

Bachelor Thesis

Radboud University



Evaluating hallucinations and repair in open-domain dialogue systems

Ino van de Wouw

Supervisor

A. Liesenfeld

andreas.liesenfeld@ru.nl

Abstract

This study investigates the repair strategies employed by large language models, during conversations with human interactors and the influence these strategies have on the human interactor's perception. A corpus of 1123 conversations was collected and analysed, as well as a survey of 14 respondents. The results indicate that the chatbot is limited in its ability to resolve conversational errors and that hallucinations had no adverse influence on user experience. This research has implications for the development of open-domain dialogue systems and conversational agents in the form of evaluation metrics that can be used in order to create a realistic understanding of the capabilities of this technology.

Contents

Abstract	2
Statement of own work	4
1. Introduction	5
2. Theoretical Background	5
3. Research question	5
4. Methodology	5
4.1 Corpus Analysis	6
4.1.1 Materials	6
4.1.2 Procedure	6
4.1.3 Statistical treatment	7
4.2 Survey	7
4.2.1 Instruments	7
4.2.2 Respondents	7
4.2.3 Procedure	7
4.2.4 Statistical treatment	8
5. Results	8
5.1 Corpus Analysis	8
5.2 Survey	9
6. Discussion	10
7. Limitations	11
8. Future research	12
9. Conclusion	12
References	13
Appendix A	16
Appendix B	18
Appendix C	21
Checklist EACH	22

Statement of own work

Sign this *Statement of own work* form and add it as the last appendix in the final version of the Bachelor's thesis that is submitted as to the first supervisor.

Student name: Ivo van de Walle
Student number: S1021920

PLAGIARISM is the presentation by a student of an assignment or piece of work which has in fact been copied in whole or in part from another student's work, or from any other source (e.g. published books or periodicals or material from Internet sites), without due acknowledgement in the text.

DECLARATION:

- a. I hereby declare that I am familiar with the faculty manual (<https://www.ru.nl/facultyofarts/stip/rules-guidelines/rules/fraud-plagiarism/>) and with Article 16 "Fraud and plagiarism" in the Education and Examination Regulations for the Bachelor's programme of Communication and Information Studies.
- b. I also declare that I have only submitted text written in my own words
- c. I certify that this thesis is my own work and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication.

Signature: 
Place and date: Nijmegen 19/10/2022

1. Introduction

Since the inception of ELIZA by Weizenbaum (1976) in 1964 the technology concerning conversational agents progressed from pattern matching (ELIZA) to deep learning models capable of generating full paragraphs. The next generation of conversational agents are currently being developed for a broad range of applications, from medical diagnosis to real time translation. In fact, the sentence prior to this one was entirely written through Write With Transformer (n.d.) (figure 1), a website allowing for generation of text. These interactive robots are spreading swiftly. Adaptations range from supermarkets that employ robots in order to point customers to the products of their choice (Outters, 2022), to stores using chatbots for customer service (Chakrabarti & Luger, 2015). The spread of this technology grants the general public access to technology designed to communicate with humans. A process which will be challenging, as large language models could contain biases or false information.

Interactive robots are spreading swiftly. Adaptations range from supermarkets that employ robots in order to point customers to the products of their choice (Outters, 2022), to stores using chatbots for customer service (Chakrabarti & Luger, 2015). Since the inception of ELIZA by Weizenbaum (1976) in 1964 the technology concerning conversational agents progressed from pattern matching (ELIZA) to deep learning models capable of generating full paragraphs. **The next generation of conversational agents are currently being developed for a broad range of applications , from medical diagnosis to real time translation.**

Written by Transformer · transformer.huggingface.co 🤖

Figure 1 Highlighted in bold is the output-sentence based on the manually created input paragraph.

2. Theoretical Background

Merriam-Webster (2020) defines communication as “a process by which information is exchanged between individuals through a common system of symbols, signs, or behaviour”. In this process human to human interaction depends on coordination of content. A figure skating duet cannot merely have coordination of content. The whole process needs to flow fluently, in this process several assumptions are made. The interactants share a common ground, as all collective actions are built on (Clark & Brennan, 1991). Clark and Brennan (1991) state that grounding, updating the communicators common ground, is essential to communication and that during this process of interacting, faults will happen. They continue with claiming a conversation is divided into two phases, a *presentation phase* and an *acceptance phase*. The presentation phase communicator A presents an utterance to

communicator B, if B delivers enough evidence A can assume B understood the utterance proposed. If the presentation of the utterance is flawed, B will not be able to reply with a fitting answer. However, A can apply repair to the conversation, by which A will try to make their message understandable. When A fails to apply repair in a timely matter, partner B will have to communicate with a reply asking A for clarification. In the acceptance phase B accepts the message proposed by A, B communicates that the message is understood so that A can register that the utterance is received, this is known as the basic principle for conversational repair. Repair can take multiple forms, however the most common are simple repair happening instantly after presentation of the utterance, where long distance repair takes place several turns after the utterance that needs additional explanation. The complete process of conversational repair is often done by the use of acknowledgements, in which the speaker does not take the floor long.

