Detecting Response Generation Not Requiring Factual Judgment

Ryohei Kamei[1]   **Daiki Shiono**[1]   **Reina Akama**[1,2]   **Jun Suzuki**[1,2]

[1] Tohoku University   [2] RIKEN

{ryohei.kamei.s4, daiki.shiono.s1}@dc.tohoku.ac.jp,
{akama, jun.suzuki}@tohoku.ac.jp

Abstract

With the remarkable development of large language models (LLMs), ensuring the factuality of output has become a challenge. However, having all the contents of the response with given knowledge or facts is not necessarily a good thing in dialogues. This study aimed to achieve both attractiveness and factuality in a dialogue response for which a task was set to predict sentences that do not require factual correctness judgment such as agreeing, or personal opinions/feelings. We created a dataset, dialogue dataset annotated with fact-check-needed label (DDFC), for this task via crowdsourcing, and classification tasks were performed on several models using this dataset. The model with the highest classification accuracy could yield about 88% accurate classification results.

1 Introduction

Large language models (LLMs) have undergone considerable development and can solve various natural language processing tasks. However, they output content that is different from the fact, i.e., hallucination, making it difficult to ensure the factuality of the output (Zha et al., 2023; Dixit et al., 2023; Huang et al., 2023).

Although hallucination in dialogue systems using LLMs has been extensively studied, they focused on methods for detecting/suppressing hallucinations and investigated the causes of their occurrence (Dziri et al., 2022b; Sun et al., 2023; Ji et al., 2023b). Wizard of Wikipedia (WoW), a knowledge-based dialogue dataset created by Dinan et al. (2019), contains many subjective opinions and feelings of the speaker. Dziri et al. (2022a) labeled utterances in WoW datasets that contained subjective opinions and feelings as hallucinations and showed that models fine-tuned on WoW datasets produce more hallucinations. However, for open-domain dialogue systems such as chatbots, unlike systems in other fields such as summarization or machine

Figure 1: Overview of the study and the collected dataset, DDFC. The existing dialogue responses based on knowledge are divided into sentences. Each sentence was annotated labels according to its type and used in a classification task.

translation, not all output in a response are based on a given input or knowledge. To promote smooth dialogue and increase engagement, expressing personal feelings and opinions is important (Huang et al., 2020). Moreover, the tolerance of factual correctness regarding the response of these contents is high (Ji et al., 2023a).

To address these issues, we propose that sentences not requiring factual correctness judgment should be detected and removed before judgment (hallucination detection) during response generation in dialogue systems. By detecting such sentences first and judging the factual correctness of remaining sentences, responses that maintain the attractiveness of the dialogue can be generated while ensuring the factuality of the dialogue.

First, we set the task of detecting sentences that do not require factual correctness judgment, and created a new dataset. Then, the dataset was
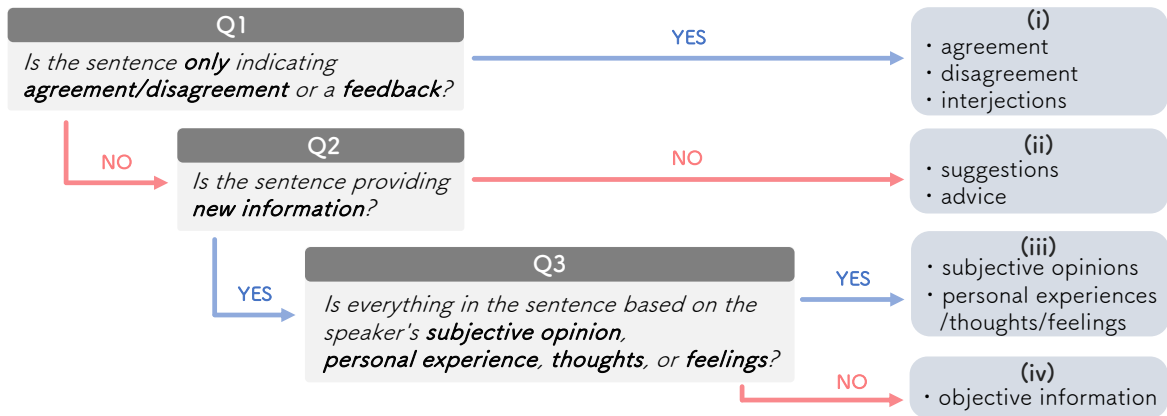
Figure 2: Flowchart of annotation by Amazon Mechanical Turk to construct DDFC.

validated using classification models. Figure 1 overviews the created dataset, **dialogue dataset annotated with fact-check-needed label (DDFC)**. The construction method and contents of DDFC are described in Section 3.

