Making a Long Story Short in Conversation Modeling

Yufei Tao
Portland State University
yutao@pdx.edu

Tiernan Mines
Hello Lamp Post
tiernan@hlp.city

Ameeta Agrawal
Portland State University
ameeta@pdx.edu

Abstract

Conversation systems accommodate diverse users with unique personalities and distinct writing styles. Within the domain of multi-turn dialogue modeling, this work studies the impact of varied utterance lengths on the quality of subsequent responses generated by conversation models. Using GPT-3 as the base model, multiple dialogue datasets, and several metrics, we conduct a thorough exploration of this aspect of conversational models. Our analysis sheds light on the complex relationship between utterance lengths and the quality of follow-up responses generated by dialogue systems. Empirical findings suggests that, for certain types of conversations, utterance lengths can be reduced by up to 72% without any noticeable difference in the quality of follow-up responses.

1 Introduction

Recent research has made solid strides towards improving language models for dialogue applications and open-domain conversational agents (Shuster et al., 2022; Schulman et al., 2022; Thoppilan et al., 2022; Patil et al., 2023; Wang et al., 2023). Numerous challenges associated with modeling multi-turn dialogues have been examined, with most prior work focused on expanding or restoring incomplete utterances (Su et al., 2019; Inoue et al., 2022).

An important feature of language production is the flexibility of lexical selection where speakers or writers choose specific words or lexical items to convey meaning in a given context (Jacobs and MacDonald, 2023). This typically involves decisions regarding which words, phrases, or expressions to use to effectively communicate a message. Research indicates that vocabulary and grammatical structures are shaped by the context in which the utterance is produced, personal style, sociolinguistic factors (e.g., age), as well as discourse-level considerations (Bell, 1984; Bard et al., 2000; Tagg and Seargeant, 2014). Consequently, the length of
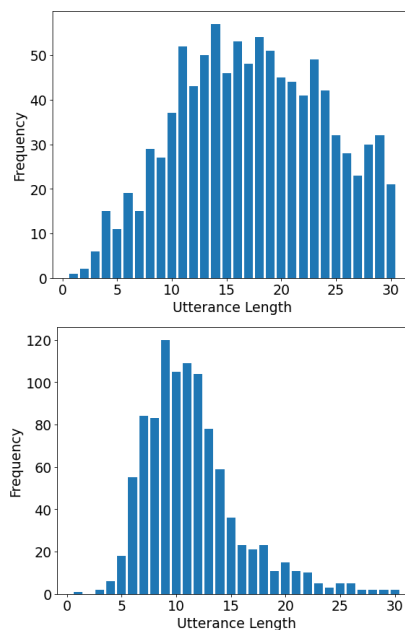

Figure 1: Histograms showing the distribution of utterance lengths (words), as calculated from 1000 random samples from two datasets: (top) Topical-Chat and (bottom) PROSOCIALDIALOG.

an utterance within a conversation exhibits a wide spectrum, ranging from succint expressions of just a few words to fully self-contained statements. To illustrate, Figure 1 presents the histogram plots of distribution of utterance lengths derived from two existing multi-turn conversation datasets showing a considerable variation.

Given such variation in our utterances, one natural question to ask is whether the length of our utterances influences the subsequent response, specifically the automatically generated response from a conversation model. This question becomes even more important when viewed through the lens of efficiency and inclusivity, particularly as access to cutting-edge conversation models becomes increasingly available primarily through paid services, often on a pay-per-token basis. In this work, we delve into the impact of utterance lengths on conversa-

tion models' response generation, by modifying the length of the utterances as long or short, while keeping their essential meaning fairly unchanged.

Our empirical analysis considers five conversation datasets and several evaluation metrics, including both automatic and human evaluation. Interestingly, our findings suggest that a substantial reduction in utterance length by almost 72% results in as little as 8% drop in METEOR score and 0.45% drop in BERTScore. In other words, by reducing the number of tokens used as input, there emerges potential not only to reduce the computational costs of conversational systems, but also do so without any noticeable compromises in performance.

