Response Generation in Longitudinal Dialogues: Which Knowledge Representation Helps?

Seyed Mahed Mousavi, Simone Caldarella, Giuseppe Riccardi
Signals and Interactive Systems Lab, University of Trento, Italy
mahed.mousavi@unitn.it,giuseppe.riccardi@unitn.it

Abstract

Longitudinal Dialogues (LD) are the most challenging type of conversation for human-machine dialogue systems. LDs include the recollections of events, personal thoughts, and emotions specific to each individual in a sparse sequence of dialogue sessions. Dialogue systems designed for LDs should uniquely interact with the users over multiple sessions and long periods of time (e.g. weeks), and engage them in personal dialogues to elaborate on their feelings, thoughts, and real-life events. In this paper, we study the task of response generation in LDs. We evaluate whether general-purpose Pre-trained Language Models (PLM) are appropriate for this purpose. We fine-tune two PLMs, GePpeTto (GPT-2) and iT5, using a dataset of LDs. We experiment with different representations of the personal knowledge extracted from LDs for grounded response generation, including the graph representation of the mentioned events and participants. We evaluate the performance of the models via automatic metrics and the contribution of the knowledge via the Integrated Gradients technique. We categorize the natural language generation errors via human evaluations of contextualization, appropriateness and engagement of the user.

1 Introduction

The state-of-the-art dialogue systems are designed for assisting the user to execute a task, holding limited chit-chat conversations with shallow user engagement, or information retrieval over a finite set of topics. The personalization in these systems is limited to a stereotypical user model. This user model is implicitly inferred from conversations with many users, or is limited to a superficial list of persona statements (e.g., "He likes dogs") (Zhang et al., 2018). The dialogue sessions are disconnected and the shared information across sessions is negligible and close to none.

Longitudinal Dialogue (LD) is one of the most challenging types of conversation for human-machine dialogue systems. LDs are multi-session interactions that encompass user-specific situations, thoughts, and emotions. Dialogue systems designed for LDs should interact uniquely with each user about personal life events and emotions over multiple sessions and long periods of time (e.g. weeks). Through each session in LDs, the dialogue system must learn about the user's personal space of events and participants and social interactions, and engage the user in personal dialogues regarding their thoughts, feelings, and personal and world events.

Figure 1 shows an example of three types of human-machine dialogues: task-based, open-domain chit-chat and LD. The user dialogues with the tasked-based dialogue system consists of either independent command-and-control exchanges such as on Day 1, or a task-driven dialogue such as on Day 2. The user model in this system is not personal as it adopts a stereotypical model -implicitly - inferred from dialogue corpora with multiple users. In the open-domain chit-chat dialogue, the dialogue does not include the execution of any explicit task, and the model engages the user in a conversation about movies and news. A common characteristic of task-based and open-domain dialogues is the fact that there is no personal information carried to the next dialogue session. The system does not update/modify the user model with each dialogue session and the level of personalization is intact from one interaction to the other (Personalization in the natural language processing and dialogue models could be added based on the voice user interface requirements and could include the exploitation of personal information such as contact directory, preferences, etc.).

In contrast, the model designed for the LD must account for three main differences compared to the other two systems; A) the contents of the LD are not about general information or knowledge matters as LDs encompass personal emotions, user

Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), pages 1–11

Time	Task-Based Dialogue	Open-Domain Chit-Chat	Longitudinal Dialogue
Day 1	**User:** Set an alarm for Monday 8 a.m.	**User:** Do you like Shining?	**User:** My boss set a meeting with me tomorrow. I am a bit worried to meet her.
	Machine: I saved the alarm.	**Machine:** *The* Shining? Yes I like most of Kubrick's movies.	**Machine:** Why? Did she say what the meeting is going to be about?
	User: Remind me to call Mike at 5 p.m.	**User:** Cool! I am going to watch the shining for the first time.	**User:** I think she wants to fire me. She has been searching for an other cashier recently.
	Machine: Reminder Saved.	**Machine:** I think you are going to enjoy it.	**Machine:** Uncertainties make us anxious. There may be another explanation.
Day 2	**Machine:** How may I help you?	**Machine:** What's the most interesting thing you've heard recently?	**Machine:** I remember you had a meeting with your boss today. How did it go?
	User: I want an expensive italian restaurant.	**User:** I learned that the Queen of England's last corgi died	**User:** Oh it went well haha! She wants me to train a new cashier. She will be my colleague.
	Machine: I found *Duomo* in the center. Should I book a table?	**Machine:** Wow! Was it a cardigan welsh corgi?	**Machine:** Good! So no bad news! Are you relaxed now?
	User: Yes, for two. Tonight at 8pm.	**User:** No it was pembroke welsh corgi!	**User:** Yeah. They are planning to expand the store. I am not fired! haha!
	Machine: Done! I booked a table for two at *Duomo* for 8 pm.	**Machine:** It was called Candy I think.	**Machine:** I am glad you feel better! Sometimes it's hard to control our emotions.