2 Related Work

2.1 Hallucination Detection

Hallucinations from an LLM output must be detected to improve the reliability of the generated output and apply LLMs to real-world applications. Guerreiro et al. (2023) detected hallucinations in machine-translated outputs by formulating them using optimal transport based on the insight that responses containing hallucinations are distant from the source sentences. Similarly, Dale et al. (2023) detected hallucinations by evaluating the contribution of the source sentence to the generated sentence. Various other methods for detecting hallucinations have been proposed in many fields such as summarization and question answering (Choubey et al., 2023; Sadat et al., 2023).

2.2 Hallucination in Dialogue System

Detection and suppression of hallucinations are crucial for constructing dialogue systems (Dziri et al., 2022a). Shuster et al. (2021) suppressed hallucinations by augmenting a dialogue system with a module that retrieved relevant knowledge. Dziri et al. (2021) also proposed a dialogue system that could modify hallucinations in the generated responses by querying the knowledge graph.

2.3 Knowledge-Grounded Dialogue Dataset

Knowledge-based dialogue datasets have been created to generate informative and reliable responses by leveraging external knowledge (Xue et al., 2023) such as WoW (Dinan et al., 2019). The WoW dataset contains dialogues between an apprentice, an information seeker, and a wizard who responds based on his knowledge of Wikipedia. CMU-DOG is another dataset containing conversations based on Wikipedia articles about movies given as knowledge (Zhou et al., 2018), and TOPICAL-CHAT is a knowledge-based dialogue dataset on eight broad topics (Gopalakrishnan et al., 2019).

3 DDFC dataset

The DDFC dataset created herein contained external knowledge, responses based on external knowledge, responses split by sentences, sentence labels based on discourse acts, and labels to determine whether factual correctness judgments are required. We used four types of labels, and crowdworkers assigned them through annotation based on the flowchart (Figure 2).

3.1 Idea

The FaithDial created by Dziri et al. (2022a) was based on WoW, wherein a response was labeled as hallucination if it contained information not supported by the given knowledge. In other words, if the speaker's subjective opinion, personal experience, thoughts, or feelings are included in the response, it is labeled as hallucination in this dataset. However, the WoW dataset was created based on this instruction: "use the given knowledge to provide an appropriate response, rather than simply parrot it, and, if possible, present relevant knowledge in a fun and engaging way" (Dinan et al., 2019). Moreover, to evaluate the chatbot system output, not only "usefulness" by providing information but also metrics such as "whether the user

	# of sample	rate(%)	included
three labels matched	815	60.0	✓
two labels matched	502	36.9	✓
no matched	42	3.1	✗

Table 1: The label match rate of Crowdworker when annotating DDFC dataset. Since there were only a few instances of no match, the validity of the data collection method was considered high. Sentences with no match were excluded from the dataset.

	explanation	# of sample	rate(%)
(i)	agreement, feedback etc.	141	10.7
(ii)	proposal, adivice etc.	110	8.4
(iii)	subjective opinions etc.	540	41.0
(iv)	objective info etc.	526	39.9

Table 2: The Number of samples and the percentage of each label in DDFC we created. Sentence label (iii) and (iv) each account for approximately 40% of the total.

wants to talk again," "whether the user is interested" are used (Inaba, 2019).

Thus, generating utterances based on given knowledge and drawing the users' interest and empathy by expressing personal opinions and feelings are crucial for dialogue systems. Therefore, the knowledge-based dialogue dataset was annotated with a new label that indicated whether a factual correctness judgment was required.

3.2 Construction of the dataset

Base dataset of DDFC. The dialogue responses based on external knowledge in the FaithDial were labeled after splitting them into sentences. Faith-Dial labels the responses of the Wizard (generates responses based on a given Wikipedia article) with hallucination and dialogue act labels in the WoW dataset.

