When Harry Meets Superman: The Role of The Interlocutor in Persona-Based Dialogue Generation

Daniela Occhipinti[1,2], **Marco Guerini**[1], **Malvina Nissim**[3]
[1]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
[2]University of Trento, Italy
[3]University of Groningen, Netherlands
dochipinti@fbk.eu, guerini@fbk.eu, m.nissim@rug.nl

Abstract

Endowing dialogue agents with persona information has proven to significantly improve the consistency and diversity of their generations. While much focus has been placed on aligning dialogues with provided personas, the adaptation to the *interlocutor*'s profile remains largely underexplored. In this work, we investigate three key aspects: (1) a model's ability to align responses with both the provided persona and the interlocutor's; (2) its robustness when dealing with familiar versus unfamiliar interlocutors and topics, and (3) the impact of additional fine-tuning on specific persona-based dialogues. We evaluate dialogues generated with diverse speaker pairings and topics, framing the evaluation as an author identification task and employing both LLM-as-a-judge and human evaluations. By systematically masking or disclosing information about the interlocutor, we assess its impact on dialogue generation. Results show that access to the interlocutor's persona improves the recognition of the target speaker, while masking it does the opposite. Although models generalise well across topics, they struggle with unfamiliar interlocutors. Finally, we found that in zero-shot settings, LLMs often copy biographical details, facilitating identification but trivialising the task.

1 Introduction

In recent years, large language models (LLMs) have proven effective in generating coherent and contextually appropriate responses. One of their applications is in role-play systems (Chen et al., 2024; Tseng et al., 2024), where they simulate conversations by adapting to a specified *target* persona in a given prompt (e.g., "You are *a young woman passionate about cinema"*). Indeed, research has amply focused on developing and studying models that adapt their speech patterns, style, and content to align to a target persona they have to represent (Lu et al., 2024; Wang et al., 2024b).

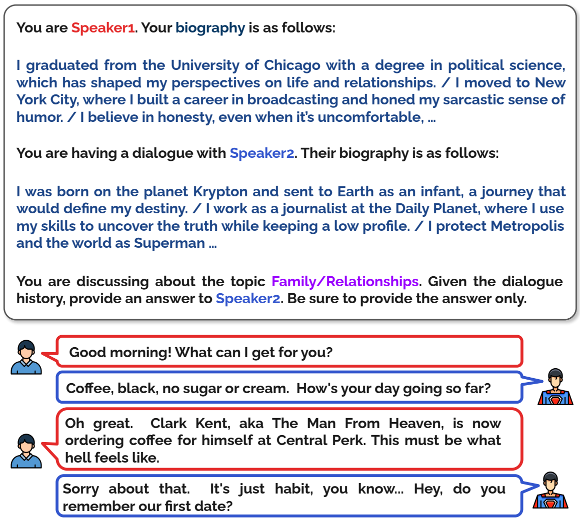

Figure 1: Example of a generated dialogue between Harry and Superman. The upper section displays the prompt, while the lower the beginning of the generated conversation.

One aspect that has received much less attention instead, is the ability of such persona-based agents to adapt to the other speaker they are engaging with (i.e., the *interlocutor*). Indeed, the impact on the model's responses of the interlocutor's persona, which plays a relevant role in yielding a more coherent and appropriate conversation, has been explored only in a few seminal works (Liu et al., 2020; Gu et al., 2021; Xu et al., 2022; Lu et al., 2022). While these works observe some degree of adjustment to the other speaker, the actual adaptation to interlocutor and topic variations is not systematically evaluated, especially regarding its effect on target speaker identification.

In the present research, we explore this aspect by addressing the following research questions:

Q1 Does the model adapt more efficiently to the target speaker's persona, or to the interlocutor's persona and dialogue turns?

