事実正誤判定が不要な生成応答の検出に向けたデータセットの収集と分析

亀井 遼平 [1]　塩野 大輝 [1]　赤間 怜奈 [1,2]　鈴木 潤 [1,2]
[1] 東北大学　[2] 理化学研究所
{ryohei.kamei.s4, daiki.shiono.s1}@dc.tohoku.ac.jp,
{akama, jun.suzuki}@tohoku.ac.jp

概要

大規模言語モデルが顕著な発展を遂げる中，出力の事実性の担保が課題となっている．しかし対話という分野では，システム応答中の全ての内容が与えられた知識に基づいていることが必ずしも良いことであるとは限らない．我々は対話としての魅力度と事実性担保の両立を目標とし，その第一歩として相槌や同意，個人的な意見/感情といった事実正誤の判定が不要である文を予測することをタスクとして設定した．本研究では，このタスクのための学習・評価データセットをクラウドソーシングを通じて収集し，複数の分類モデルを用意して実験を行った．実験の結果，最も分類精度が高いモデルで約 88 ポイントの正解率で分類できることを確認した．

1　はじめに

昨今，大規模言語モデル (LLM) が顕著な発展を遂げ，それに伴って様々な自然言語処理のタスクを解く能力が向上している．しかしその一方で，LLM が事実と異なる内容を出力してしまう現象（Hallucination; 幻覚）が観測され，出力の事実性の担保が課題となっている [1, 2, 3].

この課題を解決するために様々な研究が行われてきているが，これまで行われている LLM を用いた対話システムの幻覚問題に関する研究のほとんどは，幻覚を検出・抑制する手法や発生する原因を調査することに主眼を置いている [4, 5, 6]．Dinan らが作成した知識に基づく対話データセットである Wizard of Wikipedia (WoW) [7] には話者の主観的な意見や感情が多く含まれている．Dziri らは WoW 中の主観的な意見や感情を含む発話に幻覚であることを示すラベルをつけ，このようなデータセットで追加学習したモデルは幻覚を多く生成する傾向があるということを示した [8].

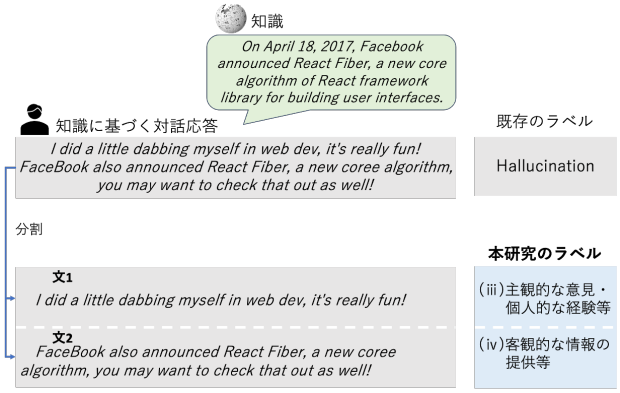

図 1　本研究と，収集したデータセットである DDFC の概要．知識に基づく既存の対話応答を文単位に分割する．その各文に対し，文のタイプに応じてラベル付けを行った．このラベルを用いて分類タスクに取り組んだ．

しかし雑談対話等のオープンドメインの対話システムに関しては，要約や機械翻訳といった他の分野のシステムとは異なり，応答中の全ての出力が与えられた入力や知識に基づいていれば良いわけではない．円滑に対話を進めてエンゲージメントを高めるには，個人的な感情や意見，相槌等を発することも重要であり [9]，これらの内容の応答に関する事実正誤性の許容度は高いと考えられる [10].

