PRODIGy: a PROfile-based DIalogue Generation dataset

Daniela Occhipinti[1,2]**, Serra Sinem Tekiroğlu**[1]**, Marco Guerini**[1]

[1]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
`docchipinti@fbk.eu, tekiroglu@fbk.eu, guerini@fbk.eu`
[2]University of Trento, Italy

Abstract

Providing dialogue agents with a profile representation can improve their consistency and coherence, leading to better conversations. However, current profile-based dialogue datasets for training such agents contain either explicit profile representations that are simple and dialogue-specific, or implicit representations that are difficult to collect. In this work, we propose a unified framework in which we bring together both standard and more sophisticated profile representations by creating a new resource where each dialogue is aligned with all possible speaker representations such as communication style, biographies, and personality. This framework allows to test several baselines built using generative language models with several profile configurations. The automatic evaluation shows that profile-based models have better generalisation capabilities than models trained on dialogues only, both in-domain and cross-domain settings. These results are consistent for fine-tuned models and instruction-based LLMs. Additionally, human evaluation demonstrates a clear preference for generations consistent with both profile and context. Finally, to account for possible privacy concerns, all experiments are done under two configurations: *inter-character* and *intra-character*. In the former, the LM stores the information about the character in its internal representation, while in the latter, the LM does not retain any personal information but uses it only at inference time.

1 Introduction

Dialogue agents capable of holding human-like interactions (i.e. coherent, natural and engaging conversations) have drawn increasing interest in the fields of AI and NLP, becoming a key topic and challenge in both industry and academia.

Unlike task-oriented systems, in which the goal is to help users solve specific tasks, open-domain dialogue systems (or chit-chat systems) are designed to talk about a wide range of topics and to possibly display a well-defined and consistent profile (Kann et al., 2022). In this work, we aim to understand the role of profile information in open-domain dialogue systems.

In fact, despite the impressive improvements that conversational agents have recently shown, thanks to the continuous development of neural models (Radford et al., 2019; Devlin et al., 2019; Scao et al., 2022; Zhang et al., 2022; Peng et al., 2022), these agents are not always able to maintain coherence, generating sometimes inconsistent or uninformative responses. This leads to a significant reduction in user engagement and trust (Li et al., 2016b, 2020). In this scenario, endowing dialogue systems with profile information is a key element to improve models' ability to produce more fluent, consistent and informative responses (Li et al., 2016a; Zhang et al., 2018; Zemlyanskiy and Sha, 2018; Song et al., 2019; Majumder et al., 2021; Mazaré et al., 2018).

The concept of *profile* in a dialogue can refer to three aspects: *personalisation*, *persona*, and *personality*. *Personalisation* refers to employing users' information to drive engagement and help them satisfy their needs (Vesanen, 2007). *Personality* is a psychological concept meant to capture how we behave and react to the world (Allport, 1937; Vinciarelli and Mohammadi, 2014). The notion of *persona* can have diverse meanings in literature. In this work, we will stick to the definition provided by Li et al. (2016a), according to which the persona is the character that an artificial agent plays during conversational interactions and includes elements such as background facts, language and interaction style.

Several approaches have been investigated to condition the generation of dialogues with the in-

formation about persona (Li et al., 2016a; Mazaré et al., 2018; Welch et al., 2022; Zhang et al., 2018; Song et al., 2021; Zheng et al., 2020; Cao et al., 2022; Majumder et al., 2020; Liu et al., 2020; Majumder et al., 2021; Zheng et al., 2019). However, most of these approaches are sporadic and disjointed, considering only one specific dimension of the persona at a time. In fact, current persona-based dialogue datasets contain either an *explicit* representation (i.e. a handful of simple and dialogue-specific sentences about the user), or an *implicit* representation (i.e. a collection of previous dialogues of the user) that is often difficult to attain. Thus, these approaches fail to model the complexity of human communication which is influenced by the interaction of multiple aspects.

In this paper, we aim to understand the impact of diverse profile representations in the development of dialogue systems by comparing and benchmarking them. To this end, we propose a unified framework in which pre-existing profile representations, such as language style, gender and personality, are brought together with novel and more complex representations of the persona, such as biographies. Therefore, we introduce the PRODIGy (PROfile-based DIalogue Generation) dataset, a new resource that aligns dialogues with all the representations considered above. PRODIGy was created starting from the Cornell Movie Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011), which includes movie script dialogues, and adopting the character IDs and binary gender labels from the original corpus. This allows us to avoid privacy issues that might derive from using real users' data, and facilitates the distribution. Moreover, the dataset has been aligned with external resources containing characters' profiles, and it can be further expanded by adding new scripts or scripts in other languages. In Figure 1, we showcase an example from PRODIGy.

We use PRODIGy with generative LMs, either via fine-tuning or instruction prompting, and test different configurations using the profile dimensions. We also experiment with two ways of partitioning the PRODIGy dataset: (i) inter-character, in which the characters in the test set are not in the training set, and (ii) intra-character, in which the same characters are both in the training and test sets. This allow to account for different privacy levels: in one case, the LM does not retain any personal information, while in the second,

Figure 1: Example of a dialogue with diverse speaker's profile information provided.

user information is stored in the LM internal representation. To validate the proposed resource, we conduct experiments by employing both automatic metrics and human evaluation. In particular, we conduct experiments both in-domain and cross-domain. The automatic results of the in-domain experiments show that training LMs with diverse aspects of a profile, both separately and jointly, significantly improves the models' predictive capabilities. Furthermore, when given as an instruction, the profile information can also improve the performances of non-fine-tuned LLMs. Finally, in the cross-domain setting, the automatic evaluation shows that the models trained on PRODIGy exhibit better generalisation abilities than those trained on other resources.