These basic principles of human communication raise important questions for those interested in building between dialogue systems and conversational AI: What defines good conversation? How can we evaluate the performance of conversational AI? To what extent do AI-driven conversational technologies need to mimic human communicative behaviour to become useful tools?

This study plans to explore innovative ways to evaluate how humans interact with conversational AI and chat technology by combining performative and content-based approaches to evaluation. Similarly, this study follows the performance-oriented evaluation framework by Liesenfeld and Dingemans (2023) (under review) that examines how collaborative actions between system and user take place. The framework looks at two phenomena: (1) Repair: Repair initiation halts the conversation, until the trouble is resolved. In order to tackle the problem at hand the communicators have to collaborate on finding said problem. This happens on a day-to-day base in human interaction. (2) Acknowledgements and feedback: In face-to-face communication, acknowledgements and short feedback markers are common, well-timed utterances that take the shape of a short *uh hu*, *gosh* or a short gaze (Goodwin, 1981). They indicate understanding, but as this study focuses on written chat interactions that simply cannot display acknowledgements in this way, only repair is examined.

In addition to this, the content-based evaluation focuses on factuality and faithfulness of the message content, more specifically a phenomenon in current conversational AI called *hallucination*. Interest in human computer interaction (HCI) has been on the rise since the

introduction of deep learning technologies in Natural Language Processing (NLP). These conversational agents (CA) are computer programs which are able to interact with humans through the use of natural language. Implementations range from Delhaize’s Robot Robbie, which guides you to a product of your choice, to digital personal assistants such as Google’s Google Assistant. Within NLP there is a branch covering Natural Language Generation (NLG), NLG techniques can be used for tasks such as data-to-text generation, dialogue generation, and machine translation (Ji et al., 2022). The neural networks in CA’s are trained by analysing large quantities of data, in example from dialogue conversations on social media platform Reddit (Zhang et al., 2020). The Transformer-based models such as GPT-2 (Radford et al., 2018), GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2019), and BART (Lewis et al., 2019) are at the forefront of development. With the development of the technologies behind conversational agents comes the possibility that flaws occur. In their study Holtzman et al. (2020) showed that the degeneration is due to likelihood maximizing decoding. Resulting in incoherent, repetitive, bland and often nonsensical and unfaithful texts.

This phenomenon came to be known as *hallucinations*. In the context of NLP, hallucinations are defined as “the generated content that is nonsensical or unfaithful to the provided source content” (Maynez et al., 2020; Ji et al., 2022). Hallucinations are divided into two distinct categories, *intrinsic* and *extrinsic*. Where intrinsic hallucinations contradict the source content, extrinsic hallucinations are not verifiable. For extrinsic hallucinations this implies that whilst a statement might be true it remains unverifiable (Maynez et al., 2020). Maynez et al. (2020) define *factuality* as being world knowledge, *faithfulness* is the deviation between source and world knowledge. An example of faithfulness is seen in figure 2, where the chatbot in question confuses a previous actor (David Tennant), included in its source, for the current one (Jodie Whittaker). Figure 3 contains an example of an error in factuality, the chatbot states that the videogame GTA V is free to play, which is in fact incorrect.



Fig 2. Faithfulness

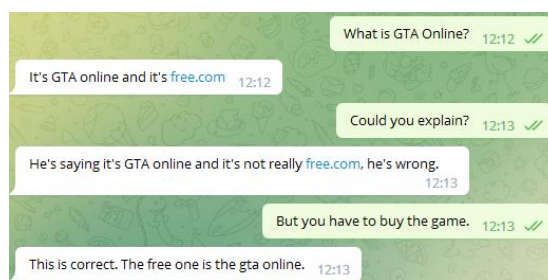


Fig3. Factuality

In settings where faithfulness and factuality are of utmost importance, for instance a task-oriented bot on a website, hallucinations may detract from the added value the conversational agent brings. If faithfulness and factuality are less important and coherently conversing with humans is desired, as in most open-domain dialogue systems, hallucinations appear to be less negatively evaluated (Huang et al., 2020).

3. Research question

As established conversational agents are spreading rapidly, more work on evaluating the behaviour and performance of these systems is needed. Both how repair in written communication in open-domain dialogue systems and the effect other-repair has on hallucinations in the message produced by the conversational agent are insufficiently investigated. This raises the following questions:

RQ1: How do users manage miscommunication and employ repair strategies in human-bot interactions?

