# Investigating the Representation of Open Domain Dialogue Context for Transformer Models

**Vishakh Padmakumar**[1]* **Behnam Hedayatnia**[2] **Di Jin**[2] **Patrick Lange**[2]
**Seokhwan Kim**[2] **Nanyun Peng**[23] **Yang Liu**[2] **Dilek Hakkani-Tur**[2]

[1]New York University, [2]Amazon Alexa AI, [3]University of California, Los Angeles

vishakh@nyu.edu
{behnam, djinamzn, patlange, seokhwk, yangliud}@amazon.com
violetpeng@cs.ucla.edu dilek@ieee.org

## Abstract

The bulk of work adapting transformer models to open-domain dialogue represents dialogue context as the concatenated set of turns in natural language. However, it is unclear if this is the best approach. In this work, we investigate this question by means of an empirical controlled experiment varying the dialogue context format from text-only formats (all recent utterances, summaries, selected utterances) as well as variants that are more structurally different (triples, AMR). We compare these formats based on fine-tuned model performance on two downstream tasks—knowledge selection and response generation. We find that simply concatenating the utterances works as a strong baseline in most cases, but is outperformed in longer contexts by a hybrid approach of combining a summary of the context with recent utterances. Through empirical analysis, our work highlights the need to examine the format of context representation and offers recommendations on adapting general-purpose language models to dialogue tasks.

## 1 Introduction

The bulk of existing work in adapting transformer models to open-domain dialogue represents the dialogue context as the concatenated set of turns in natural language (Zhang et al., 2019b; Roller et al., 2021; Shuster et al., 2022). While the self-attention mechanisms of these models are able to capture the context from these flat representations, it remains unclear if this is the best approach (Li et al., 2021). Studying the format of context representation would help improve performance on downstream tasks such as response generation and external knowledge selection and could also potentially inform the pretraining of general-purpose dialogue models. Additionally, as the length of conversations increases (Gopalakrishnan et al., 2019; Xu, 2021), these are truncated based on the limit imposed by the positional encodings on transformers. We also know that not all of the utterances are equally relevant so succinctly representing the relevant information in the context given the current conversation state and filtering out the noise from prior interactions would help to model provide more coherent responses.

In this work, we empirically investigate the dialogue context representation in the text space for using sequence-to-sequence models. To prioritize broad coverage, we vary the the format of the context using both natural language-only formats (e.g., using all recent utterances or summaries) as well as formats that are more structurally different (e.g., extracting knowledge triples from the utterances) (Section 2) and compare these based on downstream task performance.

We find that concatenating all recent utterances is a strong baseline. However, in longer dialogues, combining recent utterances with a summary of the past context obtains the best performance. This shows the benefit of the complementary long and short view of dialogue context. We also observe that improving summary quality and introducing external elements about the coherence of the context result in a further gain of downstream performance. This study and related findings can be extended to combine with elements from the broader definition of context (Bunt, 1999), such as social cues and guidelines (Gupta et al., 2022b), which were previously not included in dialogue datasets.

## 2 Approach

We study the effect of the representation of dialogue context on downstream dialogue tasks—knowledge selection and response generation. In order to do so, we run a controlled experiment fine-tuning sequence-to-sequence models on the two tasks verbalised into the text-to-text setup, while varying only the format in which the dialogue con-

* Work done during summer internship at Amazon Alexa AI.



*Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 538–547
September 11–15, 2023. 

text is represented.

The first broad category of representations consists of directly using the dialogue utterances. We include the concatenated past dialogue utterances, truncated when necessary, as ***Plaintext*** representation. This includes all the past turns delineated using a special token when applicable. We also include ***Windows*** of recent turns where we only use the most recent $n$ utterances as the context.

To test if models require only the knowledge items within the dialogue utterances, we extract (subject, object, relation) ***Triples*** from the utterances as the context. To see if models benefit from more structured information, we convert the utterances into ***AMR*** graphs (Banarescu et al., 2013).

Finally, we examine if the information from the context can be distilled using summarization (Feng et al., 2021; Gliwa et al., 2019a; Khalifa et al., 2021). One method is to convert the utterances from both speakers into an abstractive ***Summary*** using a separate summarization model.[1] And while a summary might contain all the required high-level information from the dialogue context, it loses the local discourse-level information from recent utterances. To mitigate this, we create a hybrid ***Summary + Utterances*** format by appending the *Summary* with *Windows of Turns*. We also include an extractive summary in the form of ***Selected Turns*** from the context using pointwise mutual information, a proxy for relevance, with respect to the most recent turn (Padmakumar and He, 2021).

We provide further implementation details about each of the methods in Appendix C and illustrate an example converted to each of them in Figure 1.

## 3 Experiments

### 3.1 Datasets and Metrics

**Knowledge Selection** To evaluate performance on knowledge selection, we report results on the Wizard of Wikipedia (WoW) (Dinan et al., 2018) dataset, which consists of dialogue between a wizard (expert) and apprentice (novice) where the wizard selects knowledge items (sentences) to form a response. In the sequence-to-sequence setup, we frame this as a classification task on individual knowledge items as follows.

**Input:** <context> </s> <knowledge item>
**Output**: `"Relevant"` for the gold knowledge item given that context, and `"Not Relevant'` otherwise.

In addition to all the context formats from Section 2, we include another baseline called ***Plaintext with Documents*** where the gold documents that were used to generate previous wizard turns were appended to the utterances in the dialogue context.