Figure 1: Examples of a task-based dialogue, a chat-chit, and a Longitudinal Dialogue (LD) in two different sessions. The dialogue system for LDs needs to learn about the user in a timely manner and engage her in a personal conversation encompassing her life events, thoughts, and emotions.

and time-specific situations, and participants; B) the sessions are not disconnected dialogues and we can not model them as stand-alone interactions. In contrast, they belong to a multi-session interaction unique to the individual user, where the information shared in each interaction creates a common ground between the machine and the user. For each interaction, the system must engage the user in a dialogue respecting the common ground based on the information shared in the previous interactions, as well as the novel information in the new dialogue history; C) the machine has to extract the personal information presented in the user responses to construct and update the user model and respond coherently. Similar to a natural interaction between human speakers, the model has to gradually become acquainted with the user throughout the dialogues and not from a superficial list of sentence-based persona descriptions.

There has been limited research on personal conversations with users over a long period of time. Engaging the user to elaborate on personal situations and emotions is a challenging task and designing appropriate collection/elicitation methodologies is not straightforward. As a result, research on multi-session dialogues resorts to crowd-sourcing datasets with superficial persona statements and pretended longitudinality (Xu et al., 2022a,b; Bae et al., 2022). Meanwhile, studies on LDs have been limited to inferring user's attributes such as age and gender (Welch et al., 2019b), or next quick-response selection from a candidate set of "yes," "haha," "okay," "oh," and "nice" (Welch et al., 2019a).

In this work, we study the task of response generation in LDs. Response generation in LDs is subject to appropriateness and accuracy as well as personalization and engagement of the user. The level of personalization in LDs is beyond a set of personal preferences and can not be learned from a limited set of persona statements ("*I like cars*" does not necessarily imply that I like to talk about cars in my interactions). The generated response needs to respect individuals' states, profiles, and experiences that vary among users and dialogue sessions. Therefore, we can not collect a massive knowledge base of user models that can suit all individuals and scenarios. The dialogue system should learn about each user and derive the individual user model through/from the previous dialogue sessions to generate a personal response that is coherent with respect to the dialogue context as well as the previous dialogue sessions.

We investigate the applicability of general-purpose Pre-trained Language Models (PLM) for grounded response generation in LDs. We study whether PLMs can generate a response that is coherent with respect to the dialogue history and grounded on the personal knowledge the user has shared in previous interactions. We conversation-

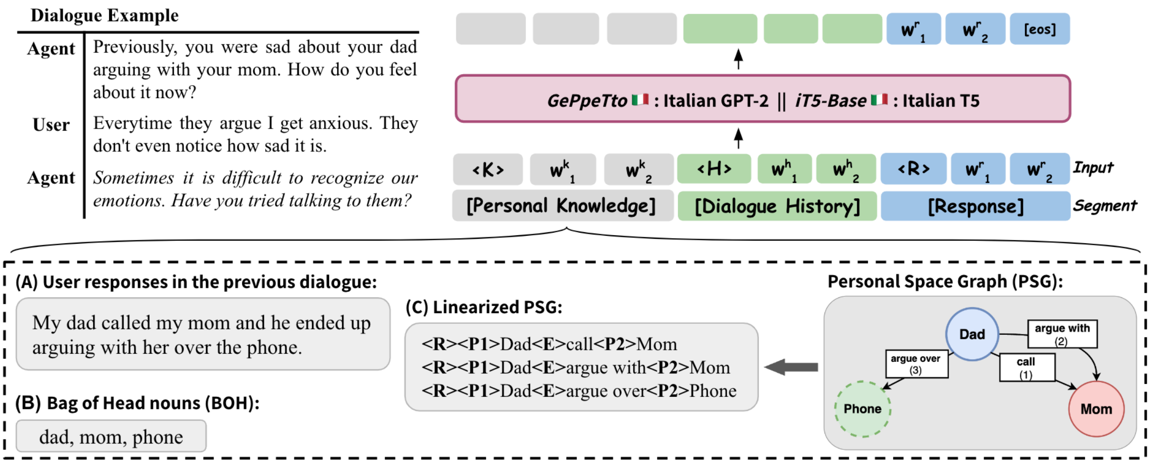

Figure 2: An example of a longitudinal dialogue. The user responses in the previous dialogue session are used as personal knowledge for grounded response generation. The knowledge is presented to the model as A) Unprocessed text (*RAW*); B) Bag of Head nouns (*BOH*); and C) Personal Space Graph (*PSG*) of events and their participants in linearized format. The model then encodes the dialogue history and the knowledge piece and generates a response candidate (the last agent turn in the dialogue example).