Sentence split for label annotation. In the DDFC, FaithDial responses were split by {'.', '!', '?', '…'} to label them in one-sentence units.

Label types. Sentence labels were created with reference to the discourse act tag in the "Corpus of Everyday Japanese Conversation" created by the National Institute for Japanese Language and Linguistics (Iseki et al., 2019). We used the following four types of labels: (i) agreement, disagreement, interjections, etc.; (ii) suggestions, advice, etc.; (iii) subjective opinions, personal experiences/thoughts/feelings, etc.; and (iv) objective information, etc. Responses that are labeled as

parameter	encoder	decoder
number of epochs	5	2
global batch sizes	64	32
optimizer	AdamW	AdamW
learning rate	5.0×10^{-4}	5.0×10^{-5}
scheduler	cosine	cosine
max length	256	1,024

Table 3: Fine-tuning settings for the classification models used in this study.

(i), (ii), and (iii) were considered dialogue acts intended to attract user interest or increase the attractiveness of the dialogue response. Therefore, they are acceptable even if they are not based on given knowledge and were labeled as not required factual correctness judgment. In contrast, responses labeled as (iv) that provided objective information must be appropriately based on the given knowledge; therefore, they were assigned the label of requiring a factual correctness judgment.

Sentence label annotation by AMT. We used Amazon Mechanical Turk (AMT) to annotate sentence labels. The task of the crowdworker was to classify the labels of sentences (i)–(iv) by answering questions about a given sentence. A YES/NO chart format, similar to the FaithDial creation method, was used, in which labels were classified by answering questions that can be answered with a YES/NO. To increase data reliability, three crowdworkers were assigned per sentence, and only sentences with matching labels from two or three annotators were included in the dataset. The following three questions were used to classify the four sentence labels. (1) "*Is the sentence only indicating agreement/disagreement or feedback?*" If the answer is YES, then assign label (i); if NO, then proceed to the second question. (2) "*Is the sentence providing new information?*" If the answer is NO, then assign label (ii); if YES, then proceed to the third question. (3) "*Is everything in the sentence based on the speaker's subjective opinion personal experience, thoughts, or feelings?*" If the answer is YES, then assign label (iii); if NO, assign label (iv). Figure 2 shows a flowchart of the annotation process, which was also presented to the crowdworker while they were working on the task.

3.3 Analysis of the dataset

Validity of dataset annotation. Table 1 shows the label match rates for the three crowdworkers assigned to each sentence during data collection.

model	architecture	parameter size	fine-tuning	accuracy	precision	recall	F1-Score
GPT-3.5	decoder	no data	✗	57.73	58.17	96.74	72.65
GPT-4	decoder	no data	✗	57.73	58.99	89.13	71.00
Llama 2$_{\text{Chat 7B}}$	decoder	7B	✗	58.99	58.60	**100.0**	73.90
Llama 2$_{\text{Chat 7B}}$	decoder	7B	✓	**88.33**	**91.53**	88.04	**89.75**
DeBERTa v3$_{\text{large}}$	encoder	434M	✓	86.75	85.83	81.95	83.85
RoBERTa$_{\text{large}}$	encoder	355M	✓	84.23	87.39	72.93	79.51
BERT$_{\text{large}}$	encoder	335M	✓	83.28	80.77	78.95	79.85

Table 4: Results of the classification of sentences that do not need to be judged as factually correct or incorrect in each model (binary classification). The highest value in each index is shown in bold.

Of the three crowdworkers assigned to each sentence, 60.0% of the sentences had all three labels in matching, 36.9% of the sentences had two labels in matching, and 3.1% of the sentences had all different labels and no match. As the percentage of no match was small, the validity of the data collection method was considered high. Sentences with no match were excluded from the dataset because labels could not be assigned to them.

Number of each labels. Table 2 shows the number of samples and percentage of each label in the dataset. (iv) Objective information etc., accounted for approximately 40% and (iii) subjective opinions, personal experiences/thoughts/feelings, etc. accounted for approximately 40%. This is possibly because when creating the base dataset WoW for FaithDial, the crowdworkers aimed to generate engaging dialogue responses by disclosing information about themselves in accordance with the statement in the instructions to "present relevant knowledge in a fun and engaging way."