**Metrics**: We report accuracy/F1-score of each label in lieu of instance-based classification performance. To report retrieval performance, we score the individual knowledge items for a particular context using the token probabilities assigned to `"Relevant"` and select the most relevant item. We then evaluate if this matches the checked sentence from the dataset, akin to Recall@1 when this is framed as a retrieval problem. We also report a more relaxed metric that evaluates if this item is from the checked document from the dataset.

**Response Generation** We report results on WoW, Multi-Session Chat (MSC) (Xu et al., 2021) and Topical Chat (TCS) (Gopalakrishnan et al., 2019) where the objective is to generate the gold response given the context. For WoW, the task is a knowledge-grounded dialogue where the responses were formed using the gold knowledge item from the dataset. The task for TCS is also knowledge-grounded response generation, but not all turns are accompanied by relevant knowledge items. For MSC, the task is for the partners to converse about their own interests and discuss information about each others' interests across multiple sessions. We concatenate utterances from all past sessions with a special token indicating a session break.[2]
**Input:** <context> </s> <optional knowledge item>
**Output**: Gold response from the dataset.

**Metrics** We report perplexity of the gold utterances w.r.t. the finetuned models and the BertScore (Zhang et al., 2019a) between the generated response and the target utterance.

### 3.2 Model Training

For each of the datasets, we convert all of the train examples into the different context representations from Section 2 and report finetuned T5 (Raffel et al., 2020) performance. We use the T5-base ($220M$ parameters) and Large ($770M$ parameters) variants. While the models trained in Zhang et al. (2019b); Peng et al. (2022) have examined further

[1] In particular, we use a BART-large model finetuned on SAMSum(Gliwa et al., 2019b).

[2] For MSC, the *Summary* baseline(s) use the released summaries for past sessions coupled with a model generated summary for the utterances in the current session.

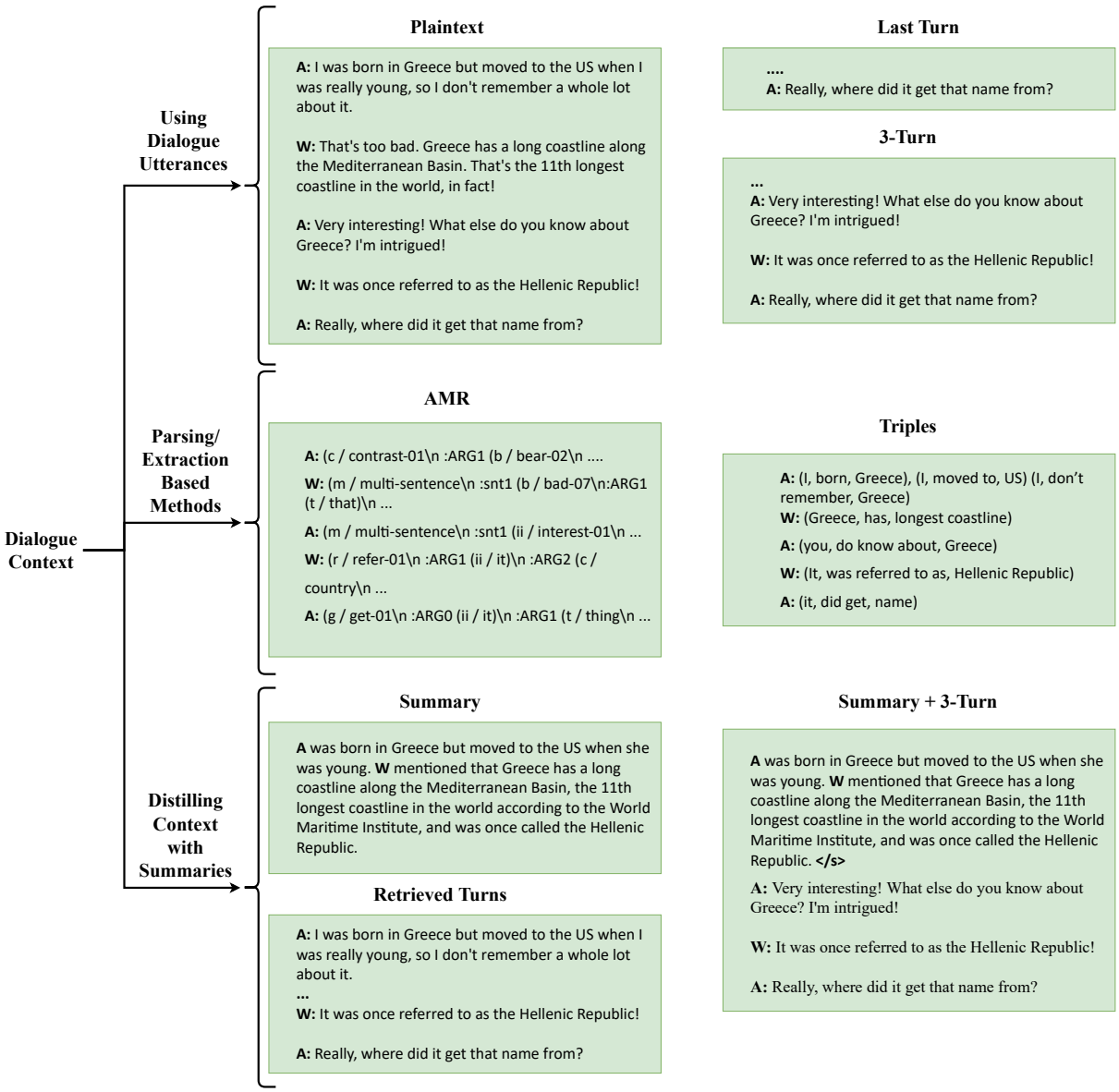


Figure 1: Example illustrating the conversion a dialogue into all the context representation methods evaluated in our experiments. The original set of utterances indicated by *Plaintext*. We perform an empirical controlled experiment evaluating the fine-tuned dialogue model performance on each of these context representation formats.