ally fine-tuned two recent PLMs, GePpeTto (GPT-2) (De Mattei et al., 2020) and iT5 (Sarti and Nissim, 2022), using a dataset of LDs about real-life events, feelings, and situations that the user has experienced. We use the responses each individual user shared in the previous dialogue sessions with the system as personal knowledge and evaluate whether grounding the generation on such knowledge results in more appropriate and personal responses. In previously published research on grounded generation, the knowledge sequence is provided to the model as-is. In this work, we experiment with three different representations of the knowledge piece; A) *Raw* as unprocessed text, similar to the previously published research; B) bag of head nouns as a distilled syntactic representation of the knowledge; C) graph representation of the events and participants mentioned in the user responses (Mousavi et al., 2021b). An example of a dialogue and different representations of the corresponding personal knowledge is shown in Figure 2.

We evaluate the performance of the models and the impact of different knowledge representations through automatic and human evaluations, as well as explainability studies using the Integrated Gradients technique (Sundararajan et al., 2017). Our contributions can be summarised as follows:

- To the best of our knowledge this is the first study on the task of response generation in LDs.
- We conversationally fine-tune two PLMs with and without grounded response generation on personal knowledge. We study the performance of the models and how different representations of knowledge can affect generation quality.
- We evaluate and compare the performance of the models using automatic evaluation, including explainability studies, and human evaluations, including studying the sub-dimensional errors made by each model.

2 Literature Review

Grounded Response Generation PLMs have achieved comparably well performance for open-domain chit-chats (Zhang et al., 2020), goal-oriented agents (Thulke et al., 2021) and question answering (Zhao et al., 2020). However, such models can generate inappropriate and/or generic responses which can lead to ethical problems and low user engagement (Zhang et al., 2020). Research to address this problem and improve the generation quality includes grounding the generation on external knowledge content. The selection of the knowledge source to ground the generation has been studied as an individual component (Hedayatnia et al., 2020), as well as a joint task along with response generation (Zhao et al., 2020; Huang et al., 2021).

Personal Dialogue Research on personalized response generation has focused on persona descriptions and synthetic sets of user preferences and profiles. Zhang et al. (2018) collected Persona-Chat dataset of open-domain dialogues using crowd workers, where the workers were instructed to impersonate as speakers with synthetic personas of 5 sentences. This dataset has been studied for personal response generation by fine-tuning PLMs (Wolf et al., 2019; Kasahara et al., 2022), by learning the users' persona from the dialogues samples rather than the persona descriptions (Madotto et al., 2019), or investigating different representations of persona statements (Huang et al., 2022). While the mentioned work focused on personalization in open-domain dialogues, Joshi et al. (2017) generated profiles consisting of gender, age, and food preference permutations for the user side in restaurant booking dialogues, which was used in another work (Siddique et al., 2022) to generate personalized responses in a task-based dialogue.

Multi-session Dialogue Studies on multi-session dialogues have been limited to simulated longitudinality and superficial persona. Xu et al. (2022a) extended the Persona-Chat dataset to a multi-session chat dataset with 4 to 5 sessions, by instructing crowd-workers to impersonate the role of returning dialogue partners in the first session (extracted from the Persona-Chat dataset) after a random amount of time. The workers were explicitly asked not to discuss any personal and real-life matters but play the role defined by the persona statements. This approach was further used by Bae et al. (2022) to extend an existing dataset of persona chats in Korean to multi-session dialogues. Xu et al. (2022b) proposed a framework for persona memory in multi-session dialogues and collected a dataset of persona chats in Chinese via crowd workers.

3 Experiments

3.1 Dataset

The dataset of LDs used in this work (Mousavi et al., 2021a) consists of two dialogue sessions for each individual user. The first dialogue session is a set of personal human-machine conversations with real users encompassing their personal life events and emotions. These dialogues are collected from a group of 20 Italian native speakers receiving therapy to handle their distress more effectively. Throughout the interaction, the machine prompts the user to engage her in the recollection of daily life events the user has experienced, while the user shares details about the events and participants that have activated her emotions by answering a set of questions.