Conversely, the human evaluation results indicate that evaluators often preferred generic responses for their broader applicability. However, when responses were consistent with both profile and dialogue they were favoured. Profile information proves to be beneficial especially in dialogues with restricted context and when this information is disclosed to evaluators, profile based responses are perceived as more appropriate.

2 Related Work

We discuss three main topics relevant to our work: (i) theories on persona and personality (ii) available datasets for persona-based generation and (iii) persona and personality based models.

Persona and Personality The way we communicate is closely related to our social status, gender and motivations, and can convey relevant information on our psychological state (Pennebaker et al., 2003). These aspects are closely related to the concepts of *persona* and *personality*, which fall under the more general concept of *profile* (Schiaffino and Amandi, 2009). *Persona* can be defined as the character that an artificial agent acts during a conversation and it is a combination of identity factors, such as background facts, language use, and communication style (Li et al., 2016a). *Personality* is a psychological concept grasping different behaviours, feelings and way of thinking (Allport, 1937; Vinciarelli and Mohammadi, 2014). It can be formalised using theoretical frameworks called *trait models*, such as *Big Five* (John et al., 1991) and the *Myers-Briggs Type Indicator (MBTI)* (Myers, 1962).

Persona-Based Dialogue Datasets Several dialogical datasets contain a persona representation, many of which were collected starting from social media such as Twitter, Reddit, Weibo or Kialo. However, these datasets have various limitations. They can include short conversations, thus failing to fully represent real dialogues (Li et al., 2016a; Mazaré et al., 2018); they can rely only on the dialogue history of the users (Qian et al., 2021); they may include only generic persona representations such as gender or age (Zheng et al., 2019; Zhong et al., 2020); finally, they may not consider linguistic style, being based on controlled and redacted conversations (Scialom et al., 2020). Other resources were collected from television series transcripts (Li et al., 2016a), but are small and, therefore, not sufficient to train open-domain dialogue models. One of the most widely used persona-based dataset is Persona-Chat (Zhang et al., 2018), collected in a controlled crowd-sourcing environment. However, in this case, the persona is limited to a generic fact-based representation (e.g. "I just got my nails done") which is specific to that single dialogue and leaves out complex aspects, such as linguistic style or biographical history.

Persona/Personality Based Dialogue Models Several approaches have been investigated to condition the dialogue generation through the persona information. On the one hand, diverse studies were based on resources in which the persona is represented by the users' previous dialogues (Li et al., 2016a; Mazaré et al., 2018; Zhong et al., 2020). On the other hand, a line of research has been built on Persona-Chat. Some approaches employed this dataset to train persona-based models in under-resourced scenarios (Song et al., 2021; Zheng et al., 2020; Cao et al., 2022). Other methodologies used Persona-Chat to test commonsense expansion (Majumder et al., 2020), mutual perception persona (Liu et al., 2020), or enriching persona information through background stories (Majumder et al., 2021). However, these studies present the same limitations of the resources they rely on. Regarding the personality-driven generation, few seminal studies have been conducted (Mairesse and Walker, 2007, 2008; Gill et al., 2012). However, they leave the interactions between personality and persona unexplored.

Dialogue models have been evaluated using word overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), that are insufficient when multiple possible responses can exist for a given context but only one gold reference is provided. Alternative metrics that do not use word overlap such as Perplexity (Jelinek et al., 1977), Conditional Perplexity (Su et al., 2021), and Accuracy at N (Welch et al., 2022) are also used. In our view, these metrics are better suited for dialogue generation, as they evaluate the model's likelihood of generating the gold response, without confining the evaluation to mere word overlap. However, human evaluation of the generated content is a preferable solution (Liu et al., 2016), hence, it has been adopted for dialogue generation evaluation (Li et al., 2016a; Song et al., 2019; Zhong et al., 2020; Song et al., 2021; Ji et al., 2022). In this paper, we will employ non overlap metrics and human evaluation.

3 Construction of the PRODIGy dataset

To build the PRODIGy dataset, we started from the Cornell Movie Dialogs Corpus, a dataset of dialogues from movie scripts that includes metadata about movie genre, release year and characters' gender (Danescu-Niculescu-Mizil and Lee, 2011). The dialogues in the Cornell Movie Dialogs Cor-

pus are between two actors and have an average length of 4 turns. The reason for using this resource as a starting point is three-fold: (i) **Data Persistency and Accessibility:** it eliminates privacy issues or ephemerality problems (Klubicka and Fernández, 2018) that would arise from collecting data from real users and, therefore, facilitate the distribution of PRODIGy to the research community; (ii) **Data Enrichment:** it is possible to enrich PRODIGy with the profile of movie characters through the alignment with external web resources containing information about characters and movie plots; (iii) **Data Expansion:** it leaves room for further development/extension; for example, it can be aligned with similar movie script resources in other languages or new movie scripts.

Dialogical Information. Following previous approaches (Li et al., 2016a; Qian et al., 2021), we provide an implicit representation of each character's persona through a collection of characters' dialogues. Thus, we can represent the characters' linguistic styles. To this end, we included in PRODIGy only the characters with at least 50 dialogues in the Cornell Movie Dialogs Corpus.

Personality Information. To associate each character with *personality* information, we cross-referenced the Cornell Movie Dialogs Corpus with the Personality Database (PDB)[1] website. PDB is a widely used social platform in which users can assign personality types from several trait models to fictional characters and real famous people. We use this platform as a provider of crowd-sourced characters' personality annotations.