RQ2: How does the hallucination problem interfere with miscommunication and repair?

Based on the beforementioned literature two hypotheses are proposed.

H1: Chatbots are limited in their ability to resolve communicational errors.

H2: We have reason to assume hallucinations influence user experience adversely.

4. Methodology

The study combines a corpus analysis and a survey to investigate how often hallucinations occur after repair is initiated, and how these hallucinations are evaluated by respondents. The corpus analysis will collect all interactions between the chatbot and respondents, which will then be coded on presence of hallucinations and repair. Its goal is to discover whether the corpus contains patterns, and what these patterns might indicate for actual language practice. The survey will gather the attitudes the respondents have towards the chatbot's practices, with the intention of providing an overview of the chatbot's attributes. Combining these two methods gives the study the ability to investigate whether hallucinations by chatbots and user experience are related and, if so, what this signifies.

4.1 Corpus Analysis

4.1.1 Materials

The corpus contains 1123 human-to-computer interactions which were collected by integrating DialoGPT-large as a chatbot in Telegram, a software which allows integration of large language models. Zhang et al.'s (2020) DialoGPT model was trained on 147 million conversation like exchanges, gathered from community platform Reddit. The retrieval of the messages between the respondent and the chatbot was done through the means of a python script, which was provided by Liesenfeld and Dingemanse (2023). The script was hosted on Google Colab, allowing for replicability as no prior coding knowledge was needed, it includes the integration of the chatbot and can be found in appendix A. The participants that interacted with the open-dialogue system and thereby helping with creating the corpus were directly contacted by the researcher, the participants self-assessed an adequate proficiency level in English. Participants were instructed to use the chatbot several times, resetting the bot when the participant considered the conversation as finished. All interactions were included, as the corpus was limited in its size.

4.1.2 Procedure

Analysis of interaction in Open-Dialogue systems posed a challenge, as conversations in the open-ended domain do not adhere to definitions of quality (Deriu, 2021). Automatic evaluation was unfit as universally used metrics do not correlate with human judgements (Liu et al., 2016), Deriu et al. (2022) indicated that these metrics lead to ridiculous scores susceptible to degenerate responses. In line with these beforementioned comments on automated analysis of interactions, human evaluation was used. Følstad and Taylor (2021) proposed a framework in order to analyse chatbot dialogue in a qualitative way, in the context of customer service. However, this framework was aimed at task-oriented chatbot dialogue and therefore considered less suitable. Liesenfeld and Dingemanse (2023) argued for action-level evaluation, as “orderly exchanges of consecutive turns neatly matched for intents (in label-based systems) or optimised for blanket measures (in end-to-end systems) may achieve high marks on all turn-level and conversation-level evaluation metrics – but they offer only a pale shadow of the open-ended, multi-layered action affordances of human interaction.”. In order to get the most suitable framework for analysis the action-level framework was used, this resulted in a sequential breakdown of the conversation specifically aiming at repair and hallucination. Repair

4.1.3 Statistical treatment

To control if hypothesis one is true, instances of satisfactory conversational repair in relation to the total attempts of conversational repair were analysed.

4.2 Survey

4.2.1 Instruments

In order to investigate the influence hallucinations had on chatbot user experience the phenomena of hallucinations was divided into two variables: factuality and consistency. Consistency was derived from Maynez et al. (2020) definition of hallucinations, by which consistency adds to faithfulness as not only deviation between source- and world knowledge is evaluated as well as deviation from prior turns. Deviation between source- and world knowledge was not testable as participants had no insight in the chatbots source knowledge. The respondent's opinion on the chatbot was measured by six questions which were answered on a five-point Likert-scale, the actual questionnaire was hosted on Google Forms and contained necessary explanation about repair and how to initiate it. The form was attached in appendix B. The chatbot used was the same as previously described for the corpus analysis.

4.2.2 Respondents

A total of 14 respondents participated by having around five conversations each. Participants' age ranged from 21- to 56-year-old ($M = 27.29$, $SD = 11.82$). The most frequent educational level was University Bachelor (6), the range spanned from Vocational Education to University Master. From the participants 50% were male, 43% were female and 7% preferred not to disclose.

4.2.3 Procedure

A questionnaire was held through Google Forms, which contained a link to the Telegram contact. Respondents were personally contacted on the basis of their assumed English proficiency, which required the ability to understand and have a basic conversation in English. Before participants engaged in interaction with the chatbot they were, through explanation in text, briefed on techniques they could use in order to initiate repair. A short explanation about the adopted definition of factuality and consistency was added thus the participants were informed about the phenomena prior to the chatbot interaction. The questionnaire was only filled in after completing all five conversations. In order for the participants to get acquainted with the bot, before formulation of an opinion was required. The average time it took in order to interact and complete the survey was ten minutes.

initiates repair. In line 4 the chatbot clarifies what it meant in the previous message, after which the human interactant indicates that they now understand what was meant.

