	32.10	29.69
CS	12.02	55.41	45.25	57.35	21.65	20.46
ITI	9.68	22.75	24.38	22.88	-9.02	0.50
DoLa-low	11.60	60.75	48.90	60.24	34.61	32.47
DoLa-high	11.16	61.15	49.23	60.86	34.65	34.51

表 2 要約タスクにおいて各復号手法に不確実性推定手法を組み合わせた場合の PRR.

復号手法	PRR-RougeL				PRR-BARTScore				PRR-FactKB			
	MSP	MTE	SentSAR	RDE	MSP	MTE	SentSAR	RDE	MSP	MTE	SentSAR	RDE
Greedy	7.05	8.02	6.01	**2.44**	14.60	16.51	14.53	16.60	24.15	27.64	26.52	-24.40
BS	6.55	7.49	6.55	2.00	12.70	13.39	12.78	**19.65**	**31.99**	**35.77**	31.68	-30.50
CS	3.73	6.33	4.49	0.73	11.49	16.35	**15.19**	13.70	19.12	21.91	18.76	**-16.89**
ITI	**19.87**	**20.96**	**19.00**	-7.36	**18.73**	**19.61**	14.09	10.31	9.35	9.94	14.60	-27.27
DoLa-low	6.42	7.40	4.75	1.38	13.77	14.91	11.79	18.96	29.51	32.35	32.78	-27.06
DoLa-high	7.47	8.09	5.74	-0.82	14.76	15.46	12.14	17.22	29.30	31.29	**33.32**	-30.04

表 3 要約タスクにおいて各復号手法を用いた場合の生成品質.

復号手法	RougeL	BARTScore	FactKB
Greedy	15.08	-277.69	52.66
BS	**15.24**	-277.74	**63.51**
CS	14.97	-279.13	44.01
ITI	13.00	-292.20	31.51
DoLa-low	15.21	-277.60	57.96
DoLa-high	15.22	**-277.35**	60.70

に反映できていないためだと考えられる.

4.2 要約

表 2 に，XSum データセットを用いた質問応答タスクでそれぞれの復号手法と 2.2 節で提示した不確実性推定手法を組み合わせて出力の不確実性スコアを得た場合の PRR を示す．また，あわせて各復号手法を用いたときの生成品質を表 3 に示す.

生成品質の観点ではビーム探索や DoLa が優れており，不確実性推定性能においても同様にビーム探索や DoLa が優れている傾向にあることがわかる．また，質問応答タスクと同様に，貪欲探索が依然として優れたベースラインであることも確認された．一方で，ITI は PRR-RougeL および PRR-BARTScore で最も優れているものの，総合的な生成品質および PRR-FactKB で著しく劣後していることから，特に事実性の観点で適切でない確信度を出力していると考えられる.

事実性を高める復号手法である DoLa は総合的な生成品質だけでなく，複数の PRR に渡って優れた不確実性推定性能を示しており，原論文 [12] で実験されていない要約タスクにおいても有効性が確認できた.

5 おわりに

本研究では，大規模言語モデルにおける復号手法が，不確実性推定性能に与える影響を質問応答タスクおよび要約タスクで評価した．その結果，質問応答タスクおよび要約タスクの双方でビーム探索や貪欲探索などのシンプルな復号手法や，ロジット差の対比を用いる DoLa が優れた不確実性推定性能を示すことが判明した．また，シンプルなベースラインである貪欲探索を用いた場合，生成品質は他手法と比較して高くないものの不確実性推定性能は高いことから，必ずしも生成品質の高さが不確実性推定性能の高さを保証するとは限らないことも確認された.

今後の課題として，各復号手法を用いた場合における出力トークンのスコア分布のより詳細な分析や，より複数のタスクやデータセット，復号手法で実験を行い網羅性を向上することが挙げられる.

参考文献

[1] OpenAI. Gpt-4 technical report, 2023.

[2] Meta. Llama: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[3] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. **Frontiers of Computer Science**, Vol. 18, No. 6, 2024.

[4] Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. A thorough examination of decoding methods in the era of LLMs. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 8601–8629, 2024.

[5] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1601–1611, 2017.

[6] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1797–1807, 2018.

[7] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[8] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In **Proceedings of the First Workshop on Neural Machine Translation**, pp. 56–60, 2017.

[9] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In **Advances in Neural Information Processing Systems**, 2022.

[10] 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 (Volume 1: Long Papers)**, pp. 3214–3252, 2022.

[11] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[12] 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.

[13] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-polygraph: Uncertainty estimation for language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 446–461, 2023.

[14] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 539–555, 2020.

[15] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5050–5063, 2024.

[16] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In **Advances in Neural Information Processing Systems**, 2018.

[17] KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 3656–3672, 2022.

[18] Peter J. Rousseeuw. Least median of squares regression. **Journal of the American Statistical Association**, Vol. 79, No. 388, pp. 871–880, 1984.

[19] Meta. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[20] Andrey Malinin. **Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment**. PhD thesis, University of Cambridge, 2019.

[21] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, 2004.

[22] Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTScore: Evaluating generated text as text generation. In **Advances in Neural Information Processing Systems**, 2021.

[23] Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 933–952, 2023.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, 2020.

A　データセットの統計情報

表 4　データセットの統計情報．訓練データとして Fadeeva ら [13] の実験設定に従い 1,000 件サンプリングされ，MD および RDE のみで使用する．

タスク	データセット名	訓練	テスト
質問応答	TriviaQA	1,000	17,944
要約	XSum	1,000	11,334

B　プロンプトのテンプレート

TriviaQA データセットにおける質問応答タスクおよび XSum データセットにおける要約タスクのプロンプトをそれぞれ図 2 および図 3 に示す．

```
# Question:
{question}

# Answer:
{answer}
```

図 2　質問応答タスクに用いたプロンプト．

```
# Instruction:
Please summarize the following document in one sentence.

# Document:
{text}

# Summary:
{summarization}
```

図 3　要約タスクに用いたプロンプト．

C　生成品質に関連する評価指標の詳細

RougeL　は，生成テキストと参照テキスト間の類似性を評価する指標であり，最長共通部分列（Longest Common Subsequence; LCS）に基づいて計算される [21]．

BARTScore　は，事前学習済み BART [24][6] を用いて生成文の品質を測定する指標であり，生成文が参照文に対してどの程度自然で意味的に近いかを評価する [22]．具体的には、生成文を参照文として再構成する際の確率を用いてスコアを計算する．

FactKB　では，知識ベースを用いて Factualy Training を行ったモデル [7] を用いて，要約が事実かどうかを [0, 1] でスコアリングする [23]．

6）https://huggingface.co/facebook/bart-large-cnn
7）https://huggingface.co/bunsenfeng/FactKB