Q2 To which aspects of the dialogue does the

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17964–17985
July 27 - August 1, 2025

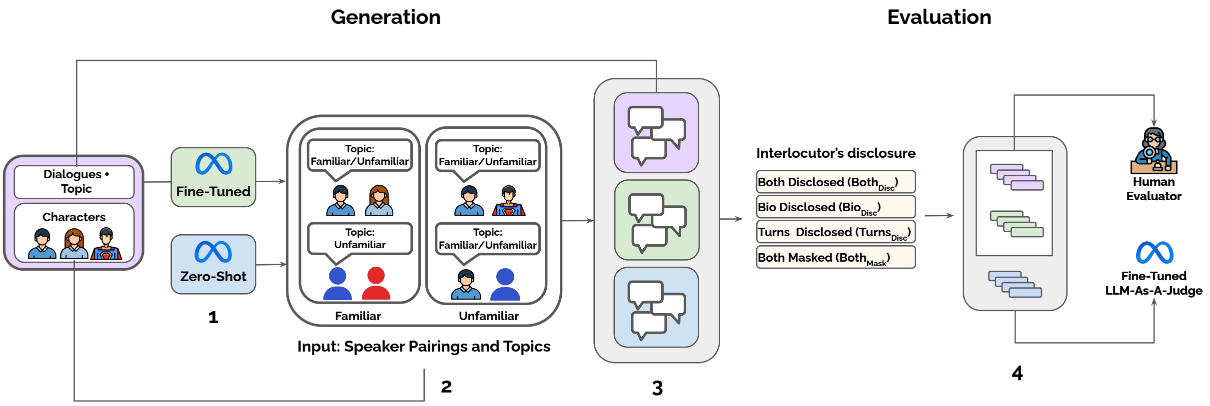

Figure 2: Pipeline: (i) Fine-tune Llama 3.1 8B Instruct on PRODIGy dialogues and speaker pairings. (ii) Create speaker pairings using PRODIGy and newly created Non-PRODIGy characters. (iii) Generate dialogues using both the fine-tuned model and a zero-shot setting providing the created pairings as input. (iv) Evaluate dialogues using LLM-as-a-Judge (for all model outputs and gold dialogues) and human reviewers (for fine-tuned model outputs and gold dialogues only), systematically masking or revealing interlocutor information.

model adapt more effectively? The aspects we consider are the following:

- *unfamiliar interlocutors* who fall outside the typical context of the target speakers (e.g., Harry from the movie *When Harry Met Sally* talking with Superman, as in Figure 1);
- *unfamiliar topics* to the target speaker (e.g., Harry from the movie *When Harry Met Sally* having a conversation about war).

Q3 Is additional fine-tuning on specific persona-based dialogues beneficial to the models' performance in preserving and adapting persona features?

To answer these questions, we generate dialogues with various speaker pairings and topics using both fine-tuned and zero-shot settings. We frame the task as author identification (Stamatatos et al., 2014), where given a dialogue and a pool of possible personas, an evaluator has to assign the correct persona to the target speaker. We introduce a novel evaluation paradigm specifically tailored to assess the influence of interlocutors and topics on dialogue generation and speaker identification. In this framework (Figure 2), we conduct extensive evaluations with human reviewers and an LLM-as-a-judge.

By systematically masking or revealing different aspects of the interlocutor's information, we measure how efficiently the LLM leverages the target speaker's information and adapts its responses to the interlocutor. The key idea is that if masking interlocutor information makes it harder to identify the target speaker's biography, it is likely that the model has appropriately adapted its responses to the interlocutor's information, while still staying coherent to the target persona. For instance, if the target speaker's persona is *a young woman passionate about cinema* and the interlocutor's is *her little daughter*, the model adjusts its language to fit a mother-child interaction rather than focusing on cinema. Masking the interlocutor's biography makes identifying the target speaker harder, as the conversation may not explicitly reference cinema.