The procedure for pairing the characters with corresponding PBD entries consists of an automatic alignment via `movie_title+year` query, followed by a manual refinement for possible mismatches over franchises (i.e. movies, tv-series and other media belonging to a common universe, such as Indiana Jones), characters names, etc. Details of the alignment procedure are provided in Appendix A.

Among the several trait models provided by PDB on each character's web page, we focused on MBTI since it is widely studied and it was the most voted model by users, thus proving a more stable and reliable crowd-annotation. The MBTI trait model takes into account 16 personality types obtained from the combination of 4 dichotomies:

introversion or extroversion, sensing or intuition, thinking or feeling, and judging or perceiving.

In line with the definition of personality traits, which posits their stability over time, we assigned a unique MBTI personality type to each character. This differs from the approach of Jiang et al. (2020), who assigned a different personality for each dialogue in which the character is present. Finally, to ensure the reliability of the annotations, we discarded the characters (and related dialogues) with less than 5 user votes and used the personality type derived from the majority of votes on each MBTI dichotomy.

Biographical Information. The third step was to provide the characters with explicit persona representations that serve as background information for all the dialogues in which the character is present. Inspired by the concept of *background story* by Majumder et al. (2021), we aim to provide a representation that goes beyond simple facts. To this end, we consider the biographical information. We scraped the biographies of the characters already annotated with the personality information, from Charactour.com, Fandom.com and Wikipedia. Then, to obtain the final biography, we employed a human-machine collaboration procedure[2] that included an automatic relevance-based ranking of biography sentences, followed by a manual intervention on the top 10 sentences to make them homogeneous to the style of Persona-Chat (Zhang et al., 2018) in order to make the resources comparable. Specifically, the sentences were turned into first-person and split/merged whenever necessary. However, they are conceptually and qualitatively different from Persona-Chat, as they are not limited to generic facts and capture more complex aspects of the persona. Details are provided in Appendix B.

Biography Expansion through Paraphrasing. To account for the diversity in the way a biography can be expressed and to add variability to the number of biography sentences used during training, we performed an additional step of paraphrasing similar to the one presented by Zhang et al. (2018). We used another human-machine collaboration procedure, where ChatGPT (OpenAI, 2022) produced two possible paraphrases for each sentence that are validated and post-edited by an an-

---

[1]https://www.personality-database.com/

[2]In all the human-machine collaboration procedures of this work, the human annotator was one of the authors.

Dialogues	20850
Turns	80604
Annotated Characters	339
Turns per Dialogue	4 ($\pm$3.28)
Dialogues per Character	78 ($\pm$31.21)
Sentences per Bio	8 ($\pm$1.57)
Token per Bio sentence	13 ($\pm$5.66)

Table 1: The main statistics of PRODIGy. The upper part of the table reports counts, while the lower reports averages.

notator (see Appendix B for details).

As a result of the aforementioned procedures, we obtained a dataset with more than 20K dialogues for 80K turns with 300 annotated characters. The dialogues are aligned with the following dimensions of one of the speakers: gender, personality type, character biography, and characters' dialogues. Character biographies have on average 8 sentences (from 5 to 10) of 13 tokens each. Each biography sentence has been paraphrased twice. Detailed statistics of the PRODIGy dataset are provided in Table 1. We distribute the dataset for research purposes at the following link: https://github.com/LanD-FBK/prodigy-dataset.

4 Baselines and Experiments

In this section, we propose several configurations to condition the dialogue generation with profile information. In particular, we represent profiles by using either the persona or the personality information, or both. Our aim is to analyse the impact of each representation in the generation process.

For all the configurations, we employed the DialoGPT model as our baseline since it is a generative transformer-based model pre-trained on conversation-like exchanges (Zhang et al., 2020), making it the most suitable baseline for the dialogue generation task. Since we aim to study the effect of each profile dimension, we ensured that the LM and the hyperparameters used in the fine-tuning process were constant, and we considered the type of profile information as the only variable. In particular, we fine-tuned all our models for 5 epochs with a learning rate of $1e - 6$ and a batch size of 2. We investigated several training configurations. As a baseline, we fine-tuned DialoGPT without any profile information, while in the remaining configurations we fine-tuned the model considering both

single profile dimensions and their combinations. Specifically, we concatenated the characters' profile information to the corresponding turns of the dialogues. The input syntax used in the experiments with DialoGPT is as follows (we use the example given in Figure 1 as a reference):

<|id|> u9999 <|mbti|> extrovert, sensor, feeler, perceiver <|gender|> female <|bio|>I am an actress, a star. I live in an old mansion, built for glamorous stars of 1920s Hollywood, just off of Sunset Boulevard. (...) <|start_dialogue|> What's the matter, Norma?<|endoftext|> u9999: Nothing. I just didn't realize what it would be like to come back to the old studio. I had no idea how I'd missed it.<|endoftext|> We've missed you too, dear.<|endoftext|> (...) u9999: *turn_to_be_predicted*

<|id|>,<|mbti|>,<|gender|>,<|bio|> and <|start_dialogue|> are special tokens added to the model vocabulary, and they are used to segment the input sequence. During training, each part of the profile input and its corresponding token are added or removed depending on the configuration under inspection.

Besides DialoGPT, we also experimented with GODEL (Peng et al., 2022), an instruction-based LLM for dialogue generation. Our aim is to assess the effect of adding profile information when given as an instruction to a non-fine-tuned LLM. Following, we provide an example of the input syntax for GODEL.

