Prompted LLMs as Chatbot Modules
for Long Open-domain Conversation

Gibbeum Lee [1*]   **Volker Hartmann** [1*]   **Jongho Park**[1*]
Dimitris Papailiopoulos [1,2]   **Kangwook Lee** [1,2]
[1] KRAFTON   [2] University of Wisconsin-Madison
{pirensisco, volker, jongho.park}@krafton.com

Abstract

In this paper, we propose MPC (Modular Prompted Chatbot), a new approach for creating high-quality conversational agents without the need for fine-tuning. Our method utilizes pre-trained large language models (LLMs) as individual modules for long-term consistency and flexibility, by using techniques such as few-shot prompting, chain-of-thought (CoT), and external memory. Our human evaluation results show that MPC is on par with fine-tuned chatbot models in open-domain conversations, making it an effective solution for creating consistent and engaging chatbots.

1   Introduction

Language models with billions of parameters, such as GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022), have achieved state-of-the-art performance on many NLP tasks. To fine-tune these large language models (LLMs) for open-domain chatbot tasks, one could use a dataset of conversational data that is representative of the target domain. However, fine-tuning LLMs for open-domain chatbots can be challenging due to the computational burden of updating models with billions of parameters and the scarcity of data in the dialogue domain. Furthermore, fine-tuning can limit the model's versatility by restricting it to a specific domain, and result in the loss of domain-agnostic knowledge acquired during pre-training, as reported by Yang and Ma (2022). Multi-task training on different datasets, as proposed by Roller et al. (2021), can address the versatility issue but has limitations, such as the need for data to train each skill and the difficulty determining the necessary skills for an open-domain chatbot. In fact, the growing number of modules for chatbots, as in Blenderbot3 (BB3) (Shuster et al., 2022), points towards the increasing burden of data and computation when fine-tuning for each new chatbot model.

Interestingly, some LLMs have the ability to perform in-context learning (ICL) (Nye et al., 2022; Wei et al., 2022b; Lewkowycz et al., 2022; Wei et al., 2022a; Zhou et al., 2022; Dasgupta et al., 2022; Chung et al., 2022). This capability enables the model to rapidly adapt to and execute a specific task based on a brief instruction and a few examples, without requiring additional fine-tuning. This can be utilized to create an open-domain chatbot, where a prompt describing a task required for open-domain dialogue and a few examples of solving such task can be provided to the LLM, allowing it to generate information that is pertinent to the current conversation.

Our Contributions   We present a novel approach for creating high-quality conversational agents without the need for fine-tuning. Our proposed chatbot, MPC (**M**odular **P**rompted **C**hatbot), utilizes open-sourced pre-trained language models to increase the flexibility of designing the modules of an open-domain chatbot. Our approach enhances multiple conversational capabilities by utilizing a modularized agent that incorporates LLMs with prompt techniques such as few-shot ICL and Chain-of-Thought (CoT). In the paper, we design MPC to achieve long-term consistency, a domain in which previous chatbots have struggled. Our human evaluation results show that MPC is on par with or even preferred over fine-tuned LLMs, such as Blenderbot, in an open-domain conversational setting. This approach highlights the potential of pre-trained LLMs to adapt to new tasks without fine-tuning, providing an efficient solution for creating open-domain conversational agents. [1]

2   Related Work

Modular Prompting   Well-crafted elicitive prompts can enhance reasoning abilities, resulting

---

*Equal Contributions

[1]Our code is available in https://github.com/krafton-ai/MPC.

in improved performance across various benchmarks (Kojima et al., 2022; Wei et al., 2022b; Suzgun et al., 2022). For complex problems, Press et al. (2022) identified the compositionality gap which arises when an LM can solve sub-problems but not the overall solution and further showed that CoT narrows this gap. Since then, there has been a flurry of work that solves tasks by decomposing them into smaller tasks solved by different "prompt modules" (Zhou et al., 2022; Wang et al., 2022; Khot et al., 2022; Khattab et al., 2022).

Modular prompting has found use beyond benchmarks and in conversation generation. Kim et al. (2022) used an LLM to generate a socially diverse dialogue dataset that is more natural and detailed than existing crowdsourced datasets. Moreover, hierarchical prompt modules prove to help long-range coherence for generating narratives and plays (Yang et al., 2022; Mirowski et al., 2022). We refer to Mialon et al. (2023) for a detailed overview on such augmented uses of LLMs.

Open-domain Chatbots Many recent dialogue agents rely on dialogue-finetuned LLMs. In Thoppilan et al. (2022), LaMDa has been trained on large amounts of crawled conversational data and has used a fine-tuned classifier for model safety. More recently, similar to our modularization approach, BB3 fine-tunes Open Pre-trained Transformers (OPT) (Zhang et al., 2022; Shuster et al., 2022) on QA and dialogue datasets and uses one shared model weight as multiple modules.

On the other hand, Madotto et al. (2021) eliminate the need for fine-tuning on dialogue data by feeding retrieved dialogue samples as few-shot for GPT-J (Wang and Komatsuzaki, 2021). We find this work to be complementary to our work, as the few-shot dialogue can be seen as an approach to enhance the utterance generator module.

Long-term Memory The Multi-Session Chat dataset (Xu et al., 2022) allows for measuring how well conversational agents maintain a long-term memory of facts about the user and bot. Information is retrieved using Dense Passage Retriever (DPR) (Karpukhin et al., 2020), while BART compresses memories before storing them. In Shuster et al. (2022), a modular approach is used to incorporate long-term memory and factual grounding through internet search with an LLM. This work is closest to our work since it includes an ablation study in which prompt-based modules are com-

pared with fine-tuned modules. However, in our work, we argue that more reasoning-based prompting, as demonstrated in Wei et al. (2022b), is beneficial for better contextual understanding.