For each user, the first session is then followed by a follow-up dialogue. These dialogues were elicited from 4 psychotherapists and 4 trained annotators supervised by the psychotherapists. In the second dialogue session, the user tends to share more details about her feelings and the possible evolution of the previously mentioned events. Meanwhile, the listener provides personal suggestions and asks questions to expand or disambiguate previously stated facts or feelings. A mock-up example of a second dialogue session and the corresponding user response in the previous dialogue is shown in Figure 2. This dataset consists of 800 2-session LDs in the mental health domain with an average of 5 turns per dialogue.

3.2 Models

We fine-tuned two state-of-the-art PLMs using the dataset of LDs.

GePpeTto: Italian GPT-2 The first model we experimented with is GePpeTto (De Mattei et al., 2020), a PLM based on GPT-2 small (12 layers of decoder, 117M parameters) (Radford et al., 2019), trained for the Italian language (13 GB corpus size). We fine-tuned the model using AdamW optimizer (Loshchilov and Hutter, 2017) with an early-stopping wait counter equal to 3 and a history window of 2 last turns.

iT5: Italian T5 The second PLM in our experiments is iT5 (Sarti and Nissim, 2022), a PLM based on T5 (Raffel et al., 2020), trained on the Italian portion of mC4 corpus (275 GB corpus size). We experimented with iT5-Small (12 layers, 60M parameters) and iT5-Base (24 layers, 220M parameters) [1]. We fine-tuned this model class using AdaFactor optimizer (Vaswani et al., 2017) with early stopping wait counter equal to 3 and a history window of 4 last turns.

3.3 Grounded Response Generation

For each user, we extracted her responses in the first dialogue session as personal knowledge to ground the response generation for the second dialogue session. We experimented with three representations of the knowledge piece:

- **(A) RAW**: We provide the responses of the user in the previous dialogue as an unprocessed knowledge piece. The average length of knowledge with this representation is 126.7 tokens.

[1] We were unable to use iT5-Large due to lack of GPU memory

- **(B) Bag of Head nouns (BOH)**: We automatically parse the user responses [2] and extract the head nouns as a distilled syntactic representation of the knowledge.
- **(C) Personal Space Graph (PSG)**: We represent the knowledge by the personal graph of the events and participants mentioned by the user Mousavi et al. (2021b). The predicates in a sentence represent an event, and its corresponding noun dependencies (subject, object) represent the participants. In this graph, the participants are the nodes while the predicates are the relations (edges) among the participants. We obtain a linear representation of the graph using an approach inspired by Ribeiro et al. (2021) in which the authors observed that providing a linearized representation of the graph to the PLMs results in outperforming the models with a graph-specific structural bias for the task of graph-to-text generation.

4 Evaluations

The fine-tuning of the models was done using 80% of the dialogues (640 second-session dialogues, 1284 samples with different turn levels), while the remaining data was split into 10% (80 dialogues, 160 samples with different turn levels) as the validation set for parameter engineering and early-stopping, and 10% as unseen test set. Each split was sampled at the dialogue level to guarantee no history overlap among splits. An example of a second dialogue session and the generated responses are presented in Appendix Table 5.

4.1 Automatic Evaluation

The results of the automatic evaluation of the models is presented in Table 1. The perplexity scores cannot be used to compare the performance between GePpeTto and iT-5 model classes as the vocabulary distributions in the pre-training phase of the two PLMs are not identical. However, the scores are comparable among iT5 variations as the same model class pre-trained using the same data. In fact, the perplexity scores indicate that iT5-Base demonstrates a better performance than iT5-Small in all combinations with knowledge representations. Therefore, we select iT5-Base among the iT5 models and focus the rest of the analysis on GePpeTto and iT5-Base.

[2]the dependency parser used is spaCy: spacy.io

Models	***nll***	***ppl***
GePpeTto	2.76	15.84
$+RAW Knowl.$	2.79	16.33
$+BOH Knowl.$	2.85	17.38
$+PSG Knowl.$	2.77	16.06
iT5-Small	2.18	8.84
$+RAW Knowl.$	2.19	8.95
$+BOH Knowl.$	2.18	8.88
$+PSG Knowl.$	2.19	8.93
iT5-Base	2.05	7.79
$+RAW Knowl.$	2.04	7.70
$+BOH Knowl.$	2.12	8.40
$+PSG Knowl.$	2.09	8.07

Table 1: Automatic evaluation of the models indicates that incorporating the knowledge slightly increases the models' perplexity (Perplexity scores can not be compared among models since the vocabulary distributions of pre-training data are not identical).

