PRODIGy: a PROfile-based DIalogue Generation dataset

Anonymous ACL submission

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 introduce the PRODIGy (PROfile-based DIalogue Generation) dataset, which brings diverse representations together, providing a more comprehensive profile dimension set for each speaker. This resource comprises more than 20k dialogues, sourced from movie scripts, aligned with speaker representations such as communication style, biography, personality and gender. Initial experiments with diverse baselines show that providing generative language models with these aspects of a profile, both separately and jointly, enhances models' performance. This improvement holds true in both in-domain and cross-domain settings, for both fine-tuned and instruction-based LLMs.

1 Introduction

Dialogue agents capable of holding human-like interactions 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 focusing on solving specific tasks, open-domain dialogue systems aim to discuss various topics, possibly maintaining a consistent profile in their responses (Kann et al., 2022). In this work, we investigate the role of profile information in open-domain dialogue systems.

Despite recent advancements in conversational agents, due 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 often struggle to maintain coherence, resulting in inconsistent or uninformative responses. This issue adversely affects user engagement and trust (Li et al., 2016b, 2020). In this scenario, endowing dialogue systems with profile information is crucial for enhancing the models' ability to generate 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*, on the other hand, 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 conversations and includes elements such as background facts, language, and interaction style.

Several approaches have been explored to integrate persona information into dialogue generation (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, these methods are typically sporadic and disjointed, addressing only one persona dimension at a time, either through an *explicit* representation (a few simple, dialogue-specific sentences about the user) or an *implicit* representation (a collection of the user's previous dialogues) that is challenging to obtain. Consequently, these approaches fail to model the complex nature of human communication, which is influenced by the interaction of multiple aspects.

In this paper, we investigate the impact of diverse profile representations in the development of

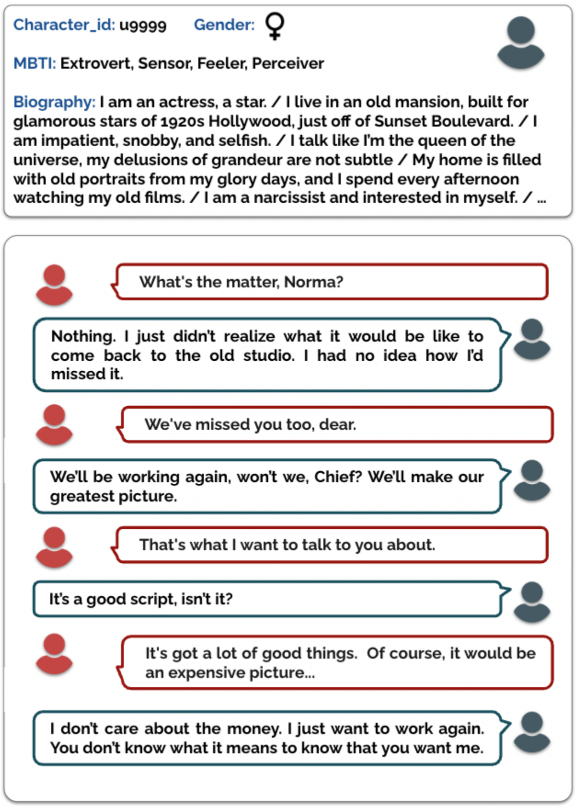

Figure 1: Example of a dialogue with diverse speaker's profile information provided.

dialogue systems by comparing and benchmarking them. To this end, we introduce a new dataset, named PRODIGy (PROfile-based DIalogue Generation)[1], that combines existing profile representations (i.e., language style, gender, personality) with novel and more complex representations of the persona, such as biographies. PRODIGy is 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 approach avoids privacy concerns related to employing real user data and simplifies 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. Figure 1 illustrates an example from PRODIGy, in which the dialogue is aligned with the target speaker's profile representation.