pretraining on dialogue, this would bias the model to additionally favor the *Plaintext* baseline. As a result, we choose T5, noting that absolute performance might improve further by adding dialogue-specific pertaining. When tokenizing the context, we allow for up to 1024 tokens and truncate earlier utterances in case of an overflow. We optimize cross-entropy loss on the output tokens in the desired format based on the dataset. We run finetuning for 10 epochs with an early stopping criteria based on validation loss. For each context representation, we select the best learning rate sweeping from $1e^{-3}$ to $1e^{-6}$. In the text-to-text setup, we run inference with greedy decoding kept uniform across the representations. Our experiments were run on a `p3.8xlarge` and a `p3.16xlarge` EC2 instances containing 4 and 8 Tesla V100 GPUs respectively.

## 4 Results

Table 1 and Table 2 show the results comparing context representation formats on knowledge selection and response generation respectively.

***Plaintext* is a strong baseline, which is outperformed by *Summaries+Utterances* on longer dialogues** From Table 1 and Table 2, we see that the *Plaintext* representation provides a strong baseline

| | | Plaintext | Plaintext w Docs | Last Turn | 3-Turn | Selected Turns | AMR | Triples | Summary | Summ + 1 Turn | Summ + 3 Turn | Summ + 5 Turn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Accuracy** | **Overall** | 0.959 / 0.963 | 0.958 / 0.960 | 0.960 / 0.962 | 0.960 / 0.963 | **0.965 / 0.966** | 0.961 / 0.965 | 0.961 / 0.963 | 0.963 / 0.964 | 0.958 / 0.960 | 0.954 / 0.958 | 0.957 / 0.961 |
| | **Relevant** | 0.331 / 0.265 | 0.355 / 0.289 | 0.278 / 0.234 | 0.307 / 0.261 | 0.282 / 0.244 | 0.265 / 0.264 | 0.268 / 0.263 | 0.286 / 0.231 | 0.301 / 0.253 | **0.369 / 0.297** | 0.353 / 0.281 |
| **F1 Scores** | **Relevant** | 0.196 / 0.170 | **0.202 / 0.188** | 0.169 / 0.150 | 0.184 / 0.165 | 0.192 / 0.172 | 0.160 / 0.158 | 0.166 / 0.163 | 0.174 / 0.155 | 0.183 / 0.159 | 0.191 / 0.167 | 0.194 / 0.170 |
| **Recall@1 of Most Relevant Item** | **Checked Sentence** | 0.159 / 0.116 | **0.171 / 0.129** | 0.114 / 0.111 | 0.120 /0.105 | 0.138 / 0.118 | 0.097 / 0.085 | 0.101 / 0.086 | 0.116 / 0.099 | 0.128 / 0.111 | 0.143 / 0.118 | 0.147 / 0.116 |
| | **Checked Passage** | 0.238 / 0.174 | **0.265 / 0.201** | 0.186 / 0.165 | 0.165 / 0.146 | 0.214 / 0.178 | 0.138 / 0.124 | 0.140 / 0.126 | 0.160 / 0.150 | 0.199 / 0.177 | 0.234 / 0.191 | 0.222 . 0.185 |

Table 1: Evaluation of context representation methods on WoW knowledge selection. Each cell has two numbers corresponding to results on the random split (left) and topic split (right) of the validation set. All metrics are rounded off to three decimal places and the highest in each row is bold. We include only the overall accuracy and classification metrics of the `Relevant` label here. For metrics on all labels see Table 7 in Appendix E.

| | | Plaintext | Last Turn | 3 Turn | 5 Turn | Selected Turns | AMR | Triples | Summary | Summ + 1 Turn | Summ + 3 Turns | Summ + 5 Turns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **WoW** | **Bertscore** | **0.905 / 0.904** | 0.903 / 0.901 | 0.903 / 0.902 | 0.904 / 0.902 | 0.903 / 0.900 | 0.895 / 0.890 | 0.898 / 0.894 | 0.902 / 0.900 | 0.903 / 0.901 | 0.905 / 0.903 | 0.904 / 0.903 |
| | **Perplexity** | **6.978 / 7.545** | 7.446 / 8.084 | 7.398 / 8.011 | 7.304 / 7.885 | 7.177 / 7.783 | 7.987 / 8.623 | 7.803 / 8.510 | 7.477 / 8.115 | 7.261 / 7.836 | 7.050 / 7.660 | 7.028 / 7.601 |
| **MSC** | **Bertscore** | **0.873** | 0.861 | 0.864 | 0.872 | 0.865 | 0.854 | 0.858 | 0.866 | 0.869 | 0.871 | **0.873** |
| | **Perplexity** | 12.246 | 15.262 | 14.701 | 14.024 | 14.565 | 16.245 | 15.782 | 13.985 | 13.69 | 13.011 | **12.205** |
| **TCS** | **Bertscore** | 0.871 / 0.868 | 0.869 / 0.867 | 0.870 / 0.869 | **0.871 / 0.869** | 0.869 / 0.869 | 0.865 / 0.864 | 0.866 / 0.865 | 0.868 / 0.866 | 0.869 / 0.867 | 0.870 / 0.868 | 0.871 / 0.868 |
| | **Perplexity** | 12.313 / 14.443 | 13.293 / 15.950 | 13.045 / 15.650 | 12.847 / 15.023 | 12.778 / 15.237 | 13.587 / 16.290 | 13.402 / 16.117 | 12.899 / 15.262 | 12.686 / 15.013 | 12.538 / 14.812 | **12.181 / 14.342** |