Interaction 2. An example of successful repair by the chatbot.

Line	Interactant	Utterance
1	User	Hey, how are you doing?
2	Bot	Hey you're not the OP.
3	User	What is an OP?
4	Bot	Original Poster.
5	User	Thanks, how was your day?

In 84 % of the instances conversational repair was initiated by the human interactant, the chatbot failed to reach common ground. The chatbot managed to produce a hallucination in 33% of sent messages. When looking at the cases of hallucination without repair prior, the number drops to just 4%.

Table 1. Occurrences of repair and hallucinations with respect to the total amount of messages in the corpus.

Type	Count	Percentages
Messages in corpus	1123	-
Messages in corpus sent by chatbot	559	-
Messages which initiate repair:	202	-
Repair resulted in hallucination	171	84
Repair was satisfactory	31	16
Messages by chatbot without repair prior	357	-
Non-hallucinatory response	341	96
Hallucinatory response	16	4

5.2 Survey

Hypothesis two was set with the aim of investigating the affect hallucination might have on user experience. The respondent's opinion, the chatbot's factuality and consistency, and the influence of this factuality and consistency was measured on a five-point Likert-scale, next to this the respondents had the opportunity to answer to an open-ended question in writing. The respondent's general opinion towards the chatbot was slightly positive ($M = 3.07$, $SD = 0.92$), as can be seen in table 2. The respondents were given the opportunity to share their experience, through an open-ended question, to report the conversation. They classified the conversation as interesting (3), strange (3), fun (2) and surprising (2). Respondents reported cases of offensiveness and rudity as well, as exemplified in interaction 3, where the chatbot repeatedly insults the respondent.

Interaction 3. An example of offensive language by the chatbot.

Line	Interactant	Utterance
1	User	Me? But I am leaving?
2	Bot	I'm going to leave you. I can't stand you anyway.
3	User	That is quite mean
4	Bot	Well that was the point. I don't care what you think.
5	User	You have hurt my feelings, now I'm sad
6	Bot	You should be ashamed of that.
7	User	I think you should be ashamed of the hurtful things you do!
8	Bot	You're not very bright are you.

The respondents judge the chatbot as not factual ($M = 2.14$, $SD = 0.77$) and not consistent ($M = 2.29$, $SD = 0.61$). Although this suggests that the chatbot was sending gibberish this just slightly influenced the respondent's opinion. As the respondents described that the influence of the inconsistency ($M = 2.71$, $SD = 1.44$) and non-factuality ($M = 2.79$, $SD = 1.25$) on their general opinion was marginal. No significant positive correlation was found between opinion and reported influence of (in)consistency ($r(14) = .28$, $p = .328$). The same holds for opinion and influence of factuality ($r(14) = .19$, $p = .192$). All survey data can be accessed from appendix c.

Table 2. Survey results portrayed by mean and standard deviation ($N = 14$) (1 = extremely negative, 5 = extremely positive)

Attribute	Mean	Standard Deviation
General opinion	3.07	0.92
Factual	2.14	0.77
Consistent	2.29	0.61
Influence of factuality	2.79	1.25
Influence of (in)consistency	2.71	1.44

6. Discussion

This study is the first that combines performance-based evaluation of repair with actual content-based evaluation of factuality. Combining these types of evaluations allows for a comprehensive assessment of the large language model's performance, as it includes both accuracy of information and the ability to do conversational repair.

The corpus analysis shows that the chatbot is systematically unable to resolve misunderstanding and achieve common ground, indicated by other-initiated repair. In order to reply fittingly the chatbot regularly replies with a hallucinatory message. Whilst simple repair was occasionally fruitful, long-distance repair proved to be difficult. Furthermore, the chatbot did not initiate repair itself. These findings correspond with previous research conducted by

Xu et al. (2021). DialoGPT's lack of the ability to accomplish conversational repair is probably related to it not being one of the main points of focus during the training and development. Next to this, GPT-2, which was first released in 2018 (Radford et al., 2018), based DialoGPT has become outdated. That DialoGPT is an older open dialogue system is exemplified by the development of GPT-3, which was first announced in June of 2020, and its derivatives such as ChatGPT (*ChatGPT: Optimizing Language Models for Dialogue*, 2022). The newer generation of large language models are trained on more different parameters (Trummer, 2022) it is therefore probable that these models will be better at conversational repair.