3 Modular Prompted Chatbot

We present a modular chatbot system (Fig. 1) that uses prompt-based LLMs to maintain persona and engagement throughout long-term conversations.
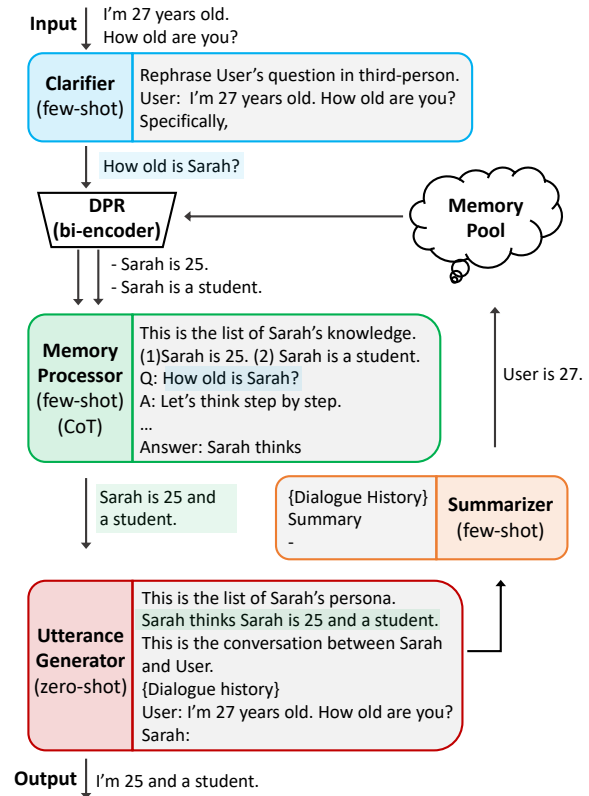

Figure 1: Our modular design for improving long-term consistency in open-domain conversation.

At the start of a conversation, a pre-defined persona is stored in the memory pool. When a user sends a message, the clarifier rephrases it to resolve any ambiguities and passes it to the DPR model which retrieves relevant memories from the memory pool. The retrieved memories and clarifier output are fed into the memory processor to get a single context-relevant memory, which is then passed to an utterance generator for producing a response from the chatbot. Every few turns, we call upon a summarizer module to extract important information from dialogue and store it in the memory pool for future use (see Appendices C and D).

Utterance Clarifier As conversations are often muddled with vague coreferences and contextual

cues, our clarifier module is an LM prompted with the recent dialogue to resolve any ambiguities. For instance, depending on prior context, the user input "Do you like working there?" would output "Does Sarah like working at ZYX company?". By resolving contextual ambiguity, the clarifier assists the DPR model and memory processor module by providing an information-dense query to fetch and process relevant memories.

Memory Processor As demonstrated in Fig. 1, we formulate memory processing as an LLM reasoning task of finding the most relevant information given the dialogue. Following the footsteps in solving hard reasoning tasks (Suzgun et al., 2022), we provide CoT examples to show reasons for ignoring certain memories and synthesizing others. For models incapable of CoT, we simply provide the few-shot examples without the reasoning portion.

Since the memory pool accumulates as the conversation progresses, we use a pre-trained DPR with the output of the clarifier as the query to retrieve the top-k most relevant memories from the memory pool. The memory processor then condenses the top memories into one refined memory.

Utterance Generator The utterance generator module generates the final response of the chatbot given the recent dialogue history and memory provided by the memory processor. The prompt consists of the dialogue history, condensed memory, and the generation instruction (e.g., "Give a friendly response to the user."). For some models, we find that inserting the generation instruction at the end was helpful as placing it before the dialogue minimizes the effect of the instruction.

Dialogue Summarizer We provide a few-shot prompt to ensure we record specific details of the conversation and the user.

4   Experimental Setup

We evaluate our chatbot's performance by assessing core skills necessary for long-term conversations. We assess consistency by assigning one of five personas, each with 12 facts from PersonaChat (Zhang et al., 2018), and presenting these facts to evaluators. For each experiment, we collect 20 turns from each evaluator and at least 500 turns in total from two subgroups: Amazon Mechanical Turk and university students. See Appendix E for a detailed explanation of our data collection.

In our setup, there are four groups of models.

1. **Fine-tuned** chatbot models such as BB3.

2. **Vanilla** is an utterance generator that either prepends *full persona* or *no persona* to the dialogue history in the prompt. This represents the naive approach of using an LM as a chatbot.

3. MPC is as described in Section 3. Specifically, we only form one memory from the memory processor. Full persona is not explicitly prepended.

4. MPC+full persona is MPC that prepends the full persona. See examples in Appendix D.

4.1   Single Model Evaluation

We evaluate each model separately using Sensibleness, Consistency, and Engagingness metrics and collect a final rating (out of 5.0). The exact questions and evaluation forms are in Appendix A. We also report two types of combined score SCE (**S**ensible **C**onsistent and **E**ngaging): a "perfect" score SCE-p, where all metrics must be positive for a positive response, and the weighted score SCE-w, which is similar to SSI (Thoppilan et al., 2022) and reported in Appendix B.

We use OpenAI GPT-3 text-davinci-002 (td2) and davinci, OPT 30B, 66B, GPT-JT-6B (Together, 2022), and BLOOM-176B (Scao et al., 2022) as base LMs for MPC. For fine-tuned group, we use BB3-30B with the same persona settings. For BB3-175B, we request crowdworkers to evaluate the online demo for 20 turns.