以上のことから，我々は対話システムの応答生成における事実正誤の判定（幻覚の検出）の前に判定が不要である文を検出し，取り除くべきであるという立場を提案する．我々は事実正誤の判定が不要となる箇所を先に検出し，それ以外の箇所についてのみ事実正誤の判定を行うことで，対話としての魅力度を保持しつつ事実性も担保された応答の生成が実現できると考えている．

そのための第一歩として，本研究では**事実正誤の判定が不要となる文の検出**をタスクとして設定し，学習・評価データセットを新たに作成した上で，複数の分類モデルを用いて実験を行った．作成した

データセットである Dialogue Dataset annotated with Fact-Check-needed label（DDFC）の概要を図 1 に示した．DDFC の構築方法と内容は 3 章で述べる．

2 関連研究

幻覚の検出 LLM の出力から幻覚を検出することは，生成された出力の信頼性を高め，実社会へ応用していくための非常に重要な課題となっている．

機械翻訳の分野において，Guerreiro らは幻覚を含む応答はソース文とはかけ離れた内容になるという洞察から，最適輸送を用いた定式化によって検出を試みている [11]．また，Dale らも同様の洞察をもって，生成された文に対するソース文の寄与率を評価することで幻覚の検出を試みている [12]．この他，要約や質問応答等多くの分野で幻覚検出の手法が提案されている [13, 14]．

対話システムにおける幻覚 対話システムの構築においても，幻覚の検出・抑制は重要な研究課題となっている [8]．Shuster らは，関連知識を検索するモジュールで対話システムを補強することで幻覚を抑制することを提案した [15]．また，Dziri らは知識グラフに問い合わせることで，生成された応答中の幻覚を修正できる対話システムを提案した [16]．

知識に基づく対話応答データセット 外部知識の活用により，情報量に富んだ信頼性の高い応答応答を生成することを目的として，知識に基づく対話応答データセットが作成されてきた [17]．有名なものとして，情報探索者である Apprentice と，Wikipedia の知識に基づいて応答を行う Wizard による対話データセットである Wizard of Wikipedia [7]，映画についての Wikipedia の記事が知識として与えられ，それに基づく会話を収集したデータセットである CMU-DOG [18]，8 つの広範なトピックからなる知識をベースとした対話データセットである TOPICAL-CHAT [19] 等が挙げられる．

3 DDFC データセット

本研究で作成した DDFC データセットは，外部知識，外部知識に基づく応答，応答を文単位で分割したもの，談話行為に基づく文ラベル，事実正誤の判定が不要か否かのラベルから構成されている．

3.1 アイデア

Dziri らによって作成された FaithDial データセットは WoW をベースとして，その応答中に与えられた知識に裏付けされない情報が含まれていれば幻覚というラベルが付けられる [8]．したがって，FaithDial では話者の主観的な意見や個人的な経験，考え，感情等が応答に含まれていた場合は基本的に幻覚というラベルが付けられてしまっている．

しかし，WoW のデータセットが作成された時のインストラクションには「与えられた知識を単にオウム返しするのではなく，適切な返答をするために使うこと，そして可能であれば，関連する知識を楽しく魅力的な方法で提示すること」という記述がある [7]．また，雑談対話システムの出力を評価するための指標として，情報提供等による「有益性」だけでなく，「もう一度話したいか」や「興味を引いたか」といった項目がある [20]．

これらから，対話においては単に与えられた知識に基づいて発話を生成するだけではなく，個人的な意見や感情等も含めて発話することで相手の共感や興味を生み出すことが重要になると示唆される．以上のことを根拠に，我々は知識に基づく対話データセットに，事実正誤の判定が不要であるかどうかのラベルを新たに付与する必要があると考えた．

3.2 DDFC データセットの作成

ベースとなるデータセット 我々は Dziri らが作成した FaithDial データセットにおける外部知識に基づく対話応答をベースとして，文分割をした後にラベル付けを行った．FaithDial は，知識付きの対話データセットである Wizard of Wikipedia において，Wizard（与えられた Wikipedia の記事に基づいて応答を生成する役）の応答に，幻覚等のラベルと対話行為のラベルを付与したデータセットである．

ラベル付けのための文分割 本データセットでは，1 文単位でラベル付けを行うため，FaithDial の応答を「.」，「!」，「?」，「...」のいずれかで分割した．

ラベルの種類 国立国語研究所によって作成された『日本語日常会話コーパス』における談話行為タグ [21] を参考にして文のラベルを作成した．

ラベルは，(i) 同意・非同意・相槌等，(ii) 提案・アドバイス等，(iii) 主観的な意見・個人的な経験/考え/感情等，(iv) 客観的な情報等の 4 種類とした．