We validated PRODIGy by benchmarking it with diverse baselines. In particular, we employed either fine-tuning or instruction prompting, and tested a range of configurations varying the profile dimensions, both in-domain and cross-domain. Evaluation involved both automatic metrics and human assessment. As for automatic metrics, in-domain experiments show that fine-tuning LMs with diverse profile aspects significantly improves their predictive capabilities. Additionally, instructing non-fine-tuned LLMs with profile information also improves their performance. In cross-domain settings, PRODIGy-based models show better generalisation than those trained on other persona-based resources. In human evaluations, evaluators had a tendency of favouring generic responses for broader applicability. However, when responses were consistent with both profile and dialogue they were clearly preferred. Profile information proves beneficial especially in dialogues with limited context, and when disclosed to evaluators, profile-based responses are deemed more appropriate.

[1]The dataset will be distributed for research purposes at the following link: [URL].

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 Our communication style is closely related to social status, gender, and motivations, and offers insights into 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 may encounter challenges related to ephemerality (Klubicka and Fernández, 2018); 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 users' dialogue history (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 not sufficient to train open-domain dialogue models. One of the most widely used persona-based datasets is Persona-Chat (Zhang et al., 2018), collected in a controlled crowd-sourcing environment. However, it provides a generic fact-based persona representation (e.g. *"I just got my nails done"*) specific to single dialogues and leaving 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 using users' past dialogues to represent the persona (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. Various 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.

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 Corpus 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, facilitates 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.

Below, we outline the profile representations and detail the methodology employed to annotate the characters within the dataset.

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)[2] 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.

To annotate the characters in the Cornell Movie Dialogs Corpus, 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. Details of the alignment procedure are provided in Appendix A.1.

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

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, for annotation reliability, 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 annotated with the personality information, from Charactour.com, Fandom.com and Wikipedia. Then, to automatically extract the most relevant sentences, we employed an extractive summarisation algorithm based on Kullback-Leibler distance (Haghighi and Vanderwende, 2009). Subsequently, a human-machine collaboration procedure followed, where a human annotator[3] modified the extracted sentences to ensure that our resource maintains qualitative and quantitative alignment with the Persona-Chat dataset (Zhang et al., 2018) for comparability purposes. To achieve this, specific guidelines were formulated and provided to the annotator:

- Re-rank the top 10 sentences in order of importance, according to the speaker's profile.
- Convert the sentences from the third to the first person singular.
- Shorten excessively long sentences.

[3]The human annotator was one of the authors and a Computer Science PhD student.

Category	Statistics
Dialogues	20850
Turns	80604
Annotated Characters	339
Biography Sentences	8498
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.

To further enhance the quality of the biographies, two additional instructions were provided:

- Enrich the sentences by incorporating any missing relevant information.
- If a character biography is not available, create one by reading the movie plot.

While PRODIGy biography sentences align stylistically with Persona-Chat (Zhang et al., 2018), they are not limited to generic facts and capture more complex aspects of the persona, making them conceptually and qualitatively different from Persona-Chat.

To increase the number and the variability of biography sentences, ChatGPT (OpenAI, 2022) 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. More details about the biographical information procedure are provided in Appendix A.2.

As a result of the aforementioned procedures, we obtained a dataset with more than 20K dialogues for 80K turns with 300 annotated characters and more than 8k biography sentences. The dialogues are aligned with the following dimensions of one of the speakers: gender, personality type, character's biography, and linguistic style modelled by character's dialogues. Character biographies consist of an average of 8 sentences, ranging from 5 to 10 sentences, with an average of 13 tokens per sentence. Each biography sentence has been paraphrased twice. Detailed statistics of the PRODIGy dataset are provided in Table 1.

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, the personality information, or both. Our aim is to analyse the impact of each representation on 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. We investigated several fine-tuning 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. In Appendix B, we provide details on the fine-tuning setup and input syntax utilised for DialoGPT.

Besides DialoGPT, we also experimented with GODEL (Peng et al., 2022), an instruction-based LLM specific for dialogue generation. Our aim is to assess the effect of providing profile information as an instruction to a non-fine-tuned LLM. The input syntax for GODEL is shown in Appendix C.

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

Plain Dialogue Driven Generation In the first configuration, we fine-tuned DialoGPT and instructed 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 to fine-tune DialoGPT and prompt GODEL, as 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, aggregating characters' dialogue lists using their IDs to 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 provided DialoGPT and GODEL with persona information 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 Zhang et al. (2018). These two configurations also address privacy concerns: in one case, the LM does not retain any personal information but uses it only at inference time, while in the second, the LM stores the information about the user in its internal representation.