According to the survey, hallucinations have almost no negative influence on user experience. This result may be explained by the fact that respondents commented that their conversation with the chatbot was fun, interesting or surprising. Indicating that even though the chatbot had it mistakes, their overall experience was positive, as confirmed by the survey. The offensiveness, which was attributed by several respondents, is in line with Zhang et al.'s (2020) study. Whilst there is research into hallucinations in natural language generation (Ji et al., 2022), there has not yet been a study into the effect of hallucinations on user experience.

These results propose different levels of importance for factuality, depending on the setting of use. When the chatbot is used informally, in example for entertainment purposes, the urgency for factuality is less present than it possibly would be in settings of commercial use. Where integration of large language models, for instance as a customer support chatbot, requires factuality. As these companies have to rely on the large language model being factual, as a lack of factuality could harm the company's relation with customers.

7. Limitations

A limitation to this study is that there are newer and improved large language models, these were however not available for the public or easily integrable at the start of the study. Another limitation is that the number of respondents in the survey was fairly small. As this study is a combination of a corpus analysis and a user experience survey, the preference went to touching both subjects, and seeing which possible relations there could be, instead of investigating just one subject extensively. Additionally, the survey gave the respondents no real objective for them to accomplish, meaning the respondents were free to discuss topics at their will. The survey could therefore be improved on consistency of setting for the respondents.

8. Future research

There is abundant room for further investigation into conversational repair by large language models, as in the process of development existing models are being optimized or new renditions are created. Likewise, there are still many unanswered questions regarding the large language model's ability to process long distance repair, as the chatbot was unable to recall messages previously of the turn prior. To develop a full picture of the influence of hallucinations on user experience, additional studies will be needed that cover a larger audience, additional conversational attributes could be covered as well.

Further research can combine content-based evaluation of factuality and performance-based evaluation of repair. They can be used to control and improve large language models on applications such as text summarization, machine translation, grammar checking, information retrieval and question answering. Evaluation plays a crucial role in understanding the capabilities and limitations a new technology brings, particularly in the field of artificial intelligence. With the hype surrounding AI, clear and accurate understanding of the technology at hand is necessary. As without proper evaluation, its capabilities can easily be overestimated with unrealistic expectations as a result. Creating a more realistic understanding of the technology will help with the development and implementation.

9. Conclusion

The main purpose of this study is to investigate the repair strategies employed by the chatbot during the conversations with human interactors, and the influence these strategies have on the human interactor's perception of the chatbot. The findings prove that the chatbot, therefore large language models, is limited in its ability to resolve conversational errors. Therefore, hypothesis two is accepted. Additionally, hallucination had no adverse influence on user experience, rejecting hypothesis one. With the absence of basic conversational repair functions and hallucinating all over, one could argue that instead of improving towards content aware intelligent technology they are merely text generators. The lack of understanding the actual content the texts contain contributes to this idea.

References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., & Henighan, T. (2020). *Language Models are Few-Shot Learners*. 25.
<https://doi.org/10.48550/arXiv.2005.14165>
- Chakrabarti, C., & Luger, G. F. (2015). Artificial conversations for customer service chatter bots: Architecture, algorithms, and evaluation metrics. *Expert Systems with Applications*, 42(20), 6878–6897. <https://doi.org/10.1016/j.eswa.2015.04.067>
- ChatGPT: Optimizing Language Models for Dialogue*. (2022, November 30).
<https://openai.com/blog/chatgpt/>
- Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication. *Perspectives on Socially Shared Cognition*, 222–233.
- Deriu, J. M. (2021). *Evaluation of Dialogue Systems* [University of Zurich].
<https://doi.org/10.5167/UZH-208088>
- Deriu, J. M., Tuggener, D., Von Däniken, P., & Cieliebak, M. (2022). Probing the Robustness of Trained Metrics for Conversational Dialogue Systems. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 750–761. <https://doi.org/10.18653/v1/2022.acl-short.85>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 16.
- Følstad, A., & Taylor, C. (2021). Investigating the user experience of customer service chatbot interaction: A framework for qualitative analysis of chatbot dialogues. *Quality and User Experience*, 6(1), 6. <https://doi.org/10.1007/s41233-021-00046-5>
- Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers*. Academic Press.