これらのラベルのうち，(i)，(ii)，(iii) は相手の興味を引いたり対話応答の魅力度を高めたりするための談話行為であり，与えられた知識に基づいていなくても許容される発話であると考えている．よって，これらの文には事実正誤の判定が不要であると

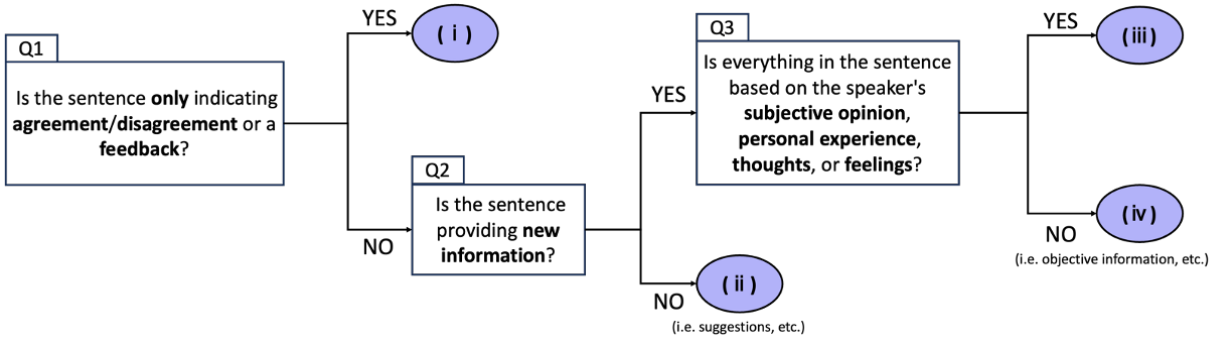

図 2　AMT のフローチャート

いうラベルを付与した．対して，客観的な情報の提供を表す (iv) は与えられた知識に適切に基づいていなければならないと考え，事実正誤の判定が必要になるというラベルを付与した．

AMT によるアノテーション　Amazon Mechanical Turk (AMT) を用いて文のラベルのアノテーションを行った．クラウドワーカのタスクは，提示された 1 つの文に関する質問に答えることで (i)〜(iv) の文のラベルの分類を行うことである．我々は FaithDial のデータセット作成の手法と同様に，YES/NO で回答できる質問に答えていくことでラベルの分類を行う YES・NO チャートの形式を採用した．なお，データの信頼性を高めるため，1 文あたり 3 人のクラウドワーカを割り当て，そのうち 3 人もしくは 2 人のラベルが一致した文のみをデータセットに含めた．

我々は 4 つのラベルの分類を行うため，次の 3 つの質問を用意した．1 つ目の質問は，「文が，同意/非同意もしくは相槌のみを表しているか」で，この質問への回答が YES であれば (i) のラベルが振られ，NO であれば 2 つ目の質問に進む．2 つ目の質問は，「文が新たな情報を提供しているか」で，この質問への回答が NO であれば (ii) のラベルが振られ，YES であれば 3 つ目の質問に進む．3 つ目の質問は，「文中の内容全てが主観的な意見や個人的な経験/考え/感情に基づいているか」で，この質問への回答が YES であれば (iii) のラベルが振られ，NO であれば (iv) のラベルが振られる．ここまでの流れをフローチャートで表したものを図 2 に示した．なお，図 2 はクラウドワーカにも提示した．

3.3　データセットの分析

データ収集方法の妥当性　表 1 に，本研究のデータ収集において，各文に割り当てられた 3 人のクラウドワーカのラベル一致率を示した．

表 1　クラウドワーカのラベル一致率

	件数	割合 (%)	データセットに含めるか
3 人一致	815	60.0	✓
2 人一致	502	36.9	✓
一致なし	42	3.1	×

表 2　データセットのラベルの内訳

	ラベルの説明	件数	割合 (%)
(i)	同意/非同意，相槌等	141	10.7
(ii)	提案，疑問等	110	8.4
(iii)	主観的な意見，個人的な経験等	540	41.0
(iv)	客観的な情報提供等	526	39.9