Our framework, revolving around masking and revealing interlocutor information during evaluation, allows us to assess more directly the impact of the interlocutor on both dialogue generation and target speaker identification. Our findings show that models adapt easily to any topic given to them, while they struggle more to adapt to unfamiliar interlocutors, highlighting the importance of *who we are speaking with* over *what we are speaking about*. Moreover, while zero-shot models can adapt to both target and interlocutor speakers, they tend to capture only surface-level persona traits: we demonstrate that fine-tuning significantly improves their ability to grasp deeper persona characteristics.

2 Related Work

Role-Play Agents Despite their impressive abilities, LLMs often fall short in capturing the deep characteristics of human traits and interactions (Shanahan et al., 2023). To address this limitation, Role-Play LLMs have been developed to simulate assigned personas, capturing their distinctive traits. A key research line is the *character persona*, where models are provided with well-known fictional or real-world figures (Chen et al., 2024). Recent research enhanced the model's ability to consistently emulate specific personas by employing techniques such as memory reconstruction, long-term memory integration, and fine-grained persona construction (Shao et al., 2023; Park et al., 2023; Wang et al., 2024a; Lu et al., 2024). However, while significant progress has been made in simulating individual characters, few studies have investigated how interlocutor information influences response generation (Xu et al., 2022; Lu et al., 2022; Liu et al., 2020; Zhang et al., 2018). None of these studies has addressed how models adapt their responses across diverse speaker pairings and topics and do not directly evaluate how interlocutor information impacts dialogue generation by assessing its role in target speaker identification.

Role-Play Agents: Evaluation Evaluation for role-play models focuses on how well they replicate given characters, considering both superficial traits (linguistic style, knowledge) and deeper ones (identity, relationships, experiences) (Wang et al., 2024a; Li et al., 2023; Zhou et al., 2024; Yu et al., 2024). Early evaluation methods predominantly relied on overlap-based metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005), which compare generated text against reference responses. However, these metrics primarily capture surface-level elements and fail to assess deeper character traits. More recently, evaluator LLMs have been employed to score responses, evaluate quality or perform pairwise comparisons based on reference answers. For instance, Wang et al. (2024b) used GPT-3.5 to evaluate role-playing agents across multiple dimensions, including memorization, values, hallucination, and stability. Similarly, Lu et al. (2024) and Wu et al. (2024) employed GPT-4 (OpenAI, 2023) to assess role consistency, knowledge accuracy, and alignment with internal thought processes. Beyond automated methods, human evaluation plays is crucial to assess both the superficial and deeper aspects. However, it demands significant effort, lacks reproducibility, and is less scalable than automated methods.

3 Experimental Design

In this section, we describe the profile-based dialogue dataset we used and the methodology for assigning topics to each dialogue. Next, we introduce the models and the fine-tuning configuration we used for dialogue generation. Finally, we detail both automatic and human evaluation strategies.

3.1 Dataset

We employ the PRODIGy dataset (Occhipinti et al., 2024), comprising dialogues and characters from movies. The dataset includes comprehensive speaker profiles, including communication styles, biographies, personalities, and gender information. Each dialogue involves two speakers, with at least one speaker annotated with a profile.

We selected 5,660 dialogues where both speakers have profile annotations. This subset allows us to explore the influence of both target speaker and interlocutor's information on dialogue generation, with a focus on assessing the interlocutor's role. Among the various profile dimensions provided in PRODIGy, we focus on character biographies, as they offer rich semantic content, making them more relevant to our analysis.

By employing Llama 3.1 8B Instruct (Dubey et al., 2024), we annotated dialogues with topic labels to guide generation and evaluate the models' ability to generalise across different themes (see Appendix A for details). The dataset was then split into training, validation, and test sets (80:10:10), ensuring the same speaker pairs did not appear across different sets, though individual speakers might recur.