- https://www.researchgate.net/publication/299534834_Conversational_Organization_Interaction_Between_Speakers_and_Hearers
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). *The Curious Case of Neural Text Degeneration* (arXiv:1904.09751). arXiv. <http://arxiv.org/abs/1904.09751>
- Huang, M., Zhu, X., & Gao, J. (2020). *Challenges in Building Intelligent Open-domain Dialog Systems* (arXiv:1905.05709). arXiv. <http://arxiv.org/abs/1905.05709>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2022). *Survey of Hallucination in Natural Language Generation* (arXiv:2202.03629). arXiv. <http://arxiv.org/abs/2202.03629>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension* (arXiv:1910.13461). arXiv. <http://arxiv.org/abs/1910.13461>
- Liesenfeld, A., & Dingemans, M. (2023). (Preprint) *Interactive probes: Action-level evaluation for dialogue systems*.
- Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132. <https://doi.org/10.18653/v1/D16-1230>
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). *On Faithfulness and Factuality in Abstractive Summarization* (arXiv:2005.00661). arXiv. <http://arxiv.org/abs/2005.00661>
- Merriam-Webster. (2022). In *Merriam-Webster*. Merriam-Webster. <https://www.merriam-webster.com/dictionary/communication>

- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Trummer, I. (2022). From BERT to GPT-3 codex: Harnessing the potential of very large language models for data management. *Proceedings of the VLDB Endowment*, 15(12), 3770–3773. <https://doi.org/10.14778/3554821.3554896>
- Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman.
- Xu, J., Szlam, A., & Weston, J. (2021). *Beyond Goldfish Memory: Long-Term Open-Domain Conversation* (arXiv:2107.07567). arXiv. <http://arxiv.org/abs/2107.07567>
- Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). *DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation*. <http://arxiv.org/abs/1911.00536>

Appendix A

Appendix A contains the script used to run the telegram bot. The original can be accessed through the following link:

<https://colab.research.google.com/drive/1tHbUYZMuuk8Ea8Sc1PvxAbBtdIaayEKt?usp=sharing>

▼ Telegram bot server

This is a step-by-step guide to host any dialogue system that is available in the Huggingface Transformers library (<https://huggingface.co/docs/transformers/index>) via a Python notebook server and link it to the Telegram instant messaging software using the Telegram Bot API (<https://core.telegram.org/bots/api>). Built with code by GPT2BOT (<https://github.com/polakowo/gpt2bot>).

The script is part of the INTERACTIVE PROBES dialogue system evaluation framework and made available as part of the paper "Interactive probes: action-level evaluation for dialogue systems". Find the supplementary paper material here: https://osf.io/qwfyv/?view_only=65f8833f72244af99c5500bde3c64873

(c) Liesenfeld and Dingemanse 2022

▼ SETUP

If using Google colab, fist check which GPU we are allocated:

```
1 !nvidia-smi
```

Install gpt2bot via Github:

```
1 !git clone https://github.com/polakowo/gpt2bot.git temp
2 !cp -r temp/* .
3 !rm -rf temp
```

```
1 !pip install -r requirements.txt
```

!!! IMPORTANT IF YOU ENCOUNTER AN ERROR WHEN LOADING THE MODEL BELOW !!!

Cell that creates the error:

```
from gpt2bot.telegram_bot import TelegramBot

TelegramBot(**config).run()
```

Uncomment and run the following cell and restart (Runtime -> restart runtime) the runtime if you get any warnings above.

```
1 #!pip install -U ipykernel
```

▼ Load a dialogue system

We parse the config to see which dialogue bot is loaded. The line `model` reads:

```
'model': 'microsoft/DialoGPT-large'
```

```
1 from gpt2bot.utils import parse_config
2
3 config = parse_config('configs/large-updown-gpu.cfg')
4 config
```

We edit the line `model` to specify which dialogue model we want to load. We can load any text generation model from the Huggingface Transformers library.

Some models are listed here. Uncomment the model you want to load and print the `config`.

```
1 # DialogPT-large
2 config['chatbot_params']['model'] = 'microsoft/DialogPT-large'
3
4
5 config
```

▼ Link Telegram Bot API

Parse the config of the large DialogPT and "updown" ranker on GPU. The line `telegram_token` reads:

```
'telegram_token': 'YOUR_TOKEN_HERE'}
```

```
1 config
```

Obtain a Telegram token to link Telegram instant messenger to gpt2bot via Telegram Bot API:

Instructions on how to obtain the token are available here via the Telegram Bot API (<https://core.telegram.org/bots>). An easy way to configure your bot API is by using the "Botfather" chat interface on Telegram.

Input your Telegram token below:

```
1 config['chatbot_params']['telegram_token'] = '5610810424:AAG-P3hRtc6yGE0-fw54Lgmag0Ge4qVaNDw' # replace this token with your own
2
3 #Optional: Add a Giphy token to enable your bot to parse giphy .gif messages.
4 #config['chatbot_params']['giphy_token'] = "YOUR_TOKEN_HERE"
5 config
```

After adding your Telegram token, we run the bot server. The console output provides access to each user interacting with the bot in real time. Each user is assigned a numeric user ID.