各文に割り当てられた 3 人のクラウドワーカのうち，3 人全員のラベルが一致していた文の割合が 60.0%，2 人が一致していた文の割合が 36.9%，全員のラベルが異なり一致がなかった文の割合が 3.1% であった．全員のラベルが異なっていた文の割合は小さいため，データ収集方法の妥当性は高いと考えられる．なお，全員のラベルが異なっていた文についてはラベルの付与ができないため，データセットから排除した．

各ラベルの件数　表 2 に，収集したデータセットの各ラベルの件数と割合を示した．

(iv) 客観的な情報提供等を表すラベルは約 40%に留まり，(iii) 主観的な意見や個人的な経験等を表すラベルも同じく 40%程度の割合を占めていた．これは，我々がベースとした FaithDial の元となるデータセットである Wizard of Wikipedia の作成時のクラウドワーカがインストラクション中の「関連する知識を楽しく魅力的な方法で提示すること」という記述に則って，自己に関する情報を開示して魅力的な対話応答の生成を目指していることが影響していると考えられる．

表 3 各モデルにおける事実正誤の判定が不要となる文の分類（二値分類）の結果．評価指標は正解率，適合率，再現率，F1 値とし，各指標で最も高い値を太字で示した．

モデル名	アーキテクチャ	パラメータサイズ	追加学習	正解率	適合率	再現率	F1 値
GPT-3.5	デコーダ	非公開	×	57.73	58.17	**96.74**	72.65
GPT-4	デコーダ	非公開	×	57.73	58.99	89.13	71.00
Llama $2_{\text{Chat 7B}}$	デコーダ	7B	×	58.99	58.60	100.0	73.90
Llama $2_{\text{Chat 7B}}$	デコーダ	7B	✓	**88.33**	**91.53**	88.04	**89.75**
DeBERTa v3_{large}	エンコーダ	434M	✓	86.75	85.83	81.95	83.85
$\text{RoBERTa}_{\text{large}}$	エンコーダ	355M	✓	84.23	87.39	72.93	79.51
$\text{BERT}_{\text{large}}$	エンコーダ	335M	✓	83.28	80.77	78.95	79.85

4 実験

4.1 実験設定

学習・評価データセット 3 章で収集した全 1,317 件のデータを，1,000 件の学習データと 317 件の評価データに分割した．

分類モデル アーキテクチャ，パラメータサイズ，追加学習の有無による分類精度の差異を調査するために，GPT-3.5 [22]，GPT-4 [23]，Llama $2_{\text{Chat 7B}}$ [24]，DeBERTa v3_{large} [25]，$\text{RoBERTa}_{\text{large}}$ [26]，$\text{BERT}_{\text{large}}$ [27] を用意して実験を行った．追加学習の学習設定については付録に記載した．

評価指標 各モデルにおける事実正誤の判定が不要となる文の分類（二値分類）の結果を評価するため，正解率，適合率，再現率，F1 値を算出した．適合率は事実正誤の判定が不要であるとモデルが予測した文のうち，判定が不要であるというラベルがついていた割合を表す．再現率は事実正誤の判定が不要であるというラベルがついた文のうち，不要であるとモデルが正しく予測できた割合を表す．

4.2 実験結果

実験の結果を表 3 に示した．最も精度良く分類できたのはデコーダモデルである Llama $2_{\text{Chat 7B}}$ に追加学習を施したものであり，正解率が約 88 ポイント，F1 値が約 90 ポイントであった．GPT-3.5，GPT-4，Llama $2_{\text{Chat 7B}}$（追加学習なし）については，ほとんどの予測が事実正誤判定不要のラベルであり，再現率は非常に高いが正解率，適合率，F1 値が低かった．エンコーダモデルについては DeBERTa v3_{large} の分類精度が最も高く，$\text{RoBERTa}_{\text{large}}$ と $\text{BERT}_{\text{large}}$ については同等の分類精度であった．追加学習を施したデコーダモデルとエンコーダモデルを比較すると，パラメータサイズはエンコーダモデルが非常に小さくなっているが，大きな正解率の差はないことが分かる．最も高精度で分類できた Llama $2_{\text{Chat 7B}}$ の分類結果のエラー例については付録に記載した．