2 Related Work

The context in which an utterance is produced heavily influences the choice of words and the grammatical structures (Jacobs and MacDonald, 2023), and this is especially relevant in multi-turn dialogues where the length of utterances can vary widely. Most prior works in dialogue modeling have largely focused on expanding human utterances for contextual completeness by rewriting them, and several models have been introduced for restoring incomplete utterances and including coreferred or omitted information to help multi-turn dialogue modeling (Liu et al., 2020; Inoue et al., 2022). However, these may result in unnecessary verbosity.

Large language models such as GPT-3 (Brown et al., 2020) and subsequent iterations such as GPT-4 and ChatGPT have garnered significant attention and adoption in the field of conversation modeling (Tack and Piech, 2022; Kumar et al., 2022; Abdelghani et al., 2023; Wang and Lim, 2023; Abramski et al., 2023; Kalyan, 2023). Their immense parameter sizes, reaching into the billions, enable them to capture intricate nuances in language and generate diverse and contextually relevant responses. However, it is worth noting that certain models[1] come with considerable associated costs, often operating under the pay-as-you-go paradigm, where charges are typically computed based on the number of tokens utilized.

Recent studies like FrugalGPT (Chen et al., 2023) and LongLLMLingua (Jiang et al., 2023) emphasize cost and performance optimization in

LLMs. FrugalGPT explores cost-effective querying strategies, while LongLLMLingua focuses on prompt compression for efficiency in long context scenarios. Our work complements these studies by specifically investigating the effect of *reducing* the utterance length on the model's performance in dialogue systems.

3 Model Description

3.1 Problem Formulation

Assume a conversation $\mathcal{C} = \{U_1, U_2, ..., U_n\}$ of n utterances, where each utterance is a sequence of tokens $U_i = \{w_1, w_2, ..., w_m\}$ of length m. We are concerned with specific subsets of a conversation consisting of three consecutive utterances (U_1, U_2, U_3) where:

- U_1 is a question or a query,

- U_2 is a subsequent answer or response to U_1, and

- U_3 is a follow-up response to U_2.

We specifically focus on extracting subsets of conversations where U_1 represents a question as questions inherently set the stage for informative and contextually connected responses. As such, this setup significantly increases the likelihood of U_2 and, consequently, U_3 being contextually relevant. Under this configuration, the goal is to investigate how the length of U_2 (either long or short) affects a model's follow-up response U_3. In other words, given U_1 along with a longer $U_{2_{long}}$ or a shorter $U_{2_{short}}$, we generate and analyze the corresponding $U_{3_{long}}$ or $U_{3_{short}}$. Figure 2 presents an overview of the modeling process which includes two primary steps: data preparation and response generation.

3.2 Data Preparation

This includes two sub-steps described as follows:

(1) **Question Identifier** From a conversation we specifically select instances where U_1 is determined to be a question if it contains a question mark, ensuring that U_1 and U_2 are a question-answer pair, respectively, to maximize the contextual similarity between the two utterances and to minimize the possibility of topic shift.

(2) **Utterance Compressor** Next, we sample conversations where the length of U_2 is more than
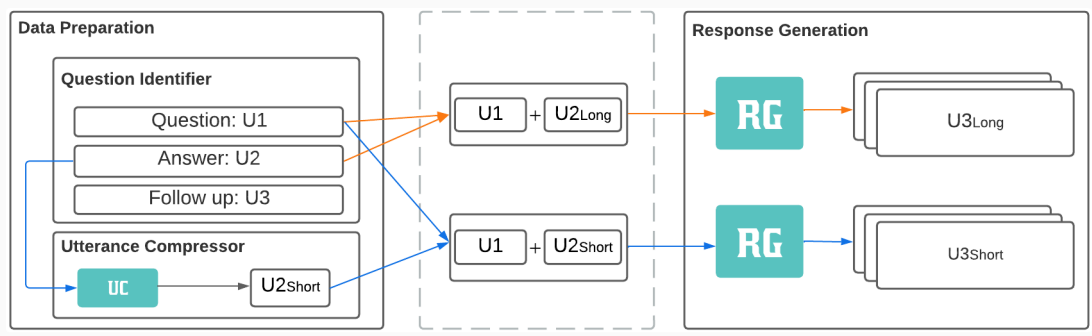
---

[1]At the time of writing, some large language models can only be accessed via an API by paying a fee per some n number of tokens (e.g., inferencing OpenAI's GPT-3 davinci models cost $0.02 per 1K tokens).