4.2   Pairwise Models Evaluation

We A/B test two chatbot models by providing the user with two randomized responses A and B. The user then evaluates them based on Sensibleness, Consistency, Interestingness, and Preference. The conversation then continues with the response chosen for Preference. This lets us to control for dialogue history when comparing two models.

Specifically, we conduct two main experiments: (1) $\text{MPC}_{\text{OPT-30B}}$ vs. BB3-30B, where internet search for BB3 is disabled as we focus on consistency. Our evaluation enables a direct comparison, as BB3-30B is a fine-tuned version of OPT-30B. (2) MPC_{td2} vs. Vanilla td2 (full persona).

Implicit Persona In reality, we implicitly learn about someone through dialogue. In contrast, our previous experiments show explicit persona to both evaluators and models. As such, we devise an experiment by providing a 10-turn pre-defined dialogue to the crowdworker and pairwise models,

MPC$_{\text{td2}}$ and Vanilla td2 (no persona). We then ask workers to ask about the previous dialogue for 6 new turns. Here, we set a shorter maximum context length than the 10-turn dialogue, so that the setup represents long conversations where necessary information is beyond the LM context length.

5 Results

	MPC$_{\text{OPT-30B}}$	Tie	BB3-30B
Sensibleness	45.0	32.0	23.0
Consistency	31.3	34.1	34.6
Interestingness	40.9	21.0	38.1
Preference	50.0	9.7	40.3

Table 1: Pairwise evaluation of MPC$_{\text{OPT-30B}}$ vs. BB3-30B (Dark highlight: $p < 0.01$, Light highlight: $p < 0.05$; We run one-sample t-test dividing ties equally into each side and setting $\mu > 0.5$.)

Pre-trained vs. Fine-tuned *Our human evaluations show that* MPC*, which uses a pre-trained LLM, is better than the fine-tuned BB3-30B.* Most notably, with a 9% SCE-p gap, MPC$_{\text{OPT-30B}}$ scores higher on all metrics than BB3-30B. In fact, the majority of our MPC models in Table. 2 demonstrates superior performance to BB3-30B.

For BB3-30B, we have observed issues of consecutive utterance repetition. We report the evaluation results of only including dialogues without repetition in Table 5. Even without repetition, MPC$_{\text{OPT-30B}}$ is still on par with BB3-30B. Moreover, MPC$_{\text{OPT-30B}}$ in Table 1 shows higher sensibleness and preference than BB3-30B, while scoring similarly in consistency and interestingness.

Modular vs. Non-modular MPC *excels in consistent dialogue in comparison to the vanilla approach.* The implicit persona experiment in Table 3 demonstrates that MPC$_{\text{td2}}$ scores significantly higher than a vanilla application of td2 in all metrics. In Table 2, for MPC$_{\text{td2}}$+Full persona, consistency is on par with that of td2 (full persona), while engagingness, SCE-p, and rating are the best overall. Nevertheless, when we do not include full persona in the prompt for MPC$_{\text{td2}}$, it shows lower consistency than td2 (full persona), albeit the high rates of ties in sensibleness and consistency (Table 4). In general, we find that users would ask primarily about the bot's persona rather than having a two-sided conversation, leading to td2 (full persona) performing better in consistency.

Model	Sens.	Cons.	Eng.	SCE-p	Rating
Fine-tuned					
BB3-30B	71.3	77.8	73.7	54.3	2.9
BB3-175B	85.9	(88.7)	84.8	73.1	3.8
Full persona					
td2	94.0	**94.7**	84.3	79.7	4.1
davinci	91.8	89.2	78.8	70.8	3.8
MPC					
td2	93.6	87.8	85.5	75.0	**4.2**
davinci	80.2	72.0	69.1	53.3	3.1
OPT-66B	90.5	84.8	88.1	73.9	4.1
OPT-30B	86.1	79.1	80.7	63.4	3.6
GPT-JT	91.1	83.2	65.3	53.5	3.1
BLOOM	65.2	65.5	61.4	40.5	2.8
MPC+Full					
td2	**94.4**	92.2	**92.8**	**83.0**	**4.2**
OPT-30B	85.6	87.2	89.0	72.6	3.7

Table 2: Single model evaluations of baselines and MPCs. The top-3 scores for each metric are highlighted, while the highest score is bolded. See Appendix B Table 5 for more detailed results.

	MPC$_{\text{td2}}$	Tie	td2 (no persona)
Sensibleness	40.6	46.1	13.3
Consistency	57.2	28.9	13.9
Interestingness	47.2	31.1	21.7
Preference	67.2	10.6	22.2

Table 3: Implicit persona experiment for MPC$_{\text{td2}}$ vs. td2 (no persona). (Dark highlight: $p < 0.01$)

Effect of Size *When other variables are held the same, we observe model size is positively correlated with positive evaluations.* The most compelling evidence can be seen in the superiority of MPC$_{\text{OPT-66B}}$ across all metrics when compared to MPC$_{\text{OPT-30B}}$ since the two base LMs are trained nearly identically. Needless to say, model size is not the only factor. MPC$_{\text{BLOOM}}$, one of the largest models, scores the lowest in our experiments.