3.2 Models and Generations

We conducted experiments with Llama 3.1 8B Instruct (Dubey et al., 2024) in two settings: fine-tuning with PRODIGy and zero-shot prompting (see Appendix B for further details). Focusing on a single model allows us to attribute any observed differences to our fine-tuning methodology rather than to differences in model architecture or size, which was a key consideration of this study. Notably, the same experiments can be reproduced under identical controlled settings using different model families, enabling broader investigations into persona-based generation and its generalisability across architectures. We also excluded few-shot prompting, as the additional contextual cues it provides could obscure the model's intrinsic ability to adapt to target and interlocutor personas. Furthermore, LLMs are known to be sensitive to the order of few-shot examples, and the increased diversity of the experimental conditions would have made consistent prompting infeasible, introducing substantial noise in the evaluation.

Since PRODIGy dialogues and characters are derived from movies and screenplays that are available online, there might be data contamination effects. To control for these effects, we designed experiments generating dialogues across diverse combinations of character pairings and topics, introducing incremental randomness to evaluate the model's generalisation capabilities. For three specific configurations, we created entirely new non-PRODIGy characters, not found in any existing resource, to assess the model's ability to generalise when dealing with completely novel and unseen personas.

The configurations are designed to vary across two key dimensions—*interlocutors* and *topics*—as outlined in Q2:

- **Speaker pairings**: Dialogues were generated using the following configurations:
  - (i) *Familiar pairings*: Characters were paired (a) as they appear in the PRODIGy dataset (e.g., Harry and Sally from the movie *When Harry Met Sally*), (b) using entirely new, non-PRODIGy characters (e.g., John Doe and Jane Doe).[1]
  - (ii) *Unfamiliar pairings*: Characters from different movies within the PRODIGy dataset were paired (e.g., Harry and Superman), or PRODIGy characters were matched with new non-PRODIGy characters (e.g, Harry and John Doe).
- **Topics**: Dialogues were generated using the following types of topics:
  - (i) *Familiar topics*: Those associated with the characters in the PRODIGy dataset (e.g., Harry and Sally discussing about love).
  - (ii) *Unfamiliar topics*: New topics unfamiliar to the target speaker (e.g., Harry having a dialogue about war).

Further details are provided in Appendix D.

Non-PRODIGy Characters Generation To generate the new non-PRODIGy characters, we employed the GPT4o-mini (OpenAI, 2024) model and the Persona Hub dataset (Ge et al., 2024), comprising 1 billion personas automatically extracted and curated from web data. We selected GPT4o-mini based on a preliminary experiment where it demonstrated the ability to generate detailed profiles while remaining cost-efficient. From the Persona Hub dataset, we randomly selected 100 personas and crafted detailed prompts using these persona descriptions (see Appendix C.1). These prompts were used to instruct GPT4o-mini to generate biographies in the style of PRODIGy biographies. Unlike the original PRODIGy biographies, consisting of fictional movie characters, the newly generated biographies represent ordinary persons.

Dialogue Generation We approached dialogue generation as a next-turn generation task. The prompt was updated at each step to include the dialogue history up to that point. To maintain consistency, we limited each character's biography to their first five sentences across all configurations, as this represented the minimum biography sentence number in PRODIGy. Each dialogue was generated with a total of 8 turns, in line with recent findings showing that persona alignment tends to decay beyond this point, with noticeable "instruction drift" occurring after 8 turns (Li et al., 2024). This choice ensures we capture the phase where persona signals are strongest and facilitates more consistent human evaluation. We generated the dialogues starting from each combination of speakers, thus obtaining a total of 4375 dialogues. Figure 1 shows an example of generated dialogue.