Table 2: Evaluation of context representation methods on response generation. For WoW, each cell has two numbers corresponding to results on the random split and topic split of the validation set. For MSC, we report results on all the turns of the validation set. For TCS, the two numbers correspond to the frequent and rare splits respectively. All metrics are rounded off to three decimal places and the highest in each row is bold.

for both knowledge selection and response generation. When we examine the *Last Turn* and *3-Turn* columns, we see the trend that increasing the window size predictably improves performance, but these lag behind *Plaintext*. This shows that transformers are able to leverage the additional information from more recent utterances in the context. However, we see that *Plaintext* is outperformed by the *Summary + 5-turn* method on the longer dialogue datasets, MSC and TCS. This shows that past the limit imposed on current transformer encoders by the positional embeddings, summarizing all available information outperforms a truncated set of recent utterances. Finally, we see that *Summary + 5-turn* outperforms *Summary* alone on all the datasets. These findings highlight the complementary *Long* and *Short* views of dialogue context from summaries and recent utterances respectively.

**Improving the quality of summaries results in better downstream performance** To observe the effect of summary quality, we point out two comparisons. On MSC, we compare the response generation performance using both the gold human-written summaries and model-generated summaries (released with the dataset). The perplexity for response generation reduces by using higher quality, human-written summaries (Table 5). Secondly, we can view the *Selected Turns* baseline as an extractive summary of the dialogue context that consistently outperforms windows of text of the same number of turns (here *Selected Turns* and *3-turn* are comparable). Combined with the observation of the complementary nature of summaries and recent turns, a future direction highlighted through our work is to use downstream task performance as a means to evaluate dialog summarization.

**Natural language-based approaches outperform the more structure-oriented variants** We observe that *AMR* and *Triples* are consistently outperformed by all the other utterance-based and summary-based variants. This is potentially explained by the higher similarity of the natural language formats to the pretraining data of sequence-to-sequence models.[3]

**Positive Scaling Trends** One of the main advantage of using sequence-to-sequence transformers is that as pretrained models get better, we can expect improved performance in downstream tasks. We observe a simple version of this when comparing results on the different context representation methods with T5-base and T5-large in Table 3 and

[3]These methods are at a disadvantage in the text-to-text format and could be improved by different methods of encoding the extracted information.

| | | Plaintext | | Last Turn | | Retrieved Turns | | Summ + 3 | | Summ + 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Base | Large | Base | Large | Base | Large | Base | Large | Base | Large |
| F1 Scores | Relevant | 0.196 / 0.170 | 0.187 / 0.170 | 0.169 / 0.150 | 0.177 / 0.159 | 0.192 / 0.172 | 0.210 / 0.185 | 0.191 / 0.167 | **0.212 / 0.189** | 0.194 / 0.170 | 0.205 / 0.181 |
| Recall@1 of Most Relevant Item | Match to 'Checked Sentence' | 0.159 / 0.116 | **0.203 / 0.156** | 0.114 / 0.111 | 0.131 / 0.110 | 0.138 / 0.118 | 0.163 / 0.135 | 0.143 / 0.118 | 0.161 / 0.135 | 0.147 / 0.116 | 0.160 / 0.131 |
| | Match to 'Checked Passage' | 0.238 / 0.174 | **0.326 / 0.258** | 0.186 / 0.165 | 0.191 / 0.162 | 0.214 / 0.178 | 0.285 / 0.231 | 0.234 / 0.191 | 0.255 / 0.219 | 0.222 . 0.185 | 0.252 / 0.199 |

Table 3: Evaluation of knowledge selection as a function of model size—T5-Base vs Large for 5 different context representations. We largely observe positive scaling trends on both retrieval metrics and classification F1-scores. Table 9 in Appendix E shows the same table with metrics for all labels.

| | | Plaintext | | Retrieved Turns | | Last Turn | | Summ + 5 Turns | |
|---|---|---|---|---|---|---|---|---|---|
| | | Base | Large | Base | Large | Base | Large | Base | Large |
| WoW | Perplexity | 6.978 / 7.545 | **5.989 / 6.371** | 7.177 / 7.783 | 6.151 / 6.574 | 7.446 / 8.084 | 6.754 / 7.226 | 7.028 / 7.601 | 6.001 / 6.412 |
| TCS | Perplexity | 12.313 / 14.443 | 9.811 / 11.279 | 12.778 / 15.237 | 10.101 / 12.980 | 13.293 / 15.950 | 11.456 / 14.374 | 12.181 / 14.342 | **9.792 / 11.113** |

Table 4: Evaluation of response generation as a function of model size—T5-Base vs Large for 4 different context representations. We observe positive scaling trends across each of the representations

Table 4. Performance improves using the scaled up model uniformly for response generation and on retrieval metrics in knowledge selection.