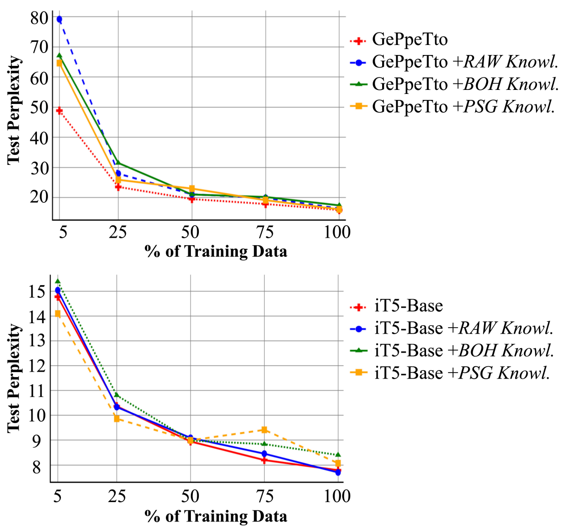

Figure 3: Perplexity score trends of the models over increasing size of the training set. The performance of GePpeTto variations is considerably improved after observing 50% of the fine-tuning training set.

Considering the small size of the LD dataset compared to the data used in the pre-training phase, we studied the impact of fine-tuning the models by optimizing the models over increasing size of the training set. The extension of the training set was gradual (the small portions are subsets of the big portions) and the performance of models was evaluated by measuring the perplexity score on the unseen test set. The results are presented in Figure 3. The performance of both models is improved considerably after observing the first 25% and 50%

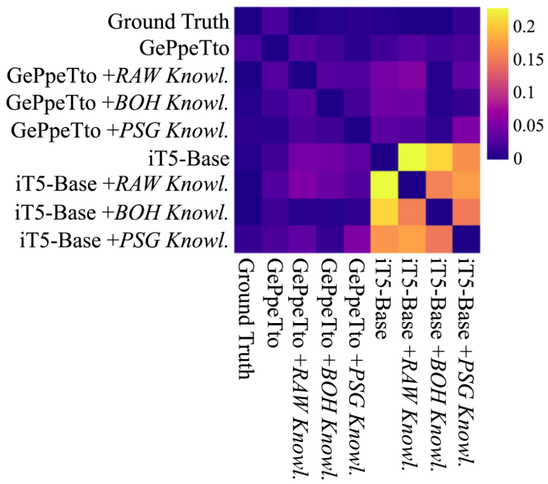

Figure 4: Lexical similarity among generated responses measured by BLEU-4 score. The results indicate a higher similarity among the responses generated by iT5-Base models.

of the train set, thus the fine-tuning has been more effective. However, in the second half of the data, both models show a steady trend while iT5-Base achieves a gradual improvement.

To investigate the impact of grounding on the response lexicalization of the models, we measured the diversity in the generated responses for the test set samples via BLEU-4 score, Figure 4. We observed that there is a higher similarity among responses generated by iT5 models, while the responses generated by GePpeTto variations are more diverse. A similar finding has been observed in the literature about the performance of auto-regressive models compared to encoder-decoder architectures regarding novelty in sequence generation (Tekiroğlu et al., 2022; Bonaldi et al., 2022). Further, responses generated by iT5-Base with *BOH* and *PSG* representations have the lowest lexical similarity. The responses with the highest lexical similarity are generated by iT5-Base with no grounding and *RAW* representation. Nevertheless, there is a negligible lexical similarity between the generated responses and the ground truth.

4.2 Human Evaluation

We sampled 50% of the unseen test set (42 dialogue histories, 80 samples with different turn levels) and evaluated the generated responses via human judges. We evaluated the responses according to four criteria using the protocol proposed by Mousavi et al. (2022):

- ***Correctness***: evaluating grammatical and syntactical structure of the response.
- ***Appropriateness***: evaluating the response to be a proper and coherent continuation with respect to the dialogue history.
- ***Contextualization***: evaluating whether the response refers to the context of the dialogue (not generic) or it consists of non-existing/contradicting information (hallucination cases).
- ***Listening***: whether the generated response shows that the speaker is following the dialogue with attention.

The annotators were asked to evaluate the response candidates and select a decision for each criterion from a 3-point Likert scale as positive (eg. Correct, Appropriate), negative (eg. Not Correct, Not Appropriate), and "I don't know". We recruited 35 native Italian crowd-workers through Prolific crowd-sourcing platform[3]. The workers were asked to perform a qualification task consisting of evaluating 5 samples (sampled from the validation set) in an identical setting to the main task. For the main evaluation, each crowd-worker annotated 3 response candidates for 10 dialogue histories, and each sample was annotated by 7 crowd-workers. We also asked the annotators to motivate their decisions for appropriateness and contextualization criteria by providing an explanation to point out possible errors in the generated response. Moreover, the ground truth was also included in the candidate set to be evaluated.