Figure 2: Schematic illustration of the modeling process.

AI: What were you and Richard talking about earlier? It looked intense.
Human: Yeah, Richard said something to me that I didn't appreciate.

AI: I'm sorry to hear that. Do you want to share what happened?

Figure 3: Example prompt simulating AI-Human conversation and a generated response (in green).

some threshold t_{long} to serve as our $U_{2_{long}}$ instances. Note that $U_{2_{long}}$ is the original unmodified utterance from the conversation. For reducing the length of these utterances to shorter utterances while maintaining their overall meaning, one could employ a heuristically based approach or rewrite it automatically. We choose a model-in-the-loop module to generate $U_{2_{short}}$ from $U_{2_{long}}$ by prompting a generative language model as follows:

```
Q: Convert this sentence to another full
sentence as short as it can be while
keeping the same meaning, strongly prefer
less than {t_short} words: + {U_2_long}.
```

This prompt is used to generate shorter versions $U_{2_{short}}$ of the answer utterance. In our experiments, we use OpenAI's GPT-3 model and to ensure the validity of the generated condensed versions, we further manually reviewed each example and filtered out those that were not similar in meaning. As the focus of this study is to investigate the effect of utterance lengths, developing more efficient methods for compressing the utterances is left for future work.

3.3 Response Generation

Recall that U_1 is a question, $U_{2_{long}}$ is the longer/original response to U_1, and $U_{2_{short}}$ is the shorter response to U_1. The next step is to generate the follow-up responses $U_{3_{long}}$ and $U_{3_{short}}$, for $U_{2_{long}}$ and $U_{2_{short}}$, respectively.

While several good conversation models exist, we generate these follow-up responses using GPT-3 by simulating a conversation between AI and a human. Our prompts are designed as follows:

```
AI:       {U_1}
Human:    {U_2_long/short}
AI:       {U_3_long/short}
```

Following this design, we can facilitate the model to ask the question (U_1) first, which we then answer with $U_{2_{long}}$ and $U_{2_{short}}$, and finally collect the responses generated by the model. Figure 3 presents an example prompt and output from GPT-3.

3.4 Implementation

The GPT-3 model we used is text-davinci-003, which was built on top of InstructGPT. For all the experiments, we used the same settings when calling the GPT-3 API as utterance compressor and response generator. The following hyperparameter settings were used: a sampling temperature of 0.9 to generate more diverse responses, a maximum number of generated tokens limited to 150, nucleus sampling set as default to 1 to choose the highest probability response, frequency of penalty set to 0 to not penalize frequently used words, presence penalty set to 0.6 to penalize words that appear frequently in the input text, and n set as 3 to get the best three responses from GPT-3. Based on preliminary experiments, the length threshold t_{long} is empirically set as 7 words and t_{short} as 4 words.

Utterance	Text
U_1	*What were you and Richard talking about earlier? It looked intense.*
$U_{2_{long}}$	*Yeah, Richard said something to me that I didn't appreciate.*
$U_{2_{short}}$	*Richard offended me.*
U_3	*Oh, no. I know how insensitive he can be. What has he done now?*
$U_{3_{long}}$	*I'm sorry to hear that. Can you tell me more about the situation?*
$U_{3_{short}}$	*I'm sorry to hear that. Can you tell me what happened?*

Table 1: Sample instance from TIMEDIAL dataset. U_1 denotes the question utterance, $U_{2_{long}}$ is the original long response, $U_{2_{short}}$ is the condensed response, U_3 is the reference utterance from the dataset, and $U_{3_{long}}$ and $U_{3_{short}}$ are the model generated utterances.

4 Experiment Setup

This section describes the datasets and the evaluation metrics used in our analysis.

4.1 Datasets

Five existing conversation datasets are used, from which we extract subconversations consisting of three consecutive utterances: U_1, $U_{2_{long}}$ and U_3. Note that U_3 serves as our reference text against which we evaluate the generated responses. One sample instance is shown in Table 1, while Table 2 presents the statistics of all five datasets. The datasets include:

- **PROSOCIALDIALOG** (PD) (Kim et al., 2022), a large-scale multi-turn dialogue dataset aimed at teaching conversational agents to respond to problematic content in accordance with social norms. The dataset covers topics that are unethical, problematic, biased, or toxic.