Instruction: given a dialog context, you need to respond as a person having the following mbti, gender and bio: "extrovert, sensor, feeler, perceiver", "female", "I am an actress, a star. I live in an old mansion, built for glamorous stars of 1920s Hollywood, just off of Sunset Boulevard. (...)" [CONTEXT] What's the matter, Norma? EOS Nothing. I just didn't realize what it would be like to come back to the old studio. I had no idea how I'd missed it. EOS We've missed you too, dear. EOS (...) EOS *turn_to_be_predicted*

Regarding the inspected configurations, we provide the description as follows:

Plain Dialogue Driven Generation In the first configuration, we fine-tuned DialoGPT and in-

structed GODEL only with the plain dialogue, without considering any profile information. This configuration will be used as a baseline to assess the improvement obtained by adding the various profile information to both models.

Personality Driven Generation In this configuration, we employ PRODIGy and the characters' MBTI types to fine-tune DialoGPT and prompt GODEL, following the insights that it is possible to generate language reflecting a certain personality type (Mairesse and Walker, 2007, 2008; Gill et al., 2012).

Persona Driven Generation In this configuration, we employ the implicit (i.e. linguistic and stylistic information) and explicit (i.e. gender and biography sentences) persona representations in PRODIGy either individually or jointly. This enabled us to analyse the effect of each representation and combination in the dialogue generation.

Firstly, we used the characters' dialogues as implicit persona representation (Li et al., 2016a; Qian et al., 2021). We fine-tuned DialoGPT on PRODIGy, using the IDs of the characters to aggregate their lists of dialogues and capture their linguistic styles. Secondly, inspired by Zheng et al. (2019) and Schwartz et al. (2013), we considered gender as another persona representation to fine-tune DialoGPT and instruct GODEL. Then, motivated by Zhang et al. (2018), we endowed DialoGPT and GODEL with information about the persona represented in the form of biography sentences. Our aim is to generate non-generic and informative responses that are consistent with both the dialogues and the biography sentences.

Inter-Character and Intra-Character Configurations Using PRODIGy, we set up two configurations to train DialoGPT: *inter-character* and *intra-character*. In the first configuration, the test characters are not used at training time. In the second configuration, at training time the system learns about the specific characters to be predicted at test time. In both cases, we use only 5 biography sentences, following the study by Zhang et al. (2018). These two configurations respond also to certain privacy concerns: in one case, the LM stores the information about the user in its internal representation, while in the second, the LM does not retain any personal information but uses it only at inference time.

5 Evaluation

In this section, we describe the automatic metrics and human evaluation design adopted for the validation of our resource and models.

5.1 Automatic Metrics

To evaluate the model performances, we use two automatic metrics: *Conditional turn Perplexity* (Su et al., 2021) and *Average Accuracy at N* (Welch et al., 2022).

Conditional Perplexity ($CPPL$) is the perplexity of a gold turn given the context. The purpose of $CPPL$ (Equation 1) is to compute the likelihood of a turn given a dialogue history and possible profile information. The $CPPL$ is the reciprocal of the product of the probability of each word in the response $R(x)$ based on the context x, where T represents the number of words in the response $R(x)$.

$$CPPL = \prod_{i=1}^{T} \frac{1}{(P(R(x)_i|x))^{\frac{1}{T}}} \quad (1)$$

With Average Accuracy at N (Acc @N), the prediction of a word from a gold turn is considered correct if it occurs within the top N most probable words given by the model. Using different values of N we can have insights for specific decoding choices, e.g. with $N = 1$ we can address greedy decoding, while with $N = 40$ we can address *top-k* decoding with its default *k* value (Radford et al., 2019). We adopted these metrics to evaluate our models in both in-domain (i.e., on PRODIGy) and cross-domain (i.e., on Persona-Chat) scenarios.

5.2 Human Evaluation

In addition to the automated metrics, we performed a human evaluation by employing six human evaluators, including four PhD students in Computer Science and two MSc students in Data Science. In a preliminary phase, outputs from both beam search and top-p decoding were qualitatively evaluated. Given the better quality of top-p generations, they are used in the subsequent phase. Each evaluator is given 100 dialogues: for 50 the speakers' profile information is provided, while for the other 50 it is not disclosed. This allows us to assess the impact of profile information on the evaluators' judgements. We focus on the output from four models trained during the inter-character experiments: the model trained

on dialogues only and the models trained also with single profile dimensions. The evaluators are presented five possible responses of the target speaker for each dialogue, including the gold response to better inspect the quality of the generations. The evaluators are asked to rank the five responses without ties, based on the perceived likelihood of being the target speaker's actual response. The ranking scale ranges from 1 (most likely) to 5 (least likely). In total, we collect 3000 evaluations. We then conducted a post-hoc qualitative interview with the evaluators.

6 Analysis and Results

In this section, we provide a detailed description of the following experiments: (i) Inter-Character Experiments, (ii) Intra-Character Experiments, (iii) Cross-Domain Experiments. In these three experimental settings, we consider the target speaker's profile, excluding the interlocutor's. Given just the dialogue context, or both context and profile information, we aim to predict the target speaker's final turn.

6.1 Inter-Character Experiments

In this setting, we partitioned PRODIGy making sure that the characters in the test set are not present in the training set in consistent with the experiments by Welch et al. (2022).

As the first set of experiments, we tested three strategies to add variability to the biographies during training: (i) *Bio*, trained using the original top-5 biography sentences, (ii) *Bio$_{rand}$*, by randomly selecting, for each dialogue, 5 biography sentences from the corresponding full set of biography sentences of the character, (iii) *Bio$_{par}$*, by randomly selecting 5 sentences for each dialogue from the original biography or from the paraphrases[3].