The Inter Annotator Agreement (IAA) level measured by Fleiss' κ, presented in Appendix Table 4, indicates high levels of subjectivity and complexity in *Contextualization* criterion, suggesting that it has been difficult for the annotators to assess this aspect of the responses.

The results of the human evaluation of responses are presented in Table 2 (the scores are obtained by majority voting). The evaluation of GePpeTto models shows that grounding generally worsens the performance of GePpeTto, regardless of the representation format, as the best performance is achieved by GePpeTto with no knowledge grounding. Nevertheless, *BOH* and *PSG* representations slightly improve the grammatical correctness of this model. The highest level of *Contextualization* among grounded GePpeTto models is achieved by *PSG* representation. Regarding iT5-Base varia-

[3]Prolific: https://www.prolific.co/

Models	*nll*	*ppl*	Human Evaluation			
			Correctness	**Appropriateness**	**Contextualization**	**Listening**
Ground Truth	-	-	97.62%	100.0%	97.62%	97.62%
GePpeTto	2.76	15.84	83.33%	**66.67%**	**69.05%**	**64.29%**
$+RAW Knowl.$	2.79	16.33	83.33%	59.52%	57.14%	57.14%
$+BOH Knowl.$	2.85	17.38	**92.86%**	45.24%	52.38%	42.86%
$+PSG Knowl.$	2.77	16.06	90.48%	54.76%	64.29%	50.00%
iT5-Base	2.05	7.79	**100.0%**	66.67%	73.81%	66.67%
$+RAW Knowl.$	2.04	7.70	85.71%	80.95%	80.95%	76.19%
$+BOH Knowl.$	2.12	8.40	92.86%	**80.95%**	85.71%	83.33%
$+PSG Knowl.$	2.09	8.07	95.24%	73.81%	**90.48%**	**83.33%**

Table 2: Human Evaluation of the fine-tuned models. The results show the impact of different representations of the knowledge source for grounded response generation in LDs. Refined representations of the knowledge (*BOH* and *PSG*) generally result in better performances than *RAW* representation.

tions, the results indicate that grounding improves the models' performance considerably with respect to *Appropriateness*, *Contextualization*, and *Listening*. However, it decreases the model's *Correctness* with the highest decrease caused by *RAW* representation. *PSG* representation achieves the highest level of *Contextualization* and *Listening* overall, besides the highest level of *Correctness* among grounded models. Therefore, refined representations of the knowledge (*BOH* and *PSG*) generally result in better performances compared to *RAW* representation. Nevertheless, there is still a huge gap between the performance of the best-performing model and the ground truth, suggesting the grounded PLMs are not suitable dialogue models for LDs in the mental health domain.

To gain better insight into the errors made by each model, we investigated the reasons provided by the annotators for their judgments. These results, presented in Figure 5, are complementary to the evaluation decisions, Table 2, and point out the errors that resulted in the negative evaluation of a response by the annotators. The analysis shows that grounding reduces the cases of genericness in rejected responses by GePpeTto, but results in more cases of hallucinations in the outputs of this model. The same trend is observed in iT5-Base with *RAW* representation. Furthermore, refined knowledge representations slightly escalate the genericness issue in rejected responses of iT5-Base. Nevertheless, grounding does have any positive impact on the cases of incoherence in rejected responses of the PLMs.

4.3 Generation Explainability

According to the human evaluation results, iT5-Base with knowledge grounding achieves the best performance among PLMs. We investigated the contribution of personal knowledge and different representations on the model's performance at inference time. We studied the attribution scores of the input tokens using the Integrated Gradients technique (Sundararajan et al., 2017; Sarti et al., 2023) based on backward gradient analysis. We experimented with two thresholds for the attribution scores:

- **Positive Contribution**: Based on the assumption that elements with positive scores have a positive influence on the model's performance, we investigated the tokens with positive attribution scores, However, tokens with small attribution scores have negligible contributions and thus this analysis can be noisy.
- **Significant Contribution**: To identify the tokens with significant contributions to the generation, we selected the top-25% of the tokens in the input sequence (knowledge and history) according to their attribution score. We then investigated what portion of these tokens belong to each segment of the input vector. For a fair comparison, the values are normalized over the segment length.

According to Positive Contribution analysis, 74% of the tokens in the *RAW* representation have a positive contribution to the generation with the majority (30%) of tokens being verbs and nouns. This percentage for *BOH* (Bag of Head Nouns) representation changes to 79.0%. This result suggests the importance of nouns for the model inference.