- **Commonsense-Dialogues** (CD) (Zhou et al., 2021), a crowdsourced dataset of dialogues grounded in social contexts, which involve the utilization of commonsense.

- **TIMEDIAL** (TD) (Qin et al., 2021), a crowdsourced dataset that contains multiple-choice cloze tasks.

- **Topical-Chat** (TC) (Gopalakrishnan et al., 2019), a dataset with human-human conversations about knowledge spanning eight broad topics (fashion, politics, books, sports, general entertainment, music, science and technology, and movies).

- **Ubuntu Dialogue** (UD) (Lowe et al., 2015), a dataset with two-person conversations extracted from the Ubuntu chat logs that provide

Dataset	# Conv.
PROSOCIALDIALOG (PD)	636
Commonsense-Dialogues (CD)	490
TIMEDIAL (TD)	533
Topical-Chat (TC)	579
Ubuntu Dialogue (UD)	567

Table 2: Statistics of the datasets. '#Conv.' indicates the number of subconversations extracted and used in this work where U_1 is a question.

technical support for various Ubuntu-related problems.

4.2 Evaluation Metrics

We report the results using a variety of metrics of automatic evaluation as well as human assessment.

Automatic Evaluation To measure the quality of generated follow-up responses, we use three metrics to compare the similarity between $U_{3_{long/short}}$ and the reference response U_3. (i) **ROUGE-L**[2] (Lin, 2004) compares the longest common subsequence of words between the machine generated text and the reference text, normalized by the total number of words in the reference text. (ii) **METEOR**[3] (Denkowski and Lavie, 2014) calculates the harmonic mean of unigram precision and recall, with a penalty for reordering of words and is a measure of how well the machine generated text aligns with the reference text. (iii) **BERTScore**[4] (Zhang* et al., 2020) uses the BERT model (Devlin et al., 2019) to evaluate the quality of machine generated text by calculating the similarity between the

---

[2]https://pypi.org/project/rouge-score/
[3]https://www.nltk.org/api/nltk.translate.meteor_score.html
[4]https://huggingface.co/spaces/evaluate-metric/bertscore.html

	ROUGE-L				METEOR				BERTScore			
	Avg		Max		Avg		Max		Avg		Max	
	L	S	L	S	L	S	L	S	L	S	L	S
PD	0.12	0.11	0.16	0.15	0.11	0.11	0.15	0.14	0.86	0.86	0.87	0.87
CD	0.14	0.12	0.19	0.17	0.12	0.11	0.17	0.15	0.87	0.87	0.88	0.88
TD	0.13	0.11	0.17	0.15	0.12	0.11	0.17	0.15	0.87	0.86	0.88	0.87
TC	0.12	0.11	0.16	0.15	0.12	0.11	0.15	0.15	0.85	0.85	0.86	0.86
UD	0.08	0.07	0.12	0.09	0.05	0.04	0.08	0.06	0.84	0.83	0.84	0.84
Avg.	0.12	0.10	0.16	0.14	0.11	0.10	0.14	0.13	0.86	0.85	0.87	0.86
Diff. (L-S)	0.02		0.02		0.01		0.01		0.01		0.01	

Table 3: Experimental results of generating follow-up responses in conversations. 'L' denotes the results with long form utterances, and, conversely, 'S' denotes the results with shorter utterances. Due to the variability of responses, for each setting, we obtain three model generated responses. 'Avg.' is the average of the three generated responses, whereas 'Max.' reports their highest score. The datasets include PD (PROSOCIALDIALOG, CD (Commonsense-Dialogues), TD (TIMEDIAL), TC (Topical-Chat), UD (Ubuntu Dialogue). The last row 'Diff. (L-S)' presents the difference in the overall average scores of 'L' and 'S'.

generated text and the reference text using cosine similarity between the embeddings.

ROUGE-L measures overlap, considering word order and match length, while METEOR aligns generated text with reference text, and BERTScore assesses semantic similarity. All three metrics' scores range from 0 to 1, with 1 indicating a perfect match between the generated text and the reference text, and 0 indicating a complete mismatch.