Table 2 shows the effect of randomly choosing 5 sentences out of the full set of biography sentences for each training example (Bio vs. Bio$_{rand}$): randomisation leads to an improvement in terms of $CPPL$. Training the models by mixing original and paraphrased biographies, thus increasing lexical variability, improves the performance even further in terms of both $CPPL$

---

[3]We employ only 5 biography sentences to ensure (i) we stay within the DialoGPT input size length of 1024 tokens, (ii) we are consistent with Persona-Chat configuration.

(98.27 for Bio$_{par}$ vs. 117.26 for Bio) and Acc @N (e.g. for Acc @40, 0.794 for Bio$_{par}$ vs. 0.780 for Bio). Thus, in the following experiments with DialoGPT, we will always use Bio$_{par}$ as the reference configuration.

Config.	CPPL	Acc @40	Acc @10	Acc @1
Bio	117.26	0.780	0.647	0.294
Bio$_{rand}$	106.24	0.750	0.653	0.302
Bio$_{par}$	**98.27**	**0.794**	**0.661**	**0.307**

Table 2: DialoGPT results of the addition of variability to biography sentences on PRODIGy test set (Inter-Character)

The performances of the models trained on the profile information are discussed in Table 3. In terms of Acc @N, these models perform better than the Plain Dialogue model, trained without the profile information, especially for Acc @40 and Acc @10. For all the Acc @N metrics, the best performing model is MBTI, achieving an Acc @40 of 0.794, an Acc @10 of 0.665, and an Acc @1 of 0.317. However, in terms of Acc @40 and Acc @10, all the models trained on single profile information display nearly equal performances. Regarding Acc @1, both Bio$_{par}$ and Gender perform similarly, deviating only by approximately 0.01 from the MBTI model's score. Also in the case of models trained by combining multiple profile dimensions, the to Acc @N scores do not differ significantly. In terms of $CPPL$, the worst performing model is Plain Dialogue (with a score of 541.16), whereas the best performances are attained by the models provided with profile information. Specifically, the model with the lowest $CPPL$ is Gender, with a score of 87.92. It has a comparable performance to the MBTI model. Although Bio$_{par}$ performs worse than Gender and MBTI, it is significantly better than the baseline, with a score of 98.27, thus proving the effectiveness of high-level character descriptions. Gender's strong performances in $CPPL$ and Acc @N might be attributed to the fact that PRODIGy's dialogues, sourced from the Cornell Movie Dialogs Corpus, may present gender-specific linguistic patterns (Schofield and Mehr, 2016), enabling the model to effectively discern and incorporate such characteristics. In general, the results show that adding profile information, either alone or jointly, strongly improves the model performance in terms of generalisation.

Config.	CPPL	Acc @40	Acc @10	Acc @1
MBTI	89.30	**0.794**	**0.665**	**0.317**
Gender	**87.92**	0.792	0.664	0.306
Bio$_{par}$	98.27	**0.794**	0.661	0.307
Plain Dialogue	541.16	0.689	0.585	0.298
MBTI + Gender	**91.50**	**0.757**	**0.660**	**0.311**
Gender + Bio$_{par}$	96.31	0.756	0.658	0.299
MBTI + Bio$_{par}$	100.35	0.753	0.653	0.296
MBTI + Gender + Bio$_{par}$	91.65	0.761	0.660	0.302

Table 3: DialoGPT results on PRODIGy test set (Inter-Character)

In Table 4 we report the results obtained by prompting GODEL with the profile information. The values of $CPPL$ and Acc $@N$ show that, even when it is given only as an instruction to a model, profile information can improve the performances. In particular, this is demonstrated by the higher $CPPL$ value of Plain Dialogue compared to the scores of MBTI and MBTI + Gender (24.00 vs 12.46). Also in terms of Acc @40 and Acc @10, MBTI + Gender turned out to be the best performing model. In terms of Acc @1, the best performing models are Bio and Plain Dialogue, with a score of 0.027, although they do not yield much better performances than the other models. These results show that profile information is beneficial also when prompted to non-fine-tuned instruction-based LLMs. It's important to state that, while GODEL may seem to outperform DialoGPT in terms of $CPPL$, a direct comparison between their metrics is not valid as these models are trained on distinct datasets and have a different vocabulary size.

6.2 Intra-Character Experiments

In the second set of experiments, we partitioned PRODIGy with the same character existing in both training and test sets. Our aim is to simulate a scenario in which we can access the information about a character already at training time, both explicitly (i.e. MBTI, gender, and biography) and implicitly (i.e. the character's dialogues, captured by the character ID, grasping their language style).

As shown in Table 5, endowing the model with the dialogical information (Char ID) provides the best results in terms of $CPPL$ (55.25). This is because the model learned the character's vocabulary and language style during training, enabling improved predictions. In terms of Acc $@N$, the

best performing model is Bio (achieving 0.833 of Acc @40, 0.712 of Acc @10, and 0.348 of Acc@1), although the other profile-based models exhibit similar performances. The Plain Dialogue model proves to be the weakest in terms of both $CPPL$ and Acc $@N$ (except for Acc @1, in which it slightly outperforms Gender), proving once again that training models through profile information is beneficial. Combining the biographical information and the Char ID further improves the efficiency of the models in terms of $CPPL$, such that, adding a high-level description of the character leads to the higher CPPL values of 54.89 for Char ID + Bio and 53.23 for Char ID + MBTI + Bio. The scores in Acc $@N$ show that, when combined with the dialogical information (Char ID), the biographical information improves the predictive ability of the model more than Gender and MBTI. Although in terms of CPPL the best performing model is Char ID, the efficiency of the models having explicit profile information do not differ significantly from the best model. Regarding the models trained with profile information jointly, the best performances are achieved by those trained with the biographical information of the characters. Generally, models perform better in the Intra-Character setup than in the Inter-Character since they are trained with the speaker's profile information and leverage it at test time.