5 Automatic Evaluation

In this section, we describe the metrics and experiments for the validation of our resource.

5.1 Metrics

We assess model performances using two 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. $CPPL$ (Equation 1) computes 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.

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 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 settings, we consider the target speaker's profile, excluding the interlocutor' profile. Given just the dialogue context, or both context and profile information, we aim to predict the target speaker's final turn.

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, consistently with the experiments by Welch et al. (2022). We opted for the Bio_{par} model as our biography-based model. This model is trained by randomly selecting five sentences[4] per dialogue from the original biography or its paraphrases. The decision to use this model is based on its demonstrated superior effectiveness, as shown in a preliminary experiment (outlined in Appendix D) focusing on biography-based models.

Table 2 presents model performances based on profile information. In terms of Acc@N, these models outperform Plain Dialogue that lacks profile information. Single-profile models show similar Acc@10 performances. Also, combining multiple profile dimensions, the Acc@N scores do not differ significantly. Regarding $CPPL$, Plain Dialogue performs the worst, while models with profile information excel. Notably, Gender attains the best $CPPL$ (87.92), comparable to MBTI. Bio_{par} performs worse than Gender and MBTI but significantly outperforms the baseline with a score of 98.27, showcasing the efficacy of high-level character descriptions. Gender's strong performance in $CPPL$ and Acc@N may stem from the gender-specific linguistic patterns in PRODIGy's dialogues sourced from the Cornell Movie Dialogs Corpus (Schofield and Mehr, 2016), enabling the model to effectively incorporate such characteristics. Overall, the results demonstrate that adding profile information, either alone or jointly, strongly improves the models performance in terms of generalisation.

In Table 3 we report the results obtained by prompting GODEL with the profile information. The $CPPL$ and Acc@N values reveal better performances even when profile information is merely provided as an instruction. In particular, Plain Dialogue exhibits a worst $CPPL$ compared to MBTI and MBTI + Gender (24.00 vs 12.46). Also in terms of 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 is important to state that, while GODEL may seem to outperform DialoGPT in terms of $CPPL$, a direct comparison between their metrics is not possible as these models are pre-trained on distinct datasets and have a different vocabulary size.

Config.	CPPL	Acc@10	Acc@1
MBTI	89.30	**0.665**	**0.317**
⚥	**87.92**	0.664	0.306
Bio_{par}	98.27	0.661	0.307
PD	541.16	0.585	0.298
MBTI+⚥	**91.50**	**0.660**	**0.311**
⚥+Bio_{par}	96.31	0.658	0.299
MBTI+Bio_{par}	100.35	0.653	0.296
MBTI+⚥+Bio_{par}	91.65	0.660	0.302

Table 2: DialoGPT results on PRODIGy test set (Inter-Character). PD and ⚥represent Plain Dialogue and Gender, respectively.

Config.	CPPL	Acc@10	Acc@1
MBTI	**12.46**	0.080	0.026
⚥	13.65	0.075	0.026
Bio	20.43	**0.082**	**0.027**
PD	24.00	0.074	**0.027**
MBTI + ⚥	**12.46**	**0.083**	0.025
MBTI + Bio	26.48	**0.083**	**0.026**
⚥+ Bio	22.50	0.081	**0.026**
MBTI + ⚥+ Bio	28.96	0.083	0.026

Table 3: GODEL results on PRODIGy test set (Inter-Character). PD and ⚥represent Plain Dialogue and Gender, respectively.

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

[4] 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.

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 4, endowing the model with the dialogical information (`ID`) provides the best results in terms of $CPPL$. This is attributed to the model learning the character's vocabulary and language style during training, enhancing predictions. In terms of Acc@N, the best performing model is Bio (0.712 of Acc@10, and 0.348 of Acc@1). The other profile-based models exhibit similar performances. The Plain Dialogue model emerges as the weakest, proving again that fine-tuning models through profile information is beneficial. Combining biographical information and `ID` further enhances model efficiency in terms of $CPPL$, with better values when a high-level character description is included. The scores in Acc@N show that, when combined with the dialogical information (`ID`), the biographical information improves the predictive ability of the model more than Gender and MBTI. Although `ID` excels in $CPPL$, models with explicit profile information show comparable efficiency. Regarding the models trained with profile information jointly, the best performances are achieved by those trained with the characters' biographical information. 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.