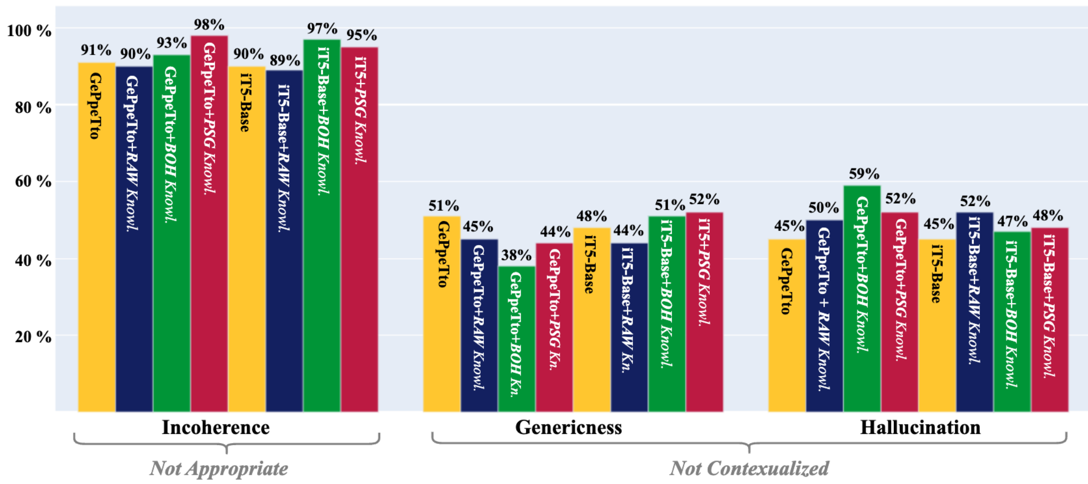

Figure 5: Explanations provided by the crowd-workers to motivate their negative judgments in *Appropriateness* and *Contextualization* criteria, represented by the percentage of the times the error category (x-axis) was selected. The figure is obtained by considering all the votes (i.e. not majority voting). Note that the labels are not mutually exclusive.

Models	***Knowl.***	***History***
iT5-Base		
$+RAW Knowl.$	44.6%	55.4%
$+BOH Knowl.$	39.5%	60.5%
$+PSG Knowl.$	38.7%	61.3%

Table 3: Percentage of tokens with significant contribution to the generation (top-25%) in knowledge and history segments of the input vector for each model.

Regarding the *PSG* representation, 55.6% of the tokens have a positive contribution to the generation (excluding the tags used for linearization), with the majority (68%) of tokens being events rather than participants.

The analysis of the tokens with significant contributions is presented in Table 3. Regarding the model with *RAW* representation, the percentage of tokens with high attribution scores is almost balanced between the knowledge and history segments. However, for the models with refined representations of knowledge (*BOH* and *PSG*), the dialogue history contains moderately more significantly contributing tokens.

5 Conclusion

We studied the task of response generation in Longitudinal Dialogues (LD), where the model should learn about the user's thoughts and emotions from the previous dialogue sessions and generate a personal response that is coherent with respect to the user profile and state, the dialogue context, as well as the previous dialogue sessions. We fine-tuned two state-of-the-art PLMs for Italian, using a dataset of LDs in the mental health domain. We experimented with grounded generation using user responses in the previous dialogue session as user-specific knowledge. We investigated the impact of different representations of the knowledge, including a graph representation of personal life events and participants mentioned previously by the user.

Our evaluations showed there is still a huge gap between the performance of the general-purpose PLMs with knowledge grounding and the ground truth. Nevertheless, we observed that a) refined representations of the knowledge (such as *BOH* and *PSG*) can be more informative and less noisy for a grounded generation; b) the encoder-decoder model exhibited more diversity in the outputs compared to the auto-regressive model; c) knowledge grounding reduces the cases of genericness in response, though it can result in more hallucinated responses.

Limitations

The dataset used in this work is in Italian and there may be language-specific limitations in the model performance. GePpeTto is the only candidate for auto-regressive models for the Italian language at the time of this research. Therefore, its performance may be limited due to the small number of parameters. We were unable to experiment with iT5-Large model due to computation power limitations.

Acknowledgements

We acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU.

References

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049. Association for Computational Linguistics.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and Lilian Tang. 2022. Personalized dialogue generation with persona-adaptive attention. *arXiv preprint arXiv:2210.15088*.