6.3 Cross-Domain Experiments

To evaluate the generalisation capabilities of the models trained on the PRODIGy dataset in a cross-domain scenario, we also analysed the model performances, trained both with no profile information and with biographical information, on the Persona-Chat test set (Zhang et al., 2018). These results are also compared with the models trained

Config.	CPPL	Acc @40	Acc @10	Acc @1
MBTI	**12.46**	0.122	0.080	0.026
Gender	13.65	**0.126**	0.075	0.026
Bio	20.43	0.121	**0.082**	**0.027**
Plain Dialogue	24.00	0.115	0.074	**0.027**
MBTI + Gender	**12.46**	**0.135**	**0.083**	0.025
MBTI + Bio	26.48	0.128	**0.083**	**0.026**
Gender + Bio	22.50	0.126	0.081	**0.026**
MBTI + Gender + Bio	28.96	0.130	0.083	0.026

Table 4: GODEL results on PRODIGy test set (Inter-Character)

Config.	CPPL	Acc @40	Acc @10	Acc @1
Bio	58.95	**0.833**	**0.712**	**0.348**
Char ID	**55.25**	0.826	0.709	0.345
Gender	58.32	0.824	0.706	0.335
MBTI	58.32	0.825	0.706	0.346
Plain Dialogue	595.14	0.720	0.368	0.337
Char ID + Bio	**54.89**	**0.834**	**0.714**	**0.347**
Char ID + Gender	58.88	0.827	0.706	0.337
Char ID + MBTI	57.82	0.826	0.704	0.343
Gender + Bio	55.73	0.799	0.708	0.343
MBTI + Bio	55.95	0.799	0.708	0.344
MBTI + Gender	58.32	0.791	0.704	**0.347**
MBTI + Gender + Bio	57.08	**0.801**	**0.710**	0.339
Char ID + MBTI + Bio	**53.23**	0.800	0.710	0.340
Char ID + MBTI + Gender	55.48	0.791	0.705	**0.344**
Char ID + MBTI + Gender + Bio	54.99	0.802	0.710	0.341

Table 5: DialoGPT results on PRODIGy test set (Intra-Character)

Train → Test	Config.	CPPL	Acc @40	Acc @10	Acc @1
PRODIGy → PC	Plain Dialogue	891.80	0.518	0.444	0.184
	Bio_{par}	**219.07**	**0.612**	**0.533**	**0.200**
PC → PRODIGy	Plain Dialogue	$1.32e$+05	**0.433**	**0.333**	**0.139**
	Bio	$3.27e$**+04**	0.386	0.309	0.119

Table 6: DialoGPT results on cross-domain experiments: training on PRODIGy and test on Persona-Chat (PRODIGy → PC) and vice-versa (PC → PRODIGy).

with the same methodology on Persona-Chat and tested on the PRODIGy test set. Results are reported in Table 6. As can be seen, adding the biography sentences drastically improves the $CPPL$ results also in zero-shot settings (both trained on PRODIGy and tested on Persona-Chat, and vice-versa). Interestingly, we also observe that using a general biography, as the one we propose, yields better generalisation capabilities in comparison to a dialogue-specific persona as in Zhang et al. (2018). In particular, the $CPPL$ performance degradation of the Persona-Chat based models in cross-domain experiments is much worse than in PRODIGy-based models (219.07 of our Bio_{par} model tested on Persona-Chat vs 3.27e+04 of the Persona-Chat Bio model tested on PRODIGy).

As regards the models trained on PRODIGy and tested on Persona-Chat, the results are in line with the in-domain experiments: Bio_{par} significantly outperforms Plain Dialogue for both $CPPL$ and Acc $@N$. On the contrary, in the scenario in which we trained the models on Persona-Chat and tested on PRODIGy, the Bio model's Acc $@N$ scores are lower than Plain Dialogue's scores. This might suggest that persona sentences do not capture personas' complex characteristics, therefore they might be less effective to generalise in a cross-domain scenario.

7 Human Evaluation Results

In this section, we provide a detailed description of the human evaluation experiments. During the analysis of the results, we consider (i) all dialogues (regardless of context length), (ii) dialogues with ≤ 6 turns, and (iii) dialogues with > 6 turns.

Table 7 presents the evaluators' average rankings. The scores are inverted for readability purposes: higher scores indicate better performances. The significant gap between the scores of gold and the generated responses indicates that there is wide room for improvement for our models. Among the models, Plain Dialogue receives the highest ratings, closely followed by the other models. In shorter contexts, profile-based models, i.e., Bio_{par}, MBTI, Gender, yield higher scores than in longer context: this suggests a decrease of the relevance of profile information as the information in the context increases. Furthermore, when the profile information is explicitly provided to evaluators, the gap between scores in shorter and longer dialogues diminishes. This suggests a positive impact of profile information on evaluators' judgements, who perceive responses generated by profile-based models as more appropriate.

Table 8 shows the percentages of times the profile-based models are preferred to Plain Dialogue. We observe that when evaluators are informed of the target speaker's profile, there is an increase in the preference for profile-based model responses over Plain Dialogue responses. Also in this case, the gap observed between responses given shorter contexts and those given longer contexts becomes less prominent.