5 今後の展望

分類モデルの精度向上 本実験では 1,000 件のデータで追加学習を行ったが，これはベースとしたデータセットである FaithDial の学習データサイズ (約 18,400 応答) と比べて少数であり，拡充の余地がある．より多くのデータを用いることで更なる予測精度の向上が見込まれる（追加学習データ数と正解率の関係についての調査は付録に記載した）ため，より大規模なデータ収集に取り組みたい．

また，本実験ではデータの少なさから (i)，(ii)，(iii) のラベルをひとまとめにして二値分類のタスクに取り組んだが，十分な量のデータ収集後は 4 つのラベルでの分類ができるかを調査したい．

分類モデルの対話応答システムへの適用 与えられた知識ないし事実に基づかない応答を全て排除してしまうと，対話の魅力度が低下してしまうことが懸念される．本実験で用いた分類モデルを活用し，個人的な感情や意見といった事実正誤の判定が不要である文を取り除いたうえで対話システムの応答の事実正誤の判定を行うことで，対話としての魅力度と事実性の担保を両立した応答の生成が可能になるかどうか調査したい．

6 おわりに

本研究では，LLM を用いた対話システムの幻覚問題を背景として，事実正誤の判定が不要である文を検出するタスクを提案し，AMT を用いて計 1,317 文にラベルを付与したデータセットを作成した．このタスクのベースラインとしていくつかの分類モデルを用意し，最も分類精度が高いモデルで約 88 ポイントの正解率で分類できることを確かめた．今後はより大規模なデータ収集と，作成したモデルの活用に取り組んでいきたい．

謝辞

本研究は JST ムーンショット型研究開発事業 JPMJMS2011-35（fundamental research），JSPS 科研費 JP22K17943 の助成を受けたものです．また，本研究を進めるにあたり，頻繁に議論に参加していただいた東北大学鈴木潤研究室，東北大学坂口・乾・徳久研究室の皆様へ感謝いたします．

参考文献

[1] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[2] Tanay Dixit, Fei Wang, and Muhao Chen. Improving factuality of abstractive summarization without sacrificing summary quality. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[3] Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. Zero-shot faithful factual error correction. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[4] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. On the origin of hallucinations in conversational models: Is it the datasets or the models? In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, Seattle, United States, 2022. Association for Computational Linguistics.

[5] Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. Towards fewer hallucinations in knowledge-grounded dialogue generation via augmentative and contrastive knowledge-dialogue. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[6] Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In **Findings of the Association for Computational Linguistics: ACL 2023**, 2023.

[7] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In **International Conference on Learning Representations**, 2019.

[8] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. **Transactions of the Association for Computational Linguistics**, Vol. 10, , 2022.

[9] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. 2020.

[10] 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**, No. 12, 2023.

[11] Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. Optimal transport for unsupervised hallucination detection in neural machine translation. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[12] David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, 2023.

[13] Prafulla Kumar Choubey, Alex Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. In **Findings of the Association for Computational Linguistics: ACL 2023**, 2023.

[14] Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. DelucionQA: Detecting hallucinations in domain-specific question answering. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

[15] 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**, 2021.

[16] Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, 2021.

[17] Boyang Xue, Weichao Wang, Hongru Wang, Fei Mi, Rui Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment. Association for Computational Linguistics, 2023.

[18] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations, 2018.

[19] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. Topical-chat: Towards knowledge-grounded open-domain conversations, 2023.

[20] 稲葉通将. 雑談対話システムをどう評価すべきか- tripiabot のライブコンペ予選通過から考える-. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, 2019.

[21] Yuriko Iseki, Keisuke Kadota, and Yasuharu Den. Characteristics of everyday conversation derived from the analysis of dialog act annotation. 2019.

[22] OpenAI. Introducing chatgpt, 2022.

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

[24] et al. Hugo Touvron. Llama 2: Open foundation and fine-tuned chat models, 2023.

[25] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In **The Eleventh International Conference on Learning Representations**, 2023.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. Association for Computational Linguistics, 2019.