4 Experiment 1: Classification

We prepared some classification models and experimentally evaluated the results of the classification (binary classification) of sentences that do not require factual correctness judgment.

4.1 Experimental Settings

Dataset. The 1,317 collected data were divided into training and test datasets containing 1,000 and 317 responses, respectively.

Classification models. To investigate the differences in the classification accuracy based on model architecture, parameter size, and fine-tuning, experiments were conducted using GPT-3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2023), Llama 2$_{\text{Chat 7B}}$ (Touvron et al., 2023), DeBERTa v3$_{\text{large}}$ (He et al., 2023), RoBERTa$_{\text{large}}$ (Liu et al., 2019), and

BERT$_{\text{large}}$ (Devlin et al., 2019). Table 3 lists our fine-tuning settings.

Evaluation Metrics. To evaluate the results of the classification of sentences that do not require factual correctness judgments (binary classification) in each model, the accuracy, precision, recall, and F1-Score were calculated. Precision is the percentage of sentences predicted by the model as do not require factual correctness judgment and labeled as judgment not required. Recall is the percentage of sentences labeled as factual correctness judgment not required that the model correctly predicted as sentences that do not require judgment.

4.2 Results

Table 4 shows the results of the experiment. The highest classification accuracy was achieved with fine-tuning on the decoder model, Llama 2$_{\text{Chat 7B}}$, with an accuracy of approximately 88 points and an F1-Score of approximately 90 points. For GPT-3.5, GPT-4, and Llama 2$_{\text{Chat 7B}}$ (without fine-tuning), most predictions were labels that did not require factual correctness; they had very high recall but low accuracy, precision, and F1-Score. For the encoder models, DeBERTa v3$_{\text{large}}$ had the highest classification accuracy, whereas RoBERTa$_{\text{large}}$ and BERT$_{\text{large}}$ had almost the same accuracy. A comparison of the decoder and encoder models with fine-tuning shows that the parameter sizes were considerably smaller for the encoder model; however, the percentage of accuracy did not differ considerably.

Tables 5 and 6 show examples of sentences that could not be correctly classified by Llama 2$_{\text{Chat 7B}}$ with fine-tuning, i.e., the model with the highest classification accuracy. Table 5 shows the examples of sentences that do not require a factual correctness judgment, but were predicted to require one, and Table 6 shows examples of sentences that required a factual correctness judgment but were

sentence	label	pred.
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

Table 5: Examples of sentences that do not require a factual correctness judgment but were predicted to require one.

sentence	label	pred.
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.	(iii)	1

Table 6: Examples of sentences that require a factual correctness judgment but were predicted to not require one.

predicted to not require one.

5 Experiment 2: Relation between train data amount and accuracy

The relation between the training data amount for fine-tuning and classification accuracy was investigated by conducting an experiment.

5.1 Experimental Settings

The decoder model, Llama $2_{\text{Chat 7B}}$, and the encoder model, DeBERTa v3$_{\text{large}}$, were used in this experiment. The same settings as in Section 4.1 were used with {100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000} as the number of training data for fine-tuning, and the accuracy was calculated.

5.2 Results

Figure 3 shows the results of each model. The accuracy rate of Llama $2_{\text{Chat 7B}}$ increases considerably when the number of training data exceeds 800, indicating that further improvement in accuracy can be expected using additional data. Overall, the accuracy of DeBERTa v3$_{\text{large}}$ gradually increased compared with that of Llama $2_{\text{Chat 7B}}$.