Plain Dialogue might be preferred over profile-based models due to its generation of generic responses easily fitting into various dialogues. However, it's worth noting that each profile-based

model learns unique patterns from the profile information during training, resulting in responses tailored to individual speakers. The example in Table 9 illustrates this phenomenon. Plain Dialogue's response is a fairly generic answer that fits the context of the dialogue well. However, we can notice that each profile-based model's generation reflected the a speaker's profile information. Bio_{par}'s output closely aligns with the Gold response concept. Given the character's biography indicating a need for psychiatric help, the model inferred a potential mental distress, responding with "*I see a skeleton.*". The MBTI response aligns with the introverted trait of the character, who is reluctant to answer the interlocutor: "*I'm sure you can tell me.*". The Gender model's response incorporates stereotypical male patterns (e.g. the use of the swear word "*shit*") which are common in the Cornell Movie Dialogs corpus (Schofield and Mehr, 2016).

These findings are consistent with a post-hoc qualitative interview with the evaluators. In fact, they reported a preference for generic answers (usually produced by Plain Dialogue) due to their wider applicability. Moreover, they noted instances where responses were coherent with profile information but not with dialogue contexts: this had a detrimental effect on the perceived quality of the answer. However, when responses were consistent with both profile and dialogue they were clearly favoured. Finally, despite a natural preference for gold responses, evaluators reported to not having recognised them, and more broadly to having seen only few of the movies whose dialogues were evaluated.

8 Conclusion

In this paper, we studied several strategies for the profile-based dialogue generation task. To this end, we built PRODIGy, a new dataset of movie dialogues aligned with characters' profile information, i.e. personality type, gender, biography, and a collection of their dialogues, useful for inferring their vocabulary and language style. This new resource served as the basis for analysing the impact of profile representations in the generation process. To this end, we tested both with fine-tuning and instruction-based LMs. Moreover, to account for different privacy levels, we used both inter-character partitions (characters in the test set are not in the training set) and intra-character par-

	All turns		≤ 6 turns		> 6 turns	
Response	**No profile**	**With profile**	**No profile**	**With profile**	**No profile**	**With profile**
Gold	4.04	3.97	3.90	3.85	4.32	4.18
Plain Dialogue	**2.90**	**2.86**	**2.89**	**2.89**	**2.92**	**2.80**
Bio_{par}	2.66	2.71	2.70	2.77	2.58	2.60
MBTI	2.67	2.75	2.77	2.77	2.49	2.70
Gender	2.73	2.71	2.75	2.71	2.69	2.72

Table 7: Human Evaluation Average Rankings: higher scores indicate better performance.

	All turns		≤ 6 turns		> 6 turns	
Response	**No profile**	**With profile**	**No profile**	**With profile**	**No profile**	**With profile**
Bio_{par}	43.14	47.60	44.30	47.85	40.95	47.14
MBTI	44.96	**49.59**	**46.33**	**50.38**	42.38	48.10
Gender	**45.36**	44.04	46.19	43.91	**44.29**	**49.52**

Table 8: Preference Percentages: responses of Profile-based Models vs. Plain Dialogue Responses

Profile	Dialogue Context	Responses
A's Biography: • I am a janitor at the Massachusetts Institute of Technology. • I live in a rough, working class Irish neighborhood in south Boston. • I am seeing a Harvard girl named Skylar. • I am cocky, brash, and combative: while I am often crude, my quick wits mark me as unmistakably intelligent. • I have to put my mathematical genius to work by helping professor Gerald Lambeau and start seeing a psychiatrist. **A's MBTI**: Introvert, Intuititive, Thinker, Perceiver **A's Gender**: Male	**B**: Okay, you're in your bed, Will. Now how old are you? **A**: Seven. **B**: And what do you see? **A**: Somethin's in my room. **B**: What is it?	[Gold] **A**: It's like a small figure, hoverin' over me. Gettin' closer. [Plain Dialogue] **A**: I don't know. It just looks like a black couch. It's too bright. [Bio_{par}] **A**: I see a skeleton. [MBTI] **A**: I'm sure you can tell me. [Gender] **A**: It's a door, I don't know which one but I want to find out for sure and put an end to this shit before it gets any worse.

Table 9: Example of Inter-Character models' generations.

titions (the same characters are both in the training and test sets). In former, the LM does not retain personal information, while in latter, character profile information is stored in the LM. Experiments showed that using profile information brings better results in all cases and with both LM approaches (fine-tuning and instruction prompting configurations). The cross-domain experiments showed that PRODIGy-based models exhibit better generalisation than those trained on similar resources. The human evaluation results indicate that generic responses are often preferred for their broader applicability. However, responses consistent with both profile and dialogue were favoured, but mismatches lowered perceived quality.

Limitations

The fact that PRODIGy includes fictional characters could imply that the roles may be stereotyped. The high predictivity of the model trained on characters' gender is a potential indicator of this hypothesis. Thus, while PRODIGy allows avoiding a number of privacy issues, it may be less realistic. However, this problem may be present in other datasets, such as Persona-Chat, where users were simulated. Moreover, as regards to Gender, PRODIGy is limited to a binary classification since it is the one originally provided by the Cornell Movie Dialogs Corpus. Finally, the human evaluation shows a strong preference for gold responses, suggesting significant room for improvement, which we plan to address in future work.

Ethics Statement

One of the potential risks of profile-based dialogue systems is that they need to collect users' information, thus creating the risk of such private data being misused or leaked (Krishnamurthy et al.; Corrigan et al., 2014). The two configurations (i.e. inter-character and intra-character) we propose in this paper have been implemented in light of this. Being able to understand the impact of each of the profile dimensions within a dialogue system can be useful to determine which are the sensitive data necessary to develop a dialogue system and which could be left out in order to preserve the users' privacy (Dudy et al., 2021). Another problem is the possible fully automated use of profile-based models. Such systems, if left to act completely autonomously, may make erroneous assumptions, even in imitating a given user, thus returning possibly misleading answers.