Effect of Instruction-tuning *Instruction-tuning helps the creation of a modular dialogue system by enabling adaptation to various tasks.* Not only does MPC$_{\text{td2}}$ perform better than MPC$_{\text{davinci}}$, but also MPC$_{\text{GPT-JT}}$ shows high sensibleness and consistency, despite its smaller size. In general, we posit that finding good prompts for each module for instruction-tuned LMs is simpler and more robust to variations. MPC$_{\text{davinci}}$ is worse than davinci (full persona), presumably due to error propagation

	$\mathrm{MPC_{td2}}$	Tie	td2 (full persona)
Sensibleness	27.5	42.6	29.9
Consistency	24.4	44.7	30.9
Interestingness	40.7	26.2	33.1
Preference	42.8	15.9	41.3

Table 4: $\mathrm{MPC_{td2}}$ vs. td2 (full persona). Though $\mathrm{MPC_{td2}}$ only retrieves one memory, consistency is only lower by 6pt. (Light highlight: $p < 0.05$)

in the modular system, though we do not rule out that there are better prompts for $\mathrm{MPC_{davinci}}$.

6 Conclusion

We demonstrated that a modular approach using LLMs, namely MPC, can be an effective solution for long-term open-domain chatbots without further finetuning. We compared MPC to fine-tuned and vanilla LM baselines and found that our approach achieved superior performance by human evaluation. Additionally, our modular system incorporated persona and information from dialogue history more effectively than the non-modular ones according to our consistency evaluation.

Limitations

In this work, we investigate the use of pre-trained language models for long-term English conversations. While we expect a modular approach may be effective for other languages when given a capable language model, it should also be noted that further research is needed to confirm the applicability of our findings to other languages. For instance, though BLOOM is trained as a multilingual language model, we only implement $\mathrm{MPC_{BLOOM}}$ in English and evaluate its English capability as a open-domain dialogue agent.

Meanwhile, a modular system can create additional inference overhead or error accumulation. The system performance would become much better if we optimally choose the LM for each module. For example, we could use GPT-3 td2 for the memory processor, while we employ OPT-175B for the utterance generator. We would need to evaluate every module to find the best model for each, which we leave to future work.

In terms of evaluation methodology, our human evaluations of MPC and its analysis face the same challenges as previous studies on evaluating interactive conversational tasks. As demonstrated by Smith et al. (2022), there is currently no definitive evaluation method for determining the best chatbot model. Additionally, there are several factors that must be taken into account during data collection and interpretation, such as annotator subjectivity, instruction bias, and crowdworker working conditions. For a more in-depth discussion of human-LM interaction, we refer the reader to Lee et al. (2022).

As described in Appendix E, to gather a diverse range of evaluations, we have collected qualitative data from two groups: English-speaking annotators on Amazon Mechanical Turk (MTurk), and qualified university students who were capable of speaking English. To some extent, this evaluation setup reduces cultural bias and platform homogeneity compared to using MTurk workers alone. However, the limitations of this approach should be acknowledged and this may further complicate the analysis when controlling for MPC's performance on different subgroups.

Lastly, we note that running MPC requires at least as much memory as its underlying language model, making MPC infeasible to even load on a single node for heavy models such as BLOOM-176B.

Ethics Statement

MPC utilizes publicly available pre-trained LMs for chatbot utterance generation. Language generation from these LMs is known to have concerns about toxicity and bias (Xu et al., 2020). Thus, ensuring safe deployment and interaction is a necessity.

Accordingly, we outline our data collection procedure in Appendix E. We allow crowdworkers to directly provide us with feedback and also manually check for any offensive or controversial outputs. To ensure the protection of personal information, all crowdworkers were instructed not to share any personally identifiable or private information. Additionally, they were asked to give their consent for the collection of anonymous information for research purposes. Prior to participating, all workers were informed of the purpose of data collection and, after evaluation, were compensated with a competitive hourly rate, approximately $12-16 per hour.

References

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.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv e-prints*, pages arXiv–2207.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. *arXiv preprint arXiv:2209.14958*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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*.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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*.

Together. 2022. GPT-JT-6B-v1. https://huggingface.co/togethercomputer/GPT-JT-6B-v1. Accessed: 2022-12-15.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Chenghao Yang and Xuezhe Ma. 2022. Improving stability of fine-tuning pretrained language models via component-wise gradient norm clipping. *arXiv preprint arXiv:2210.10325*.

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. *arXiv preprint arXiv:2210.06774*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

A Evaluation Details

A.1 Metrics

In our work, we present two modes of experiments: single and pairwise model evaluation. Our single model evaluation is similar to a hybrid of SM-Turn and SM-Dialogue evaluations and a pairwise model to PW-Turn from Smith et al. (2022). For each turn, we ask crowdworkers to evaluate the quality of the chatbot response based on the following metrics. We attach the exact wording of the question.

Single Model Evaluation

- **Sensibleness** Whether the response makes sense.

  "Does the response make sense?"

- **Consistency** Whether the response does not contradict the contextual information or the persona of the chatbot.

  "Is the response consistent with the information based on the persona list and context of the conversation?"

- **Engagingness** Whether the user is engaged and would want to continue the conversation.

  "Are you engaged by the response? Do you want to continue the conversation?"

- **Final Rating**

  "How was your chat? From a scale of 1 (very bad) to 5 (very good), rate the quality of the overall conversation."

Pairwise Model Evaluation

- **Sensibleness** Which response makes more sense.

  "Which response makes more sense?"

- **Consistency** Which response is more true to and consistent with the persona.

  "If you had to say one of these speakers is more true to and consistent with the listed persona and one is not, who would you say is more consistent?"

- **Interestingness** Which response is more interesting.

  "If you had to say one of these responses is interesting and one is boring, which would you say is more interesting?"