Config.	CPPL	Acc@10	Acc@1
Bio	58.95	**0.712**	**0.348**
`ID`	**55.25**	0.709	0.345
⚥	58.32	0.706	0.335
`MBTI`	58.32	0.706	0.346
PD	595.14	0.368	0.337
`ID`+Bio	**54.89**	**0.714**	**0.347**
`ID`+⚥	58.88	0.706	0.337
`ID`+`MBTI`	57.82	0.704	0.343
⚥+Bio	55.73	0.708	0.343
`MBTI`+Bio	55.95	0.708	0.344
`MBTI`+⚥	58.32	0.704	**0.347**
`MBTI`+⚥+Bio	57.08	**0.710**	0.339
`ID`+`MBTI`+Bio	**53.23**	0.710	0.340
`ID`+`MBTI`+⚥	55.48	0.705	**0.344**
`ID`+`MBTI`+⚥+Bio	54.99	0.710	0.341

Table 4: DialoGPT results on PRODIGy test set (Intra-Character). PD and ⚥represent Plain Dialogue and Gender, respectively.

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 with the same methodology on Persona-Chat and tested on the PRODIGy test set. The results, presented in Table 5, show a significant improvement in $CPPL$ scores when incorporating biography sentences, even in zero-shot settings (both trained on PRODIGy and tested on Persona-Chat, and vice-versa). Interestingly, using a general biography, as the one we propose, yields better generalisation capabilities than a dialogue-specific persona as in Zhang et al. (2018). When models trained on PRODIGy are tested on Persona-Chat, the results are in line with the in-domain experiments: Bio$_{par}$ consistently outperforms Plain Dialogue in 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.

Train → Test	Config.	CPPL	Acc@10	Acc@1
PROD. → PC	PD	891.80	0.444	0.184
	Bio$_{par}$	**219.07**	**0.533**	**0.200**
PC → PROD.	PD	1.32e+05	**0.333**	**0.139**
	Bio	**3.27e+04**	0.309	0.119

Table 5: DialoGPT results on cross-domain experiments: fine-tuning on PRODIGy and test on Persona-Chat (PROD. → PC) and vice-versa (PC → PROD.). PD represents Plain Dialogue.

6 Human Evaluation

Besides the automatic evaluation, we also run an human evaluation study to validate PRODIGy.

This evaluation involved six subjects, comprising four PhD students in Computer Science and two MSc students in Data Science. Evaluators received 100 dialogues each, 50 with profile information disclosed and 50 without profile disclosure, so to enable an assessment of profile information's impact on judgements. We focused on output generated using top-p decoding by four models trained during inter-character experiments: the model trained on dialogues only and the models trained with one profile dimension. Evaluators ranked five possible responses for each dialogue, including the gold response used as a control condition, on a scale from 1 (most likely) to 5 (least likely) based on perceived likelihood of being the target speaker's response. In total, we collected 3000 evaluations. Subsequently, we conducted post-hoc qualitative interviews with the evaluators.

Response	All turns		$\leq$ 6 turns		$>$ 6 turns	
	No	**With**	**No**	**With**	**No**	**With**
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
⚥	45.36	44.04	46.19	43.91	44.29	**49.52**

Table 6: Preference Percentages: responses of profile-based Models vs. Plain Dialogue Responses. ⚥represents Gender. No/With indicates profile information disclosure to evaluators.

6.1 Results

Analysing the human evaluation results, we consider (i) all dialogues, (ii) dialogues with $\leq$ 6 turns, and (iii) dialogues with $>$ 6 turns.