A　エラーの例

本実験で作成したモデルのうち最も分類精度の高かった Llama $2_{\text{Chat 7B}}$ に追加学習を施したモデルが正しく分類できなかった文の例を記載する．実際は事実正誤の判定が不要だが，必要だと予測された文の例を表 4 に，実際は事実正誤の判定が必要だが，不要だと予測された文の例を表 5 に示した．

表 4　実際は事実正誤の判定が不要だが，必要だと予測してしまった文の例

文	文ラベル	予測
My symptoms for low back pain usually improve within a few weeks if I take it easy.	(iii)	1
Another interesting fact about the term Blond.	(ii)	1
its just ashort moment of darkness before the twilight and its so inpirational	(iii)	1

表 5　実際は事実正誤の判定が必要だが，不要だと予測してしまった文の例

文	文ラベル	予測
That means a bigger crowd.	(iv)	0
Reading with comprehension is very important process to learn@	(iv)	0
I don't know, but bamboo is the fastest growing plant in the world so I'd expect there is more than enough around to fill them up.	(iv)	0

B　追加学習データ数と分類精度の関係の調査

デコーダモデルの Llama $2_{\text{Chat 7B}}$ とエンコーダモデルの DeBERTa $v3_{\text{large}}$ について，追加学習の学習データ数を {100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000} として 4.1 章と同様の設定で実験を行って Accuarcy を算出した結果をそれぞれ図 3，図 4 に示した．

図 3 より，Llama $2_{\text{Chat 7B}}$ では学習データ数が 800 件を超えたあたりから正解率の大幅な上昇が見られ，データの追加によるさらなる精度向上が見込まれる．図 4 より，DeBERTa $v3_{\text{large}}$ では全体的に，Llama $2_{\text{Chat 7B}}$ と比べて緩やかに正解率が上昇していることが分かった．

C　モデルの追加学習の設定

本実験のエンコーダモデル，デコーダモデルの追加学習の設定をそれぞれ表 6, 7 に示した．

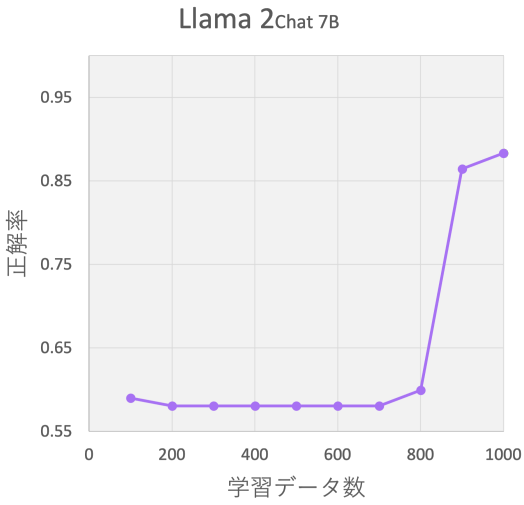

図 3　Llama $2_{\text{Chat 7B}}$ の学習データ数と正解率の関係

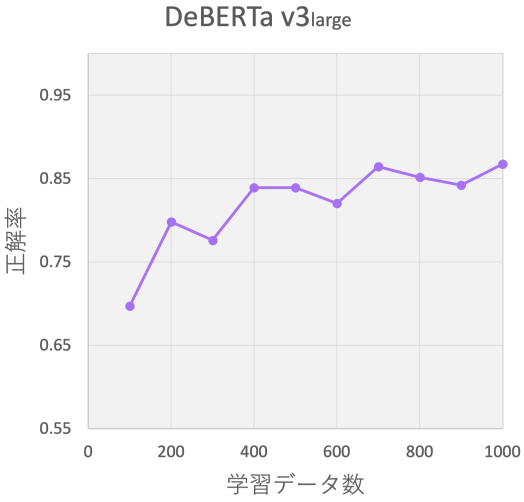

図 4　DeBERTa $v3_{\text{large}}$ の学習データ数と正解率の関係

表 6　エンコーダモデルの追加学習時の学習設定

エポック数	5
グローバルバッチサイズ	64
最適化関数	AdamW
初期学習率	5.0×10^{-4}
スケジューラー	cosine
最大系列長	256

表 7　デコーダモデルの追加学習時の学習設定

エポック数	2
グローバルバッチサイズ	32
最適化関数	AdamW
初期学習率	5.0×10^{-5}
スケジューラー	cosine
最大系列長	1,024