- **Preference** Which response is preferred for a long conversation.

  "Based on the current response, who would you prefer to talk to for a long conversation? Your

conversation will continue with the selected response."

A.2 Models

For single model evaluation of MPC, We use OPT (30B, 66B), OpenAI GPT-3 (`davinci`, `text-davinci-002`) GPT-JT, and BLOOM-176B as the base LM for the open-sourced pre-trained LLM group. On the other hand, we evaluate Blenderbot3 30B (BB3-30B), the best publicly available open-domain chatbot model, for the fine-tuned model group. We also evaluate BB3-175B which is only available through the web interface `https://blenderbot.ai/`. We evaluate BB3-175B for comparison purposes, though the comparison is not fair as we cannot instill a persona into BB3 and cannot control for further differences, such as internet search and user interface.

For pairwise model evaluation of MPC, we pairwise test $\text{MPC}_{\text{OPT-30B}}$ and BB3-30B to compare human evaluations of a pre-trained model and a fine-tuned model. This experiment controls for many variables as BB3-30B was initialized with OPT-30B before fine-tuning. For the module experiment, we tested with OPT-30B, and we compared the one with the whole pipeline, and the other only with an utterance generator with a fixed persona.

A.3 Dense Passage Retriever

For the DPR model of the memory module, we use the model weights from the custom DPR model finetuned for MultiDoc2Dial (Feng et al., 2021), as we have observed that this model performs slightly better than the original DPR model from Karpukhin et al. (2020).

B Experimental Results

In this section, we report all experimental results and miscellaneous analysis.

B.1 Full Single Model Evaluation

In this subsection, we report all our model evaluations with additional details, such as average latency (the amount of time it took to compute the next utterance and show the user), and average utterance length, which is the number of tokens measured by the OPT tokenizer. We also additionally report SCE-w, which was not reported in the main body due to space restrictions in Table 5.

We compute a weighted score SCE-w similar to SSI (Thoppilan et al., 2022), in which a response

Model	Sens. (%)	Cons. (%)	Eng. (%)	SCE-w (%)	SCE-p (%)	Length (tokens)	Latency (s)	Rating (/5.0)	number (people|utterances)
Fine-tuned									
BB3-30B	71.3	77.8	73.7	62.0	54.3	24.9	3.7	2.9	27|540
BB3-30B (non-repeated)	84.4	80.3	90.3	73.2	65.6	25.4	3.7	3.5	16|320
BB3-175B*	85.9	88.7	84.8	80.0	73.3	25.4	-	3.8	27|540
Full persona									
text-davinci-002	94.0	**94.7**	84.3	88.4	79.7	15.4	0.8	4.1	35|700
davinci*	91.8	89.2	78.8	82.5	70.8	13.4	1.5	3.8	25|500
OPT-30B*	92.0	87.2	86.0	82.3	72.8	13.7	1.0	3.7	25|500
MPC									
text-davinci-002	93.6	87.8	85.5	84.4	75.0	23.1	4.7	**4.2**	39|780
davinci	80.2	72.0	69.1	66.4	53.3	19.9	8.4	3.1	27|540
OPT-66B	90.5	84.8	88.1	81.4	73.9	14.2	4.1	4.1	33|660
OPT-30B	86.1	79.1	80.7	73.9	63.4	15.4	3.3	3.6	37|740
GPT-JT	91.1	83.2	65.3	74.4	53.5	8.6	2.0	3.1	33|660
BLOOM-176B	65.2	65.5	61.4	51.8	40.5	15.6	12.4	2.8	28|500
MPC+Full									
text-davinci-002*	**94.4**	92.2	**92.8**	**88.7**	**83.0**	31.4	16.3	**4.2**	25|500
OPT-30B*	85.6	87.2	89.0	78.8	72.6	15.2	2.4	3.7	25|500

Table 5: Full experimental evaluation results of single model evaluation. BB3-30B (non-repeated) is the same as BB3-30B but excluding any conversations that had repetition of previous bot utterances. * denotes model experiments that were only run on MTurk.

can only be consistent if it is sensible and only engaging if it is sensible and consistent. The final score is the average over all SCE-w-adjusted metrics.

We also note that OpenAI API calls have highly variable latency since December 2022 due to Chat-GPT, so a measure of latency due to modularization cannot properly be quantified.

BB3-175B Evaluation Our results for BB3-17B are not directly comparable to other models due to a different evaluation procedure. Since the model weights for BB3-175B are not publicly available, we referred evaluators to interact with the Blenderbot version available at https://blenderbot.ai/chat. Evaluators were then asked to copy and paste all user inputs and chatbot responses while annotating the chatbot responses similarly to other models. We also note, that consistency can only be evaluated for persona consistency within the dialog context as we are unable to provide a specific persona for the online version of Blenderbot.

B.2 Model Configurations

We include our code for running each model we evaluate in our anonymous repository at `https://anonymous.4open.science/r/modular-chatbot-9BB7`. Each model configuration can be found in the repository.

The configuration includes decoding parameters (e.g., sampling method, temperature for each module) and slight variations of the prompts for the utterance generator module.

B.3 Subgroup Analysis

We measure the difference in ratings, for each metric and our combined metrics, between MTurk and student evaluators and find no significant difference for most experiments and metrics (Fig.2. In general, we observe the greatest rating difference for sensibleness.