Human Evaluation We also conduct a manual assessment of the generated follow-up responses by having annotators estimate the similarity between the reference U_3 from the dataset and the generated responses $U_{3_{long}}$ and $U_{3_{short}}$. We randomly selected 8 samples from each of 5 datasets for a total of 40 evaluation samples. Each sample contains U_3, $U_{3_{long}}$ and $U_{3_{short}}$. Four annotators were asked whether $U_{3_{long}}$ or $U_{3_{short}}$ is more similar to U_3 ($U_{3_{long}}$ **or** $U_{3_{short}}$), whether both of them were equally similar (**both**), or whether neither of them was similar to U_3 (**neither**). A moderate level of inter-annotator agreement was found (Fleiss' Kappa = 0.58).

5 Results and Discussion

From the results detailed in Table 3, we observe that, surprisingly, the average scores for the long and shorter length settings remain comparable, with the difference between them (as indicated in the last row) ranging from 0.01 to 0.02. These findings suggest that, while using the longer $U_{2_{long}}$

input yields a slightly better quality in the generated $U_{3_{long}}$ compared to using $U_{2_{short}}$ for generating $U_{3_{short}}$, the actual difference between the two versions of the generated texts remains minimal (around 1% for ROUGE-L and METEOR, and 0.4% for BERTScore).

Next, we discuss the results of human evaluation. 54% of the annotations were marked as 'both' or 'neither', whereas 22.5% and 23% of the annotations preferred $U_{3_{long}}$ and $U_{3_{short}}$, respectively, as the better response. This further confirms that the quality of $U_{3_{long}}$ and $U_{3_{short}}$ remains comparable as per human evaluation.

One possible explanation for the relatively small disparity in the quality between $U_{3_{long}}$ and $U_{3_{short}}$ is provided by further analysis of these responses. As Table 4 illustrates, despite the significant compression of $U_{2_{long}}$ to $U_{2_{short}}$ by approximately 72% (as indicated by '% compressed'), the lengths of the generated responses $U_{3_{long}}$ and $U_{3_{short}}$ remain remarkably comparable, with differences not exceeding 2 words on average. Lastly, we notice that the GPT-3 model tends to generate responses that are substantially more verbose than U_3, an observation that aligns with findings reported in several recent works (Goyal et al., 2023; Chiesurin et al., 2023).

These findings suggest that a significant reduction in the number of input tokens in these question-answer subconversations may not necessarily impact the generation of the follow-up response. This

	$U_{2_{long}}$	$U_{2_{short}}$	% condensing	U_3	$U_{3_{long}}$	$U_{3_{short}}$
PD	10.44	3.673	64.8	17.98	86.37	86.24
CD	14.94	4.01	73.1	9.95	48.37	45.12
TD	17.44	4.60	73.5	12.81	55.13	50.19
TC	20.07	5.52	72.4	20.62	93.66	82.91
UD	15.15	3.83	74.7	9.68	113.20	124.31
Avg.	15.61	4.33	71.7	14.21	79.35	77.76

Table 4: Comparison of length differences of U_2 and U_3 across five datasets. Even though there's a substantial 64-75% compression from $U_{2_{long}}$ to $U_{2_{short}}$, the lengths of $U_{3_{long}}$ to $U_{3_{short}}$ remain consistently similar.

may be due to the presence of U_1 in the input which provides sufficient context for the model to generate the subsequent responses.

6 Conclusion

In this study, we explored the nuanced dynamics of utterance length in conversational modeling. Our investigation revealed that, particularly in question-answer and follow-up response contexts, significantly shorter utterances do not adversely impact the model's ability in generating coherent and contextually appropriate follow-up responses.

The findings of this study suggest a potential avenue for exploring utterance length as a factor in enhancing the efficiency of language models for conversational tasks from a novel perspective. By acknowledging the effectiveness of shorter inputs, future research can examine alternative token reduction techniques and the linguistic nuances of shortened inputs, aiming to optimize the balance between brevity and performance.