The human evaluation reveals that the gold responses are preferred by far over the generated responses, indicating clear room for future improvement over the baselines we employed. Notably, Plain Dialogue was the favored model, with only marginal rating differences compared to other models. From the post-hoc interviews it emerged that Plain Dialogue's ability to produce generic responses that easily fit into various dialogues was often the reason for this preference. However, an interesting shift occurs when evaluators are made aware of the speaker's profile. In such cases, there is a noticeable increase in the preference for profile-based model responses over Plain Dialogue responses. This shift is shown in Table 6, which outlines the percentages of times evaluators favored profile-based models over Plain Dialogue. This trend can be attributed to a clear preference towards generations that exhibit coherence with both profile information and dialogue context, emphasising the significance of the profile in the generation process. Finally, profile-based models receive more favorable evaluations in shorter contexts, suggesting that the inclusion of profile information is advantageous when the dialogue context provides limited information about the speaker.

As stated, these findings are consistent with the feedbacks from evaluators that we gathered in a post-hoc interview. Evaluators expressed a preference for generic answers, typically generated by Plain Dialogue, due to their broader applicability. This was particular evident for those cases where responses generated by profile-based models matched the profile information of the speaker but not dialogue context, thus negatively impacting perceived answer quality. However, when profile information was provided to evaluators, the preference for responses consistent with both profile and dialogue clearly emerged. At a closer inspection of such cases we found that these sentences, consistent with both profile and dialogue, were often preferred even to gold responses. Conversely, the overarching inclination for gold responses was not given because they were familiar to evaluators: they reported not recognising them, and more broadly to having seen only few of the movies whose dialogues were evaluated. See Appendix E for additional details and discussion on the human evaluation and Table 9 for some generation examples.

7 Conclusion

In this paper we introduced PRODIGy, a new dataset of movie dialogues aligned with characters' profile information, i.e. personality type, gender, biography, and a collection of speakers' dialogues, useful for inferring their vocabulary and language style. Derived from movie scripts, PRODIGy also mitigates privacy concerns associated with real user data. To validate this resource, we conducted several experiments using diverse baselines, both via fine-tuning and instruction prompting. Results indicate that including profile information in both approaches improved models' performance. Moreover, the cross-domain experiments showed that PRODIGy-based models exhibit better generalisation than those trained on similar resources. Results from the human evaluation showed that, despite a preference for generic responses due to their broader applicability, responses consistent with both profile and dialogue are clearly favoured. Moreover, the results highlight the value of incorporating profile information, especially when speaker's information provided within the dialogue context is limited.

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.

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.

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*.

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 Association 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 Annotation of PRODIGy Characters Algoritms

A.1 Annotation with Personality Information

Algorithm 1 outlines the annotation process to assign MBTI personality types to the Cornell Movie Dialogue Corpus (CMD). We selected only CMD characters appearing in at least 50 dialogues. For each character we used the query `movie_title+year` to extract from the Personality Database (PDB) the related movie metadata, containing the list of the movie characters' names and IDs. If the character was present in the movie metadata, we used a query `PDB_characterID` to extract the MBTI type and votes. If the MBTI type has at least 5 votes, the character was annotated. If the character was not found in the movie metadata, a manual check within PDB for character metadata is performed. In case the mismatch could be manually resolved, we replicated the above procedure to annotate the character.

Algorithm 1: MBTI Annotation

```
for character in CMD characters do
    if nr_dialogues ≥ 50 then
        PDB_query (movie_title + year) →
          movie_metadata
        if movie_metadata found then
            if character in movie_metadata then
                PDB_query (PDB_character_id) →
                  character_metadata
                if character_metadata found then
                    extract MBTI type and n_votes
                    if n_votes ≥ 5 then
                        annotate character
            else
                manual_check in PDB →
                  character_metadata
                if character_metadata found then
                    extract MBTI type and n_votes
                    if n_votes ≥ 5 then
                        annotate character
```

A.2 Annotation with Biographical Information

Algorithm 2 describes the process for scraping, revising, and enriching biographies of annotated characters. For each character annotated with MBTI, a biography was scraped from external sources. If a biography was successfully retrieved, an extractive summarisation algorithm based on Kullback-Leibler divergence (Haghighi and Vanderwende, 2009) (KL_{based}) was applied to extract the most relevant biography sentences and human revision was applied to the sentences. If no biography was found during the scraping process, the human annotator created a new biography from scratch. Next, an LLM (i.e. ChatGPT) was given the post-edited biography sentences and asked to generate two sets of paraphrased sentences ($sents_{par}$ 1 and $sents_{par}$ 2). Finally, human revision was again applied to the generated sentence sets ($sents_{par}$ 1 and $sents_{par}$ 2), producing the final enriched and revised version of the character's biography.