**Providing additional content as part of the context improves performance** Augmenting the *Plaintext* baseline with document level information for WoW results in further improvement in both classification and retrieval scores. In this work, we only considered the utterances in the dialogue itself to be a part of the context. However a broader definition of context for dialogue includes not just the turns but also discourse information, social context, or the relationship between the speakers, and even physical context, or cues from the relative physical positions and actions of the speakers (Bunt, 1999). Our work indicates that a promising future direction of dialogue research could involve collecting and summarizing all this additional rich information to be used by dialogue models.

We present additional results in Appendix E and discuss some limitations that inform future directions in Appendix A.

## 5 Related Work

When adapting transformers to dialogue tasks, the most common approach is to simply concatenate dialogue utterances (Zhang et al., 2019b; Adiwardana et al., 2020; Roller et al., 2021; Bao et al., 2021; Gupta et al., 2022a; Shuster et al., 2022). For longer dialog datasets where the entire conversation cannot be encoded, summaries of past sessions are a helpful way to provide all the relevant information needed to continue the conversation (Xu et al., 2022). While *AMR* graphs have been used to perturb individual utterances in order to evaluate coherence in dialogue (Ghazarian et al., 2022), to the best of our knowledge, *AMR* and *Knowledge Triples* have not been used to represent the context. We include them for wider coverage. In the dialogue space, retrieval has largely been used to identify relevant knowledge items to be included for response generation (Shuster et al., 2021). Prior work has examined matching candidate responses with multiple utterances for selection, the weighting learned in effect attending to 'relevant' turns (Wu et al., 2016; Zhang et al., 2018), however, we explicitly select turns as a means of representing the dialogue context across both of our open domain dialogue tasks. To our knowledge, ours is the first controlled experiment to evaluate different textual context representation methods for sequence-to-sequence models.

## 6 Conclusion

In this work, we present an empirical controlled study examining dialogue context representation for transformer models on open-domain dialogue tasks. While concatenating all previous turns, as is often adopted, is a strong baseline, combining summaries of the overarching context with recent utterances yields the best results in longer dialogues. Additionally improving the quality of the summaries being used and introducing further background information into the context further improve performance. This provides us with new directions to work on including dialogue summarization and considering the broader definition of context for use in open-domain dialogue.

## References


Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards building an open-domain chatbot via curriculum learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.

Harry Bunt. 1999. Context representation for dialogue management. In *Proceedings of the Second International and Interdisciplinary Conference on Modeling and Using Context*, CONTEXT '99, page 77–90, Berlin, Heidelberg. Springer-Verlag.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019a. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019b. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P Bigham. 2022a. Improving zero and few-shot generalization in dialogue through instruction tuning. *arXiv preprint arXiv:2205.12673*.

Prakhar Gupta, Yang Liu, Di Jin, Behnam Hedayatnia, Spandana Gella, Sijia Liu, Patrick Lange, Julia Hirschberg, and Dilek Hakkani-Tur. 2022b. Dialguide: Aligning dialogue model behavior with developer guidelines. *arXiv preprint arXiv:2212.10557*.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, Online. Association for Computational Linguistics.

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

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*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian McAuley, and Furu Wei. 2021. Beyond preserved accuracy: Evaluating loyalty and robustness of BERT compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10653–10659, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Yang Xu. 2021. Global divergence and local convergence of utterance semantic representations in dialogue. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 116–124, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

# A Limitations

**Coverage of Context Representations** We acknowledge that the list of context representation formats we examine is non-exhaustive and each particular context format could be further optimized. For instance, for *AMR* we only cover the semantic representations within a single utterance. There are other types of structural aspects in dialogues like discourses, turn-taking, and so on which could be incorporated. We report results comparing these in order to inform subsequent model training/pre-training as well as subsequent analysis of a similar nature.

**Using Only Verbalized Representations** In this work, we only cover context representation formats that are verbalized in natural language. It is unclear if encoding the information either into a specialized dialogue transformer architecture or as a graph would result in improved performance. We choose the verbalized format as it is the most general purpose which can be used to adapt many different language models (Liang et al., 2022).

**Adapting Retrieval Tasks for Text-to-Text Models** Adapting language models to retrieval tasks such as knowledge selection can be done either by running inference on individual examples or by combining all candidates along with the input context. Liang et al. (2022) perform a comparison of these variants in the few-shot setting for a large number of tasks and observe no clear winning format so we proceed with separate inference on knowledge items. Here, to isolate the role of the context representation, we fix the format of the task and study the effect of dialogue context on performance

**Evaluation** We acknowledge that we report performance using automatic metrics on a single run for both sets of tasks and human evaluation would allow for a more holistic understanding of the capabilities of models, particularly on response generation. Human evaluation and running multiple sets of fine-tuning runs for each of the different formats would be expensive. In this work, we restricted ourselves to the same in order to focus on comparing and identifying trends in performance between a wider range of different context representation formats.

## B Potential Risks

Our work discusses ways to adapt sequence-to-sequence transformer models for open-domain dialogue. The main associated risk comes from the black-box nature of these models. The text that is generated is pretty heavily influenced by the pretraining data. The models fine-tuned in this paper are open-sourced T5 checkpoints which may contain biases from the C4 (Raffel et al., 2020) corpus. Additionally, the advent of closed-access and limited-access language models such as GPT3 and Anthropic-LM comes with more uncertainty as the pretraining and training processes of these models are not as well documented (Liang et al., 2022).