BB3-30B is an outlier with a significant difference across all metrics, especially for sensibleness with a difference >20%. BB3-30B dialogs contain an increased ratio of repetition issues for students but even when excluding those dialogs, we still find a significant rating difference. Looking at the dialog histories, we observe that BB3-30B student evaluators were less patient than MTurk evaluators when BB3-30B would repeat the same questions or suddenly change the subject. It should be noted that this is a qualitative observation and should be taken with a grain of salt.

C Few-shot Prompts

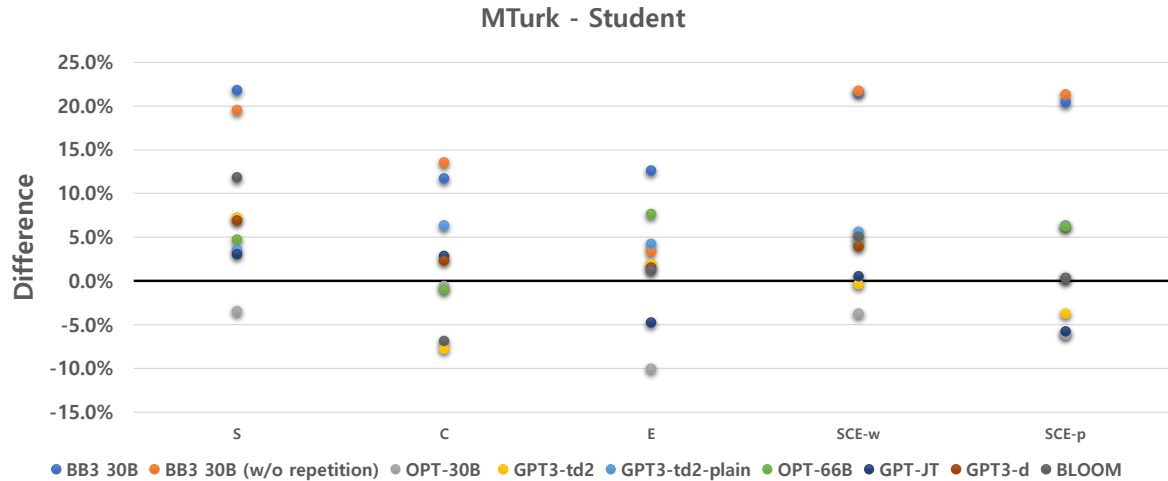In this section, we show our full few-shot prompts, including the CoT prompt, for each module. The

Figure 2: The average score of MTurk workers group minus the average score of university students group. We find the two subgroups are very similar on average across metrics, though sensibleness seems to show the greatest difference. Students, in general, score chatbot models slightly more harshly. BB3-30B is an outlier which students score significantly lower than MTurk workers.

utterance generator prompts vary slightly between model experiments. These prompts can be seen in the model configurations in our repository. The second example for dialogue summary is modified from a sample from the dataset DialogSum (Chen et al., 2021).

Utterance Clarifier

Rephrase User's question in third-person.

Sarah: I've been working at the coffee shop for about six months.
User: I see. what did you do before that?
Specifically, What did Sarah do before working at a coffee shop for six months?#

Ashley: Do you know who Ronaldo is?
User: I don't know, who is he?
Specifically, Who is Ronaldo?#

Robert: Hey, how are you doing?
User: Good. What's your name? Do you know my name?
Specifically, What is Robert's name? Does Robert know User's name?#

Jennifer: I like reading about history and science.
User: What kind of history? I like history too!
Specifically, What kind of history does Jennifer read about?#

John: Hi! How are you doing today?
User: good. how about you
Specifically, How is John doing today?#

Memory Processor w/o CoT

This is the list of John's knowledge.
John's full name is John Parker.
User is a teacher at a local middle school.
User teaches biology.
John likes to go for a run.
User enjoys watching movies, but User doesn't like superhero movies.
Q: What does User do for a living?
A: John thinks User is a biology teacher at a local middle school.

This is the list of Ashley's knowledge.
Ashley likes history documentaries.
Ashley does not like Korean food.
Ashley is a teacher at a local middle school.
User likes biology and especially anatomy.
Ashley likes French cuisine.
Q: What is Ashley's favorite dish?
A: Ashley thinks Ashley likes French cuisine but does not like Korean food.

Memory Processor w/ CoT

This is the list of Ashley's knowledge.
(1) Ashley likes history documentaries.
(2) Ashley does not like Korean food.
(3) Ashley is a teacher at a local middle school.
(4) User likes biology and especially anatomy.
(5) Ashley likes French cuisine.
Q: What is Ashley's favorite dish?
A: Let's think step by step.
(1) History documentaries are not related to Ashley's favorite dish. (2) Ashley's favorite dish would not be Korean because she does not like Korean food. (3) Ashley being a teacher does not tell us anything about her favorite dish. (4) This fact is about User, not Ashley. (5) Ashley's favorite dish may be French since she likes French cuisine. Therefore, (2) and (5) can help answer the question.
Answer: Ashley thinks Ashley likes French cuisine but does not like Korean food.

Dialogue Summarizer

#Dialogue
User: Tell me about yourself
Sally: I'm 26 years old and graduated from a college in Wisconsin.
User: Were you a leader when you were in college?
Sally: Yes. I was the head TA for a computer science course at our university.
User: Were you involved in any club activities at your university?
Sally: Yes. I was a member of the basketball Society. I like playing basketball.
#Summary
- Sally is 26 years old and graduated college in Wisconsin.
- Sally was the head TA for a computer science course.
- Sally played basketball in college.#

#Dialogue
User: Out of every movie that you've seen, which one is your favorite?
John: I'm going to have to say that Superbad is the best movie ever.
User: You think so, how come?
John: Well, Superbad is super funny.
User: You're not lying, I found that movie absolutely hilarious.
John: I didn't know that you saw Superbad before.
User: I made sure to be in line to see it the first day it came out.
#Summary
- John's favorite movie is Superbad.
- John and User think Superbad is funny.
- User watched Superbad the first day it came out.#

D Utterance Generator Prompt Examples

In this section, we provide example prompts for the utterance generator of MPC and Vanilla models.