References

Gordon Willard Allport. 1937. Personality: A psychological interpretation.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002, Dublin, Ireland. Association for Computational Linguistics.

Hope B Corrigan, Georgiana Craciun, and Allison M Powell. 2014. How does target know so much about its customers? utilizing customer analytics to make marketing decisions. *Marketing Education Review*, 24(2):159–166.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alastair Gill, Carsten Brockmann, and Jon Oberlander. 2012. Perceptions of alignment and personality in generated dialogue. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 40–48, Utica, IL. Association for Computational Linguistics.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 34, pages 13821–13822.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology*.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165, Dublin, Ireland. Association for Computational Linguistics.

Filip Klubicka and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *4REAL 2018 Workshop on Replicability and Reproducibility of Research Results in Science and Technology of Language*, page 16.

Balachander Krishnamurthy, Konstantin Naryshkin, and Craig Wills. Privacy leakage vs. protection measures: the growing disconnect.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic. Association for Computational Linguistics.

François Mairesse and Marilyn A Walker. 2008. A personality-based framework for utterance generation in dialogue applications.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 585–592, Online. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–

9206, Online. Association for Computational Linguistics.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Isabel Briggs Myers. 1962. The myers-briggs type indicator: Manual (1962).

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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*.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2470–2477.

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.

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*.

Silvia Schiaffino and Analía Amandi. 2009. Intelligent user profiling. In *Artificial intelligence an international perspective*, pages 193–216. Springer.

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, San Diego, California, USA. Association for Computational Linguistics.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Thomas Scialom, Serra Sinem Tekiroğlu, Jacopo Staiano, and Marco Guerini. 2020. Toward stance-based personas for opinionated dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2625–2635, Online. Association for Computational Linguistics.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. 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 167–177, Online. Association for Computational Linguistics.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188*.

Hsuan Su, Jiun-Hao Jhan, Fan-yun Sun, Saurav Sahay, and Hung-yi Lee. 2021. Put chatbot into its interlocutor's shoes: New framework to learn chatbot responding with intention. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1559–1569, Online. Association for Computational Linguistics.

Jari Vesanen. 2007. What is personalization? a conceptual framework. *European Journal of Marketing*.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Yury Zemlyanskiy and Fei Sha. 2018. Aiming to know you better perhaps makes me a more engaging dialogue partner. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 551–561, Brussels, Belgium. 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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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*.

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 As-*

sociation for Computational Linguistics: System Demonstrations, pages 270–278, Online. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

A Personality Information

To annotate the characters in the Cornell Movie Dialogs Corpus, we selected the ones appearing in at least 50 dialogues. Then, we used the query `movie_title+year` to extract from PDB the metadata related to each movie, containing the list of the characters' names and IDs. If the character was present in the metadata, we used the query `PDB_characterID` to extract the MBTI type and related votes. If the MBTI type had at least 5 votes, the character was annotated. If the character was not in the metadata, a human annotator performed a manual check within PDB to verify if there was an actual match. In case the mismatch could be manually resolved, we replicated the above procedure to annotate the character. (See Algorithm 1 for details.)

Algorithm 1: MBTI Annotation

---
for character **in** CMD characters **do**
  if nr_dialogues $\geq$ 50 **then**
    PDB_query (movie_title + year) $\rightarrow$
     movie_metadata
    if movie_metadata found **then**
      if character in movie_metadata **then**
        PDB_query (PDB_character_id) $\rightarrow$
         character_metadata
        if character_metadata found **then**
          extract MBTI type and n_votes
          if n_votes $\geq$ 5 **then**
            ⌊annotate character
      else
        manual_check in PDB $\rightarrow$
         character_metadata
        if character_metadata found **then**
          extract MBTI type and n_votes
          if n_votes $\geq$ 5 **then**
            ⌊annotate character

B Biographical Information

To obtain biography information, we scraped characters' pages from Charactour.com, Fandom.com and Wikipedia.org. Then, to automatically extract relevant sentences, we compared four extractive summarisation strategies and chose an algorithm based on Kullback-Leibler distance (Haghighi and Vanderwende, 2009). A human annotator modified the extracted sentences according to our guidelines. First, the most relevant sentences were re-ranked in the order of importance. Then, they

were post-edited as follows: changed from the third to the first person singular and shortened if excessively long (so to mimic Persona-Chat style Zhang et al. (2018)), or enriched with missing relevant information. If a character biography was not found, the annotator wrote it from scratch reading the movie plot. To increase the number of biography sentences, ChatGPT was given the original sentences and asked to produce two paraphrases. These new sentences were given to the annotator for post-editing to correct errors or further paraphrase those still too similar to the original biographies. Finally, we obtained a total of 8498 biography sentences. (See Algorithm 2 for details.)

Algorithm 2: Biographies Scraping, Revision and Enrichment

---
for character **in** annotated_characters **do**
  scrape bio from sources
  if bio exists **then**
    KL_{based}(bio) $\rightarrow$ bio_sents
    human_revision(bio_sents) $\rightarrow$
     bio_sents$_{revised}$
  else
    ⌊bio_sents written from scratch
  LLM(bio_sents$_{revised}$) $\rightarrow$ (sents$_{par}$ 1,
   sents$_{par}$ 2)
  human_revision(sents$_{par}$ 1, sents$_{par}$ 2) $\rightarrow$
   (sents$_{par}$ 1, sents$_{par}$ 2)$_{revised}$