## C Context Representations

**Context Representation Formats** We vary the context in the following ways in an attempt to ensure coverage of different formats. The first broad category of representations consists of directly using dialogue utterances.

- **Plaintext**: The simplest, and most widely used, manner in which we can represent the dialogue context is just the concatenated set of past dialogue utterances. This includes all turns from past sessions delineated using a special token when applicable. [4]

- **Windows of Turns**: Here we only use the most recent $n$ utterances as the context. As we increase $n$, we provide more local context about the dialogue.

Aside from including the utterances themselves, to evaluate if models benefit from more structured information we include the following representations:

- **AMR**: We convert each utterance into an AMR graph (Banarescu et al., 2013) and use the verbalised form as the context. The AMR parses the text into a directed acyclic graph, explicitly conveying the relationships as edges between the various concept nodes in the text. We use the `model_parse_xfm_bart_large` model from amrlib to convert the utterances into the corresponding AMR. We acknowledge that performance in our experiments could be affected by the quality of AMR conversions. We refer readers to the original library for performance benchmarking of the text-to-AMR model.

- **Knowledge Triples**: To test if models require only the knowledge items within the dialogue context, and not the whole utterance, we extract `(subject, object, relation)` triples from the utterances as the context. We use OpenIE5 to extract triples and use a simple unigram overlap heuristic to filter out duplicates. If two triples have a unigram overlap of over 0.7, only one is selected.

Finally, we examine if the information from the dialogue context can be distilled while retaining the natural language format using summarization.

- **Summary**: We summarise all of the dialogue utterances from both speakers abstractively using a finetuned transformer model. In particular, we use a BART-large model finetuned on SAMSum(Gliwa et al., 2019b). As indicated in Section 4, performance depends on the quality of the summarization model. This model was not trained by the authors of this work. We refer readers to the model card on HuggingFace for evaluation of the model itself.

- **Summary + Utterances**: While a summary might contain all the high-level information from the dialogue context, it loses the local discourse-level information from recent utterances which provide cues on how to use the high-level information. We create this hybrid short+long form context representation by appending the *Summary* with *Windows of Turns*.

- **Retrieved Turns**: While the aforementioned setups contain abstractive summaries of the dialogue context, we also include an extractive summary generated by selecting relevant turns using pointwise mutual information to the most recent turn (Padmakumar and He, 2021). In order to select relevant turns, we calculate the PMI of all utterances with respect to the *Last Turn* and combine the 2 most relevant turns, in order to obtain an extractive summary of the context.

An example converted to each of the above formats is provided in Figure 1.

[4] Dataset specific details are provided in Section 3.1

## D Details for Responsibility Checklist

### D.1 License and Usage of Scientific Artifacts

The Wizard of Wikipedia (Dinan et al., 2018) and MSC (Xu et al., 2021) datasets made available through ParlAI that is shared under the MIT License which permits usage of the data for research such as our work. Topical Chat (Gopalakrishnan et al., 2019) is shared using the Community Data License Agreement - Sharing, Version 1.0 which also permits the usage of the data in this manner. These datasets are commonly used in the community and are collected while ensuring that it was properly anonymized and does not contain any offensive language. We do not perform additional checks for either of the same. T5 (Raffel et al., 2020), used for all our finetuning experiments, is released under the Apache 2.0 license which permits its use for research. The model used for dialogue summarization and `amrlib` are both shared under the MIT license which permits such usage as does OpenIE which is shared under the Open IE 5 Software License Agreement. All of the artifacts, both models and datasets, were used as intended by the original authors.

### D.2 Coverage and Statistics of the Data

All of the datasets contain only English data, largely collected from American English speakers conversing in a one-on-one conversation. The specifics of the settings where the conversations are collected are well documented and can be referred to in the original works (Dinan et al., 2018; Xu et al., 2021; Gopalakrishnan et al., 2019). Wizard of Wikipedia consists of $18,430$ documents ($166,787$ utterances total, $74,092$ of which were wizard turns used in knowledge selection) in the train set. The results were reported on the random split ($981$ documents, $3,939$ wizard turns) and topic split ($967$ documents, $3,927$ wizard turns) of the validation data. For MSC, there are $4000$ train conversations (spread across multiple sessions) with $161,440$ turns and we report results on the validation set ($1001$ conversations, $53,332$ turns). In Topical Chat, there are $8628$ train conversations consisting of $188378$ utterances and we report results on the frequent ($539$ conversations, $11681$ turns) and rare ($539$ conversations, $11692$ turns) splits of the validation data.

| | | Human Written Summary | Model Generated Summary |
|---|---|---|---|
| All Turns | Perplexity | **12.129** | 12.205 |
| First Response in Session | Perplexity | **10.199** | 10.257 |

Table 5: Performance on MSC improves when using the gold, human-written summaries as opposed to model-generated summaries.

| | Truncated Examples | Examples wo Truncation |
|---|---|---|
| Perplexity | 14.381 | 12.564 |
| Bertscore | 0.8641 | 0.8722 |

Table 6: Response generation performance on MSC examples adapted into the *Plaintext* representation and divided based on whether these are truncated.

## E Additional Results

We report a more comprehensive version of the knowledge selection results from Table 1 in Table 7 and response generation from Table 2 in Table 8.

**Effect of Scaling Model Size** Table 9 and Table 10 contain the full comparison of results when we switch from T5-Base to T5-Large.