Algorithm 2: Biographies Scraping, Revision and Enrichment

```
for character in annotated_characters do
    scrape bio from sources
    if bio exists then
        KL_based(bio) → bio_sents
        human_revision(bio_sents) → bio_sents_revised
    else
        bio_sents written from scratch
    LLM(bio_sents_revised) → (sents_par 1, sents_par 2)
    human_revision(sents_par 1, sents_par 2) →
      (sents_par 1, sents_par 2)_revised
```

B DialoGPT Fine-tuning Details

In this section we report the details of the fine-tuning of each model employed during both inter-character and intra-character experiments and the input syntax.

B.1 Fine-tuning Setup

To investigate the impact of individual profile dimensions, we opted to employ DialoGPT medium for all fine-tuning experiments. To maintain consistency across our trials, we kept the hyperparameters constant throughout the fine-tuning process, 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. The fine-tuning was performed on a single Tesla V100 GPU.

B.2 Input Syntax

When fine-tuning DialoGPT, we concatenated the characters' profile information to the corresponding turns of the dialogues. The input syntax employed in the experiments conducted with DialoGPT is delineated 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 fine-tuning, each part of the profile input and its corresponding token are added or removed depending on the configuration under inspection.

C GODEL Prompt Syntax

During the experiments with GODEL, we prompted the model with an instruction and a context including the profile information and the dialogue context, respectively. We tasked GODEL to predict the last turn in the dialogue. Following, we provide an example of the input syntax.

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*

D Biography-based Models experiment

In order to understand what is the best strategy to input biographies to inter-character models, we conducted a preliminary experiment. In particular, we tested three strategies to add variability to the biographies during fine-tuning: (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.

Table 7 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 $CPPL$. Fine-tuning the models by mixing original and paraphrased biographies, thus increasing lexical variability, improves the performance even further in terms of both $CPPL$ (98.27 for Bio$_{par}$ vs. 117.26 for Bio) and Acc@N (e.g. for Acc@10, 0.661 for Bio$_{par}$ vs. 0.647 for Bio). Thus, in the inter-character experiments with DialoGPT, we will always use Bio$_{par}$ as the reference configuration.

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

Table 7: DialoGPT results of the addition of variability to biography sentences on PRODIGy test set (Inter-Character)

E Analysis of Human Evaluation Rankings

Table 8 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 that profile information is beneficial when dialogue context does not provide sufficient information about the speaker. 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.

While Plain Dialogue may be favored for its generation of generic responses adaptable to various dialogues, it is worth noting that each profile-based model learns unique patterns from the profile information during training, resulting in responses

Response	All turns		≤ 6 turns		> 6 turns	
	No 👤	**With 👤**	**No 👤**	**With 👤**	**No 👤**	**With 👤**
Gold	**4.04**	**3.97**	**3.90**	**3.85**	**4.32**	**4.18**
PD	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
⚥	2.73	2.71	2.75	2.71	2.69	2.72

Table 8: Human Evaluation Average Rankings: higher scores indicate better performance. PD and ⚥represent Plain Dialogue and Gender, respectively. No/With 👤 indicates profile information disclosure to evaluators.

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).

<table>
<tr><th>Input</th><th>Responses</th></tr>
<tr><td>
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.
<hr>
A's <code>MBTI</code>: Introvert, Intuitive, Thinker, Perceiver
<hr>
A's ⚥: Male
<hr>
Dialogue Context:
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?
</td><td>
• [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.
• [<code>MBTI</code>] A: I'm sure you can tell me.
• [⚥] 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.
</td></tr>
</table>

Table 9: Example of Inter-Character models' generations. PD and ⚥ represent Plain Dialogue and Gender, respectively.