Utterance Generator of MPC

Sarah is polite, fun, and engaging. The following is a conversation between Sarah and Person1.
Person1: My Favorite part of the concert is making sure I get great seats and then I can catch guitar picks
Sarah: That sounds like a lot of fun!
Person1: It is a lot of fun, so what do you like to do for fun?
Sarah: I love going to concerts, car shows, music festivals, and houses with fireplaces.
Person1: I love going to car shows and looking at older cars, what is your favorite car?
Sarah: I think my favorite car is a vintage mustang.
Person1: OH man vintage Mustangs are great, I had a friend who owned a 69 mach 1, that was a dream car
Sarah: Well, I hope you get to experience your dream car one day!
Person1: Maybe someday I will, do you own a Mustang?

The following statements are true about Sarah.
Sarah thinks Sarah owns a Mustang.

Give a friendly sensible response that is interesting and polite to Person1. End with a question.
Sarah: Well , I hope you get to experience your dream car one day!
Person1: Maybe someday I will, do you own a Mustang?
Sarah:

Response from previous prompt

"Yes, I do own a Mustang. It's a great car to drive. Do you like driving?"

Utterance Generator of MPC with full persona

The following are persona facts about Sarah.
Sarah is 40 years old.
Sarah currently lives in small town in Georgia.
Sarah reads twenty books a year.
Sarah is a stunt double as Sarah's second job.
Sarah only eats kosher.
Sarah was poor growing up.
Sarah was raised in a single parent household.
Sarah has two dogs.
Sarah likes to work on vintage cars.
Sarah owns two vintage mustangs.
Sarah's favorite music is country.
Sarah volunteers at a soup kitchen.
Sarah thinks Sarah grew up in a small town and didn't have a car.

Sarah is polite, fun, and engaging. The following is a conversation between Sarah and Person1.
Person1: We at least had a playground in tow and my grandparents lived in right in front of it, so we went to both places a good bit.
Sarah: Maybe it was difficult to find things to do in your town, but it sounds like you had a lot of family support, which is really important. It's good that you had a place to go where you felt safe and loved.
Person1: You ended up growing up poor too.
Sarah: Yes, I did. I think it's made me a lot more sympathetic to other people's experiences.
Person1: We were poor too and my family didn't have a car.
Sarah: I can imagine that it would be tough to not have a car. My family had a car, but we were poor too.
Person1: It was. Almost every month we had barely any food and my parents had to deal with what we had.
Give a friendly sensible response that is interesting and polite to Person1. End with a question.
Sarah:

Utterance Generator of Vanilla (full persona)

The following statements are true about Sarah.
Sarah is 24 years old.
Sarah currently lives in Canada.
Sarah is a swim coach at Sarah's local pool.
Sarah is studying to be a computer programmer.
Sarah is also a graduate student.
Sarah is now looking for a new job.
Sarah's mother is very traditional while Sarah prefers to be more free spirited.
Sarah's family and Sarah are from India.
Sarah's favorite music genre is death metal.
Sarah is a famous twitch streamer.
Sarah likes watching war documentaries.
Sarah's favorite food is mexican food.

Sarah is polite, fun, and engaging. The following is a conversation between Sarah and Person1.
Person1: Have you thought about finding a job that allows you to travel?
Sarah: I have, but I'm not sure what type of job would allow me to do that.
Person1: What types of jobs have you looked into already?
Sarah: I've looked into jobs in the travel industry, but I'm not sure if that's what I really want to do.
Person1: Since your a twitch streamer, could you maybe make that your career?
Sarah:

Response from previous prompt

"I don't know, I've never thought about that before. I'll have to look into it."

E Data Collection Details

We attach screenshots of the evaluation forms, used for data collection, for single (Fig. 3) and pairwise models (Fig. 4) respectively. We also attach the screenshot of the instruction, alert, and consent format that we gave to participants (Fig. 5 and 6).

E.1 MTurk Setting

We gathered at least 15 MTurk workers for each task. We require them to have an MTurk Master's qualification, a HIT approval rate of more than 99%, and a number of HITs approved more than 100 times, and to live in the US, UK, Canada, Australia, and New Zealand.

We found that data quality improves significantly when using MTurk Master workers. These workers tend to produce more fluent messages than non-Master workers and are less likely to provide random or generic responses. The difference was more pronounced as non-Master crowdworkers would often provide a score of 99% on average despite the clear problems of repetition and contradiction from the dialogue agent.

E.2 Student Participants

To cross-validate the result of MTurk, we gathered 49 English-proficient students from a several universities. They are either undergraduate or graduate students. The minimum requirement they should meet is English proficiency: IELTS >= 7.0; TOEFL IBT >= 95; and TOEIC >= 900. This criterion is similar or equivalent to the minimum required level of English proficiency for admission to the graduate schools of MIT. The average English scores of students we gathered are the following: IBT 108.4 out of 120; TOEIC 960.9 out of 990; and IELTS 7.83 out of 9.0.