**Quality of Summaries** In order to ablate the quality of summaries used, we compared response generation performance on the MSC dataset, comparing the *Summary + 5-Turn* baseline when the gold, human-written summaries are used as opposed to the model generated summaries released in the original dataset. From Table 5 we observe that the higher quality summaries result in further improvement in performance.

**Effect Of Truncation** Here we aim to empirically verify that truncation of context has an adverse effect on model performance. We select those examples in the second session of the MSC dataset when adapted using the *Plaintext* representation and divide these into whether or not the context was truncated. This particular set of examples was chosen because, out of all the sessions, this was the one which had a relatively large fraction of examples in both of these buckets—$27.6\%$ of examples were truncated. From Table 6 we clearly see that those examples which suffer from truncation have a drop in performance.

| | | Plaintext | Plaintext w Docs | Last Turn | 3-Turn | 5-Turn | Selected Turns | AMR | Triples | Summary | Summ + 3 Turn | Summ + 5 Turn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Item Classification Accuracy** | **Overall** | 0.959 / 0.963 | 0.958 / 0.960 | 0.960 / 0.962 | 0.960 / 0.963 | 0.960 / 0.962 | **0.965 / 0.966** | 0.961 / 0.965 | 0.961 / 0.963 | 0.963 / 0.964 | 0.954 / 0.958 | 0.957 / 0.961 |
| | **NR** | 0.969 / 0.973 | 0.965 / 0.966 | 0.970 / 0.972 | 0.970 / 0.972 | 0.969 / 0.972 | **0.975 / 0.976** | 0.973 / **0.976** | 0.970 / 0.972 | 0.974 / 0.975 | 0.963 / 0.967 | 0.966 / 0.971 |
| | **R** | 0.331 / 0.265 | 0.355 / 0.289 | 0.278 / 0.234 | 0.307 / 0.261 | 0.318 / 0.277 | 0.282 / 0.244 | 0.265 / 0.264 | 0.268 / 0.263 | 0.286 / 0.231 | **0.369 / 0.297** | 0.353 / 0.281 |
| **Item Classification F1 Scores** | **NR** | 0.979 / 0.981 | 0.977 / 0.979 | 0.979 / 0.981 | 0.980 / 0.981 | 0.979 / 0.980 | **0.982 / 0.982** | 0.977 / 0.980 | 0.978 / 0.980 | 0.981 / **0.982** | 0.977 / 0.978 | 0.978 / 0.980 |
| | **R** | 0.196 / 0.170 | **0.202 / 0.188** | 0.169 / 0.150 | 0.184 / 0.165 | 0.187 / 0.171 | 0.192 / 0.172 | 0.160 / 0.158 | 0.166 / 0.163 | 0.174 / 0.155 | 0.191 / 0.167 | 0.194 / 0.170 |
| **Recall@1 of Most Relevant Item** | **Match to 'Checked Sentence'** | 0.159 / 0.116 | **0.171 / 0.129** | 0.114 / 0.111 | 0.120 /0.105 | 0.127 / 0.106 | 0.138 / 0.118 | 0.097 / 0.085 | 0.101 / 0.086 | 0.116 / 0.099 | 0.143 / 0.118 | 0.147 / 0.116 |
| | **Match to 'Checked Passage'** | 0.238 / 0.174 | **0.265 / 0.201** | 0.186 / 0.165 | 0.165 / 0.146 | 0.179 / 0.153 | 0.214 / 0.178 | 0.138 / 0.124 | 0.140 / 0.126 | 0.160 / 0.150 | 0.234 / 0.191 | 0.222 . 0.185 |

Table 7: Evaluation of context representation methods on knowledge selection. Each cell has two numbers corresponding to results on the random split and topic split of the validation set. All metrics are rounded off to three decimal places and the highest in each row is bold.

| | | | Plaintext | Last Turn | 3 Turn | 5 Turn | Selected Turns | AMR | Triples | Summary | Summ + 1 Turn | Summ + 3 Turns | Summ + 5 Turns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **WoW** | | **Bertscore** | **0.905 / 0.904** | 0.903 / 0.901 | 0.903 / 0.902 | 0.904 / 0.902 | 0.903 / 0.900 | 0.895 / 0.890 | 0.898 / 0.894 | 0.902 / 0.900 | 0.903 / 0.901 | 0.905 / 0.903 | 0.904 / 0.903 |
| | | **Perplexity** | **6.978 / 7.545** | 7.446 / 8.084 | 7.398 / 8.011 | 7.304 / 7.885 | 7.177 / 7.783 | 7.987 / 8.623 | 7.803 / 8.510 | 7.477 / 8.115 | 7.261 / 7.836 | 7.050 / 7.660 | 7.028 / 7.601 |
| **MSC** | ***All*** | **Bertscore** | **0.873** | 0.861 | 0.864 | 0.872 | 0.865 | 0.854 | 0.858 | 0.866 | 0.869 | 0.871 | **0.873** |
| | | **Perplexity** | 12.246 | 15.262 | 14.701 | 14.024 | 14.565 | 16.245 | 15.782 | 13.985 | 13.69 | 13.011 | **12.205** |
| | ***1st*** | **Bertscore** | 0.875 | 0.868 | 0.862 | 0.863 | 0.863 | 0.859 | 0.863 | **0.876** | 0.873 | 0.874 | 0.875 |
| | | **Perplexity** | 10.386 | 15.627 | 15.118 | 13.998 | 14.409 | 16.109 | 15.704 | **10.143** | 10.988 | 10.876 | 10.257 |
| **TCS** | | **Bertscore** | 0.871 / 0.868 | 0.869 / 0.867 | 0.870 / 0.869 | **0.871 / 0.869** | 0.869 / 0.869 | 0.865 / 0.864 | 0.866 / 0.865 | 0.868 / 0.866 | 0.869 / 0.867 | 0.870 / 0.868 | 0.871 / 0.868 |
| | | **Perplexity** | 12.313 / 14.443 | 13.293 / 15.950 | 13.045 / 15.650 | 12.847 / 15.023 | 12.778 / 15.237 | 13.587 / 16.290 | 13.402 / 16.117 | 12.899 / 15.262 | 12.686 / 15.013 | 12.538 / 14.812 | **12.181 / 14.342** |