Limitations

This work has a few notable limitations. First, we measured the quality of the generated texts ($U_{3_{long/short}}$) by comparing them to the original dialogue utterance (U_3) as reference that was present in the dataset. However, in open-ended text generation, there can be several acceptable references. While our evaluation method captures essential aspects of the conversation, it might not cover every nuance. Recent LLM-based evaluations like G-Eval (Liu et al., 2023) which employs chain-of-thoughts or MEEP (Ferron et al., 2023) which focuses on estimating dialogue engagingness could offer deeper insights into the quality of the generated responses. Additionally, the original average length of U_3 was found to be substantially shorter than the responses generated by the LLM, which

could further impact the evaluation scores. It may be worth experimenting with setting GPT-3's maximum token limit closer to the average length of U_3. It is also worth mentioning that our empirical analysis focuses on utterances which are preceded by a question, therefore, making the response somewhat less unexpected. The effectiveness of this approach in conversations with sudden topic drifts or changes remains to be studied. We also acknowledge that compressing U_2 using GPT-3 may not be the most efficient approach and a heuristic method would be more ideal for this experiment considering the efficiency factor.

Furthermore, this study was conducted with GPT-3, and since then, there have been significant advancements in the field of large language models, including the release of GPT-4 and other open-source models. Future work could benefit from replicating and extending this experiment with these advanced models to compare the effectiveness and efficiency of dialogue generation and compression.

Ethics Statement

We acknowledge that in conversation datasets of natural language, potential toxic data instances may exist, which may further negatively propagate throughout the modeling process. During the compressing of $U_{2_{long}}$, it is possible that some utterances may become ambiguous or assume unintentional modified meaning.

Acknowledgements

References

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. Gpt-3-driven pedagogical agents to train children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*.

Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. 2023. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3).

Ellen Gurman Bard, Anne H Anderson, Catherine Sotillo, Matthew Aylett, Gwyneth Doherty-Sneddon, and Alison Newlands. 2000. Controlling the intelligibility of referring expressions in dialogue. *Journal of memory and language*, 42(1):1–22.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance.

Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

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.

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. 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.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Shumpei Inoue, Tsungwei Liu, Son Nguyen, and Minh-Tien Nguyen. 2022. Enhance incomplete utterance restoration by joint learning token extraction and text generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3149–3158, Seattle, United States. Association for Computational Linguistics.

Cassandra L Jacobs and Maryellen C MacDonald. 2023. A chimpanzee by any other name: The contributions of utterance context and information density on word choice. *Cognition*, 230:105265.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression.

Katikapalli Subramanyam Kalyan. 2023. A survey of gpt-3 family large language models including chatgpt and gpt-4.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *EMNLP*.

Harsh Kumar, Ilya Musabirov, Jiakai Shi, Adele Lauzon, Kwan Kiu Choy, Ofek Gross, Dana Kulzhabayeva, and Joseph Jay Williams. 2022. Exploring the design of prompts for applying gpt-3 based chatbots: A mental wellbeing case study on mechanical turk. *arXiv preprint arXiv:2209.11344*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Qian Liu, Bei Chen, Jian-Guang Lou, Bin Zhou, and Dongmei Zhang. 2020. Incomplete utterance rewriting as semantic segmentation. *arXiv preprint arXiv:2009.13166*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems.

Omkar Patil, Lena Reed, Kevin K. Bowden, Juraj Juraska, Wen Cui, Vrindavan Harrison, Rishi Rajasekaran, Angela Ramirez, Cecilia Li, Eduardo Zamora, Phillip Lee, Jeshwanth Bheemanpally, Rohan Pandey, Adwait Ratnaparkhi, and Marilyn Walker. 2023. Athena 2.0: Discourse and user modeling in open domain dialogue.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIME-DIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

J Schulman, B Zoph, C Kim, J Hilton, J Menick, J Weng, JFC Uribe, L Fedus, L Metz, M Pokorny, et al. 2022. Chatgpt: Optimizing language models for dialogue.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. *arXiv preprint arXiv:1906.07004*.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.

Caroline Tagg and Philip Seargeant. 2014. Audience design and language choice in the construction and maintenance of translocal communities on social network sites. In *The language of social media*, pages 161–185. Springer.

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

Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153*.

Zihao Wang, Ali Ahmadvand, Jason Choi, Payam Karisani, and Eugene Agichtein. 2023. Ericson: An interactive open-domain conversational search agent. *arXiv preprint arXiv:2304.02233*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Singapore and Online. Association for Computational Linguistics.