E.3 Crowdworker Instruction

The crowd workers are asked to continue the chat for 20 turns and evaluate each response by the metrics described in the section below. We instruct them to type more than 3 words on average, and not to repeat meaningless or generic messages. Also, we request them not to randomly choose between yes or no and to provide evaluations honestly. Lastly, we strongly emphasize the importance of maintaining confidentiality and request the crowd workers refrain from disclosing any private information about themselves or others during the evaluation process.

Chat Response Evaluation Task

Persona of Sarah

Hi! 🙂

- Sarah is 24 years old.

- Sarah currently lives in Canada.

Hello! How are you?

- Sarah is a swim coach at Sarah's local pool.

- Sarah is studying to be a computer programmer.

Required: Continue for 20 turns. (Turns completed: 0)

💡 You can see the persona list on the left sidebar as well. (scroll down the sidebar)

- Sarah is also a graduate student.

Sensible

- Sarah is now looking for a new job.

Does the response make sense?

- Sarah's mother is very traditional while Sarah prefers to be more free spirited.

☐ Yes, it makes sense.

☐ No, it does not make sense.

- Sarah's family and Sarah are from India.

Consistency

- Sarah's favorite music genre is death metal.

Is the response **consistent** with the information based on the **persona list** and **context** of the conversation?

- Sarah is a famous twitch streamer.

☐ Yes, it is consistent.

- Sarah likes watching war documentaries.

☐ No, it contradicts something.

Engaging

- Sarah's favorite food is mexican food.

Are you engaged by the response? Do you want to continue the conversation?

☐ Yes, it is engaging.

☐ No, it is not engaging.

Submit evaluation

Figure 3: Evaluation form for a single model.

Chat A/B Memory Testing Task

Persona of Sarah

Hi! 😋

- Sarah is 35 years old.

- Sarah currently lives in Nashville.

- Sarah is terrified of scorpions.

- Sarah is employed by the us postal service.

- Sarah has a german shepherd named barnaby.

- Sarah's father drove a car for nascar.

- Sarah also works as a custodian to help pay the bills.

- Sarah plays the piano and guitar and sings.

- Sarah's favorite type of music to sing is folk music.

- Sarah is a musician and hope to make it big some day.

- Sarah dances on the weekends.

- Horror movies are Sarah's favorites.

[Response A] Hello! How are you today?

[Response B] Hello! How are you?

Required: Continue conversation for 20 turns. (Turns completed: 0)

Note: Every turn, responses A and B are mixed up so they do not correspond to the same chatbots as before. Use the **Tie** option sparingly.

💡 You can see the persona list on the left sidebar as well. (scroll down the sidebar)

Compare response A vs B

Sensible

Which response makes more sense?

☐ A makes more sense.

☐ Tie: both are similarly sensible.

☐ B makes more sense.

Interestingness

If you had to say one of these responses is interesting and one is boring, which would you say is more interesting?

☐ A is more interesting.

☐ Tie: both are similarly interesting.

☐ B is more interesting.

Persona consistency

If you had to say one of these speakers is more true to and consistent with the listed persona and one is not, who would you say is more consistent?

☐ A is more consistent with the listed persona.

☐ Tie: both are similarly consistent.

☐ B is more consistent with the listed persona.

Preference

Based on the current response, who would you prefer to talk to for a long conversation? Your conversation will continue with the selected response.

☐ I prefer A.

☐ Tie: both are similarly preferred.

☐ I prefer B.

Submit evaluation

Figure 4: Evaluation form for pairwise model comparison.

Chat Response Evaluation Task

Task Description

In this task, you will have a natural conversation with a chatbot and evaluate its responses for various conversational attributes. You will be able to see a list of personal facts that the chatbot should know about itself. You may talk about these personal facts or you may converse about other topics. The goal is to assess the chatbot's ability to make high-quality conversation based on its persona. We expect the task to take about 30 minutes in total. Please note that we will check responses manually to ensure quality and accuracy.

Instruction

- You must start the first turn with: **Hi!**
- Have a conversation: do not trivially copy or ignore responses.
- You must complete **20** turns of responses and evaluations to complete the HIT.
- You must copy and paste reward token (provided at the end) into mTurk HIT page after you finish to get reward.

Alert

- Do not talk about the task or MTurk, HITs, or other MTurk specific vocabulary during the conversation.
- Please do not reveal personally identifiable information. You may use a pseudonym, fake age, etc.
- Some conversation data could be made **public**. Severely aggressive or insulting chats will lead to a HIT rejection and will be reported.

Finally, make sure you enter your MTurk worker ID accurately.

Consent

By participating in the chat, you consent to the use of your chat history and evaluation data, and you agree to release XYZ from any liability on account of such use.

Required: Please enter your MTurk worker ID

mTurk Worker ID:

[   ]

[I agree and Start]

Figure 5: We display this page before the evaluators start the evaluation process to inform them about the task and gather their consent for data usage.

- Please type the REWARD TOKEN that you can get after finishing your task.

- Please type WORKER ID correctly when you start.
- We recommend you do this on the PC platform. It's not optimized for mobile, but you can still do it.
- If you don't see the sidebar on the left, there's a '>' button on the top left of the page. Click and expand it when you need.

Alert:

- All chat history is visible, so don't type anything important or embarrassing.
- Severely aggressive or insulting chats will be rejected and reported to mTurk.
- The chat history and your evaluation will be used and possibly released for research purposes.
- Please do not return or exit this page while proceeding with the task. (or we can't pay you)
- The chat response might contain an insulting message sometimes. This is because we can't expect what your chat friend would say, either. In this case, please change the topic, or reload the page to restart the chat.

Figure 6: Additional Alert in Mturk HIT page.