Table 8: Evaluation of context representation methods on response generation. For WoW, each cell has two numbers corresponding to results on the random split and topic split of the validation set. For MSC, we report results on all the turns (*All*), and for the first turn in each session (*1st*). For TCS, the two numbers correspond to the frequent and rare splits respectively. All metrics are rounded off to three decimal places and the highest in each row is bold.

| | | Plaintext | | Last Turn | | Retrieved Turns | | Summ + 3 | | Summ + 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Base** | **Large** | **Base** | **Large** | **Base** | **Large** | **Base** | **Large** | **Base** | **Large** |
| **Item Classification Accuracy** | **Overall** | 0.959 / 0.963 | 0.944 / 0.950 | 0.960 / 0.962 | **0.967 / 0.968** | 0.965 / 0.966 | 0.964 / 0.965 | 0.954 / 0.958 | 0.964 / 0.965 | 0.957 / 0.961 | 0.957 / 0.961 |
| | **NR** | 0.969 / 0.973 | 0.951 / 0.954 | 0.970 / 0.972 | **0.977 / 0.979** | 0.975 / 0.976 | 0.973 / 0.976 | 0.963 / 0.967 | 0.973 / 0.975 | 0.966 / 0.971 | 0.965 / 0.970 |
| | **R** | 0.331 / 0.265 | **0.445 / 0.402** | 0.278 / 0.234 | 0.245 / 0.212 | 0.282 / 0.244 | 0.329 / 0.275 | 0.369 / 0.297 | 0.333 / 0.255 | 0.353 / 0.281 | 0.382 / 0.305 |
| **Item Classification F1 Scores** | **NR** | 0.979 / 0.981 | 0.971 / 0.968 | 0.979 / 0.981 | **0.983 / 0.984** | 0.982 / 0.982 | 0.981 / 0.982 | 0.977 / 0.978 | 0.981 / 0.982 | 0.978 / 0.980 | 0.978 / 0.978 |
| | **R** | 0.196 / 0.170 | 0.187 / 0.170 | 0.169 / 0.150 | 0.177 / 0.159 | 0.192 / 0.172 | 0.210 / 0.185 | 0.191 / 0.167 | **0.212 / 0.189** | 0.194 / 0.170 | 0.205 / 0.181 |
| **Recall@1 of Most Relevant Item** | **Match to 'Checked Sentence'** | 0.159 / 0.116 | **0.203 / 0.156** | 0.114 / 0.111 | 0.131 / 0.110 | 0.138 / 0.118 | 0.163 / 0.135 | 0.143 / 0.118 | 0.161 / 0.135 | 0.147 / 0.116 | 0.160 / 0.131 |
| | **Match to 'Checked Passage'** | 0.238 / 0.174 | **0.326 / 0.258** | 0.186 / 0.165 | 0.191 / 0.162 | 0.214 / 0.178 | 0.285 / 0.231 | 0.234 / 0.191 | 0.255 / 0.219 | 0.222 . 0.185 | 0.252 / 0.199 |

Table 9: Evaluation of knowledge selection as a function of model size. We report performance on T5-Base and Large for 5 different context representations. We observe positive scaling trends, where the larger model performs better, uniformly for retrieval metrics and generally across the classification metrics for the *Relevant* label.

| | | Plaintext | | Retrieved Turns | | Last Turn | | Summ + 5 Turns | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Base** | **Large** | **Base** | **Large** | **Base** | **Large** | **Base** | **Large** |
| **WoW** | **Bertscore** | 0.905 / 0.904 | **0.907 / 0.906** | 0.903 / 0.900 | 0.904 / 0.902 | 0.903 / 0.901 | 0.904 / 0.902 | 0.904 / 0.903 | 0.906 / 0.904 |
| | **Perplexity** | 6.978 / 7.545 | **5.989 / 6.371** | 7.177 / 7.783 | 6.151 / 6.574 | 7.446 / 8.084 | 6.754 / 7.226 | 7.028 / 7.601 | 6.001 / 6.412 |
| **TCS** | **Bertscore** | 0.871 / 0.869 | 0.873 / 0.872 | 0.869 / 0.869 | 0.872 / 0.871 | 0.869 / 0.867 | 0.871 / 0.870 | 0.871 / 0.868 | **0.874 / 0.873** |
| | **Perplexity** | 12.313 / 14.443 | 9.811 / 11.279 | 12.778 / 15.237 | 10.101 / 12.980 | 13.293 / 15.950 | 11.456 / 14.374 | 12.181 / 14.342 | **9.792 / 11.113** |

Table 10: Evaluation of response generation as a function of model size. We report performance on T5-Base and Large for 4 different context representations. We observe positive scaling trends, where the larger model performs better particularly on perplexity scores.