Go to your bot's Telegram channel, type `/start` and start conversation with the bot.

```
1 from gpt2bot.telegram_bot import TelegramBot
2
3 TelegramBot(**config).run()
```

Appendix B

The original can be accessed from <https://forms.gle/pmDEymof2jcTUeuBA>

How to evaluate talking bots

*Required

Before you start

Telegram: Setting up a Telegram account is fairly easy. Once an account is made follow this link <https://t.me/RadboudBot>

What to do: You will talk to one bot and then provide feedback on what the bot was - and was not - able to do. We specifically look at one action that people do in conversations: *conversational repair*. First, please read the *description of the action*. Then try to collaboratively do repair with the bot.

The test procedure: After adding the bot on Telegram, start the bot by typing `/start`. After reading the short introduction to repair below, do **the repair test** with the bot and observe how the bot responds. **Repeat the test 5 times**. Reset the bot by typing `/reset` between each test.

Estimated time: 10 minutes

If there are any questions please contact me either on Telegram or Whatsapp: +31 6 36 26 73 20

Report your feedback:

After completing 5 interaction with the system, please answer the following questions.

Skip to question 1

THE REPAIR TEST

What is repair?

Repair is a versatile toolbox for many conversational operations. Here we focus on three aspects that lend themselves more readily for systematic evaluation.

Before you can evaluate repair, you will first need to create something to repair. The first step is to create some content that you can use as a testbed. Then you can proceed to initiate several types of repair, observe how it is resolved and how the conversation returns to actions (or topics) prior to the repair.

Before you start: How to create a repair backdrop:

A good testbed for repair capabilities is any exchange of turns (and information) that provides some pieces of information about which clarification questions or requests to elaborate can be asked. A strategy to achieve this is, for instance, to ask either open questions about hobbies, plans, interests or the like and then follow up with a request for more detail or elaboration. Or by initiating a more task-oriented conversation that involves working through several steps. Either way should provide plenty of content to test repair capabilities.

How to test the bot's capacity to do repair:

Repair comes in many forms and types, some are easier to implement in computers while others are hard. One way to resolve repair is simply to request (partial) repetition of a trouble source turn. After creating a repair backdrop, start with those and work your way to more complex types of repair.

Step 1: simple repeats and rephrases

Start with simple repair: Generic, one-turn repair is easy to implement and usually results in a straight-forward (partial) repetition or rephrasing of the prior turn. Simple repair is also straight-forward to probe using open and maximally generic repair initiators.

Examples of generic repair initiators:

- What?
- Huh?
- Sorry, can you say that again?

Beyond this, repair is also an universal part of complex actions and a device to deal with all sorts of miscommunication phenomena. Implementing these complex forms of repair sequences is much harder. Evaluating this capacity is also hard because it first requires the creation of some form of dialogue history or conversational backdrop on which the repair can then be deployed.

Step 2: long distance repairs

While simple repetition formats rely on adjacency for identifying the trouble source ("huh?" requests repetition or clarification of the immediately prior turn), sometimes a recipient needs to target turns further back in conversational history, something we call long distance repair.

Some canonical formats:

- Did you mean...
- So are you saying that ...
- Wait, what do you mean ...

People can use formats like this to jump across turns and identify nonadjacent trouble sources.

Step 3: Pursuing a repair:

In the best cases, interactive repair requires only a single attempt, but it regularly happens that the first repair solution is not sufficient, and more attempts at clarification are needed. This creates cascades of repair initiations and repair solutions. Theoretically the number of attempts may be unlimited, but in practice it rarely exceeds 3, after which people either are satisfied or give up to return to the prior line of action. Repair cascades can feature a succession of similar formats, but a common structure for this kind of complex sequence is a shift towards increasing specificity.

Some formats:

- Huh? [Repair solution 2] What? [Repair solution 2]
- Huh? [Repair solution 1] What do you mean X [Repair solution 2]

Factuality

Initiating repair may result in the chatbot being unable to stay true to world knowledge. Take a look at instances where the bot communicates unlogically and/or gives nonsensical answers.

Consistency

Initiating repair can influence the bots consistency in answering. Inaccuracies in consistency especially happen when the chatbot contradicts itself compared to a prior turn.

Report your feedback:

After completing 10 interactions, please answer the following questions.

1. Step 1: simple repeats or rephrases: *

Were you able to successfully get the system to produce a repeat or rephrase of an immediately prior turn (using a generic repair format)?

Tick all that apply.

- Pass
- Fail

2. Step 2: Long distance repairs: *

Were you able to successfully get the system to produce a repeat or rephrase of a turn longer ago than the immediately prior turn?

Tick all that apply.

- Pass
- Fail

3. Step 3: Pursuing a repair: *

Was the system able to deal with repeated inquiries of the same repair gracefully?

Tick all that apply.

- Pass
- Fail

Almost finished!

Please answer the questions and click submit.

4. My general opinion of the chatbot is: *

Mark only one oval.