6 Future Directions

Improving the performance of classification models. Herein, fine-tuning was performed on 1,000 data, which is a small amount compared to the training data size (about 18,400 responses) of the base dataset, FaithDial. Thus, the dataset can be expanded. As further improvement in classification accuracy can be expected by expanding the dataset, future studies will involve large-scale data collection. It may also clarify the reason for the sudden increase in accuracy when the number of training
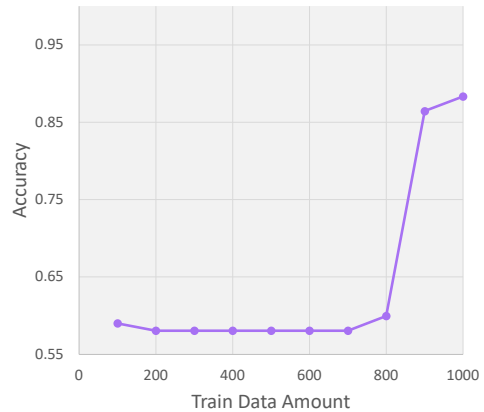
(a) Llama $2_{\text{Chat 7B}}$
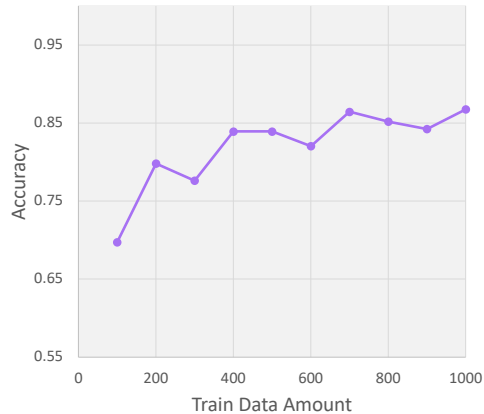
(b) DeBERTa v3$_{\text{large}}$

Figure 3: Relationship between the amount of training data and accuracy. The accuracy of Llama $2_{\text{Chat 7B}}$ significantly improves with over 800 training data, suggesting that more data will lead to even higher accuracy. Overall, DeBERTa v3$_{\text{large}}$ showed a steady increase in accuracy compared to Llama $2_{\text{Chat 7B}}$.

data exceeds 800 in Figure 3(a), and whether the trend of gradual increase in accuracy in Figure 3(b)

continues when training data is increased. Moreover, because our dataset was small, the sentence labels (i), (ii), and (iii) had to be treated as a single label, "not required factual correctness judgments," for the binary classification task. After collecting sufficient data, we would like to investigate whether the four labels can be used for classification.

Application of classification models to dialogue response systems. If all responses that are not based on given knowledge or facts are eliminated, the attractiveness of the dialogue will be reduced. By applying the classification models used herein, we would like to investigate whether factual and attractive dialogue responses can be generated by removing sentences related to personal feelings and opinions that do not require factual correctness judgment and then judging.

7 Conclusion

In this study, a task to detect sentences that do not need to be judged as factually correct or incorrect was proposed against hallucinations in a dialogue system using LLMs. We created a dataset containing 1,317 sentences labeled with sentence types using the Amazon Mechanical Turk. Several classification models were developed as a baseline for this task. Results revealed that the best model could classify with an accuracy of approximately 88%. In the future, we would like to collect data on a larger scale and apply the several models trained in this study to the dialogue system.

Ethics Statement

In this study, we created datasets by human workers using a crowdsourcing platform. In all crowdsourcing processes, identities of workers were kept anonymous and only their IDs were disclosed. Moreover, following the terms of use of the crowdsourcing platform, an appropriate exchangeable point reward was provided for workers.

References

Prafulla Kumar Choubey, Alex Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2023. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10755–10773, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. 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 (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tanay Dixit, Fei Wang, and Muhao Chen. 2023. Improving factuality of abstractive summarization without sacrificing summary quality. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 902–913, Toronto, Canada. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. 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, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.

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

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. 2023. Zero-shot faithful factual error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5660–5676, Toronto, Canada. Association for Computational Linguistics.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3).

Michimasa Inaba. 2019. How should we evaluate chat-oriented dialogue systems? *The Japanese Society for Artificial Intelligence Special Interest Group on Spoken Language Understanding and Dialogue Processing (in Japanese)*.

Yuriko Iseki, Keisuke Kadota, and Yasuharu Den. 2019. Characteristics of everyday conversation derived from the analysis of dialog act annotation. In *2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023b. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.

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

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bin Sun, Yitong Li, Fei Mi, Fanhu Bie, Yiwei Li, and Kan Li. 2023. 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 (Volume 2: Short Papers)*, pages 1741–1750, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

122

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7829–7844, Singapore. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

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