Xinxian Huang, Huang He, Siqi Bao, Fan Wang, Hua Wu, and Haifeng Wang. 2021. PLATO-KAG: Unsupervised knowledge-grounded conversation via joint modeling. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 143–154, Online. Association for Computational Linguistics.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 96–105, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021a. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, Online. Association for Computational Linguistics.

Seyed Mahed Mousavi, Roberto Negro, and Giuseppe Riccardi. 2021b. An unsupervised approach to extract life-events from personal narratives in the mental health domain. In *CLiC-it*.

Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. 2022. Evaluation of response generation models: Shouldn't it be shareable and replicable? In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2022)*. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.

Gabriele Sarti, Ludwig Sickert, Nils Feldhus, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models.

AB Siddique, MH Maqbool, Kshitija Taywade, and Hassan Foroosh. 2022. Personalizing task-oriented dialog systems via zero-shot generalizable reward function. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1787–1797.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Adapting document-grounded dialog systems to spoken conversations using data augmentation and a noisy channel model. *arXiv preprint arXiv:2112.08844*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Charles Welch, Verónica Pérez-Rosas, Jonathan K Kummerfeld, and Rada Mihalcea. 2019a. Learning from personal longitudinal dialog data. *IEEE Intelligent systems*, 34(4):16–23.

Charles Welch, Verónica Pérez-Rosas, Jonathan K Kummerfeld, and Rada Mihalcea. 2019b. Look who's talking: Inferring speaker attributes from personal longitudinal dialog. *arXiv preprint arXiv:1904.11610*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. 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.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

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? *arXiv preprint arXiv:1801.07243*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Appendix

Models	Inter Annotator Agreement Level measured by Fleiss'κ				
	Appropriateness	*Contextualization*	*Correctness*	*Listening*	IAA per Model
GePpeTto	0.27	0.14	0.64	0.15	0.32±0.10
$+RAW Knowl.$	0.42	0.22	0.36	0.27	0.36±0.11
$+BOH Knowl.$	0.23	0.05	0.31	0.11	0.27±0.05
$+PSG Knowl.$	0.30	0.39	0.34	0.26	0.42±0.06
iT5-Base	0.24	0.19	0.06	0.18	0.27±0.04
$+RAW Knowl.$	0.18	0.03	0.30	0.21	0.19±0.06
$+BOH Knowl.$	0.21	0.17	0.58	0.24	0.26±0.09
$+PSG Knowl.$	0.17	0.06	0.27	0.14	0.19±0.12
IAA per Dimension	0.31±0.09 **Fair**	0.20±0.06 **Poor**	0.43±0.20 **Moderate**	0.25±0.10 **Fair**	-

Table 4: Inter-Annotator Agreement (IAA) level calculated by Fleiss' κ for each model and criterion. Low IAA level for *Contextualization* suggests a high level of subjectivity in this criterion.

Dialogue Example		
Knowledge	**User Responses from Previous Dialogue:**	I only worked in the morning, after lunch I went to the bank with Martina to activate her prepaid card since the one she is using now is about to expire. I accompanied Paolo to fix the bike and then I went to the dentist. When I came back, I made sauce for the week and then I ironed. I ran a lot but it was not a very stressful day. I hope to be able to go to bed early to be able to rest!"
Dialogue History	**Dialogue Stimulus:**	I remember you had a positive feeling about what I read in your notes [yesterday] "*I only worked in the morning, after lunch I went to the bank with Martina to activate her prepaid card since the one she is using now is about to expire. I ran a lot but it was not a very stressful day. I hope to be able to go to bed early to be able to rest!*". Do you have any new thoughts on this?"
	User:	I think that working in the morning and in the afternoon was not tiring, actually it was pleasant. I was also able to go to bed early enough, and I am well rested.
Response Candidates	**Ground Truth**	Good! Did you even manage to spend time with your daughter?
	GePpeTto	I see. Do you often find yourself away from each other?
	+RAW Knowl.	I see. Maybe you could try doing more flexible jobs to increase your anxiety.
	+BOH Knowl.	I'm glad to hear that! Have a nice day!
	+PSG Knowl.	Good, it's important to know yourself.
	iT5-Base	I'm glad you feel rested enough. Have you tried any relaxation exercises?
	+RAW Knowl.	I'm glad you were able to go to bed early. If you want to write me about this, we can talk about it anytime.
	+BOH Knowl.	I'm glad you rested well. Did you try any relaxation exercises?
	+PSG Knowl.	I'm happy to you rested well. If you want to write more about this, we can talk about it anytime.

Table 5: Example of a second dialogue session, the corresponding user turns in the first session as personal knowledge, and the generated responses (English translation).