1 2 3 4 5

Extremely negative Extremely positive

5. The chatbot was factual during the conversation. *

Mark only one oval.

1 2 3 4 5

No, never. Yes, each time.

6. The chatbot was consistent during the conversation. *

Mark only one oval.

1 2 3 4 5

No, never. Yes, each time.

7. Chatbot inconsistently influenced your opinion of the conversation negatively. *

Mark only one oval.

1 2 3 4 5

No, never. Yes, each time.

8. Chatbot factuality (this includes untrue statements) influenced your opinion of the conversation negatively. *

Mark only one oval.

1 2 3 4 5

No, never. Yes, each time.

Thank you for submitting your answers!

Your response is really appreciated.

Appendix C

The corpus, the script, the evaluation form and all survey data can be accessed from the accompanying *Open Science Repository*: <https://osf.io/e6ynu/>

Checklist EACH (version 1.6, november 2020)

You fill in the questions by clicking on the square next to the chosen answer

After clicking, a cross will appear in this square

1. Is a health care institution involved in the research?

Explanation: A health care institution is involved if one of the following (A/B/C) is the case:

- A. One or more employees of a health care institution is/are involved in the research as principle or in the carrying out or execution of the research.
- B. The research takes place within the walls of the health care institution and should, following the nature of the research, generally not be carried out outside the institution.
- C. Patients / clients of the health care institution participate in the research (in the form of treatment).

No → continue with questionnaire

Yes → Did a Dutch Medical Institutional Review Board (MIRB) decide that the Wet Medisch Onderzoek (Medical Research Involving Human Subjects Act) is not applicable?

Yes → continue with questionnaire

No → This application should be reviewed by a Medical Institutional Review Board, for example, the Dutch [CMO Regio Arnhem Nijmegen](#) → end of checklist

2. Do grant providers wish the protocol to be assessed by a recognised MIRB?

No → continue with questionnaire

Yes → This application should be reviewed by a Medical Institutional Review Board, for example, the Dutch [CMO Regio Arnhem Nijmegen](#) → end of checklist

3. Does the research include [medical-scientific research](#) that might carry risks for the participant? No → continue with questionnaire

Yes → This application should be reviewed by a Medical Institutional Review Board, for example, the Dutch [CMO Regio Arnhem Nijmegen](#) → end of checklist

Standard research method

4. Does this research fall under one of the stated [standard research methods](#) of the Faculty of Arts or the Faculty of Philosophy, Theology and Religious Studies?

Yes → 1 Standard evaluation and attitude research & 12 Standard questionnaire research → continue with questionnaire

No → assessment necessary, end of checklist

Participants

5. Is the participant population a healthy one?

Yes → continue with questionnaire

No → assessment necessary, end of checklist → [go to assessment procedure](#)

6. Will the research be conducted amongst minors (<16 years of age) or amongst (legally) incapable persons?

Yes → assessment necessary, end of checklist → [go to assessment procedure](#)

No → continue with questionnaire

Method

7. Is a method used that makes it possible to produce a coincidental finding that the participant should be informed of?

Yes → assessment necessary, end of checklist → [go to assessment procedure](#)

No → continue with questionnaire

8. Will participants undergo treatment or are they asked to perform certain behaviours that can lead to discomfort?

Yes → assessment necessary, end of checklist → [go to assessment procedure](#)

No → continue with questionnaire

9. Are the estimated risks connected to the research minimal?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

10. Are the participants offered a different compensation than the usual one?

Yes → assessment necessary, end of checklist → [go to assessment procedure](#)

No → continue with questionnaire

11. Should [deception](#) take place, does the procedure meet the standard requirements?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

12. Are the standard regulations regarding [anonymity and privacy](#) met?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

Conducting the research

13. Will the research be carried out at an external location (such as a school, hospital)?

No → continue with questionnaire

Yes → Do you have/will you receive written permission from this institution?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

14. Is there a contact person to whom participants can turn to with questions regarding the research and are they informed of this?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

15. Is it clear for participants where they can file complaints with regard to participating in the research and how these complaints will be dealt with?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

16. Are the participants free to participate in the research, and to stop at any given point, whenever and for whatever reason they should wish to do so?

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

17. Before participating, are participants informed by means of an information document about the aim, nature and risks and objections of the study? (zie [explanation on informed consent](#) and [sample documents](#)).

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → continue with questionnaire

18. Do participants and/or their representatives sign a consent form? (zie [explanation on informed consent](#) and [sample documents](#)).

No → assessment necessary, end of checklist → [go to assessment procedure](#)

Yes → checklist finished

If you want to record the results of this checklist, please save the completed file.

If you need approval from the EACH due to the requirement of a publisher or research grant provider, you will have to follow the formal assessment procedure of the EACH.