3.3 Evaluation

A model can be considered successful in generating responses that align with a target biography and adapt to the interlocutor's information, if the speakers' biographies are identifiable from the generated dialogue. Recognising profiling indicators in text has been shown to be a complex task, especially for humans (Youyou et al., 2015; Flekova et al., 2016; De Mattei et al., 2020). To address this challenge, we employed an LLM-as-a-judge approach in addition to human evaluations. Moreover, we framed the task as author identification, where given a dialogue and three possible biographies, the evaluators (both human and automatic) were tasked to identify which biography belongs to the target speaker. While traditional statistical classifiers could serve as an additional baseline, we chose not to include them, as such models typically rely on surface-level textual features, which are not well-suited to assess the models' ability to

[1]We treat new characters as familiar because the unfamiliar category comprises cases where we forced a highly unexpected interlocutor onto a speaker, which is not the case for new characters as there are no expectations.

adapt to speakers' personas. In contrast, our framework, grounded in both human judgment and LLM evaluation, is designed to assess deeper aspects of persona consistency that a statistical method might overlook.

We designed four distinct configurations, varying in the degree of interlocutor information disclosed to the evaluator (see Figure 3 for examples):

1. **Interlocutor's biography and turns disclosed (Both$_{Disc}$)**: Evaluators were provided with the full dialogue, the interlocutor's biography and dialogue turns. This configuration examines whether the combination of the interlocutor's information with the target speaker's turns enhances the ability to accurately recognise the correct target speaker's biography (Figure 3a).

2. **Interlocutor's biography disclosed and turns masked (Bio$_{Disc}$)**: Evaluators received the interlocutor's biography, but their dialogue turns were hidden. This setup tests whether the interlocutor's biography is more relevant than their turns in order to correctly recognise the target speaker's biography (Figure 3b).

3. **Interlocutor's biography masked and turns disclosed (Turns$_{Disc}$)**: Evaluators were given the full dialogue with the interlocutor's biography hidden. This configuration examines whether the interlocutor's dialogue turns alone are more relevant than their biography for accurately identifying the target speaker's biography (Figure 3c).

4. **Both interlocutor's biography and turns masked (Both$_{Mask}$)**: Neither the interlocutor's biography nor their dialogue turns were provided. By masking all interlocutor information, we assess whether the target speaker's biography can still be accurately identified without any clue about the other speaker (Figure 3d).

To make recognition challenging and prevent evaluators from relying on overly obvious cues in the biographies, we selected the two biographies most semantically similar to the target speaker's biography as alternative options. For this, we used SBERT (Thakur et al., 2021), computing the cosine similarity of the biography embeddings generated with the all-MiniLM-L6-v2 model[2]. Additionally, we crafted four instruction templates to handle the four disclosure configurations (see Appendix C.3).

Automatic Evaluation We fine-tuned Llama 3.1 8B Instruct as LLM-as-a-judge utilising the dialogue partitions from the PRODIGy dataset as described in Section 3.1. We opted for a fine-tuned model to specialise the evaluator for the domain of movie dialogues, which exhibit distinctive stylistic and structural features that our dialogue model consistently manages to emulate. At inference time, we employed a greedy decoding mechanism to generate the LLM-as-a-judge's predictions, ensuring consistency and determinism in the evaluation process (Song et al., 2024). Our methodology involved automatic assessment of both gold and generated dialogues, with each dialogue evaluated twice, once for the target speaker and once for the interlocutor.

Human Evaluation For high-quality human evaluation, we recruited native English speakers from Prolific[3], a platform specifically designed for research experiments, granting fair payment to annotators. Each dialogue was evaluated by three evaluators, who conducted a detailed review of the dialogues across the four disclosure configurations, following the same assessment methodology as the LLM-as-a-judge to maintain consistency in evaluation criteria. Additionally, we held a post-hoc qualitative interview with the evaluators to gain further insights into the evaluation. The evaluation guidelines are reported in Appendix F.

4 Automatic Results

In this section, we present the LLM-as-a-judge's performance on the gold test set and on model-generated dialogues, including both the fine-tuning and zero-shot configurations. Results are shown for the various biography- and turn-disclosure conditions. We employ the following metrics: accuracy (Acc), precision (Prec), recall (Rec), and F1 Score (F1). Detailed results are reported in Appendix E.

Baseline As a baseline, we take a random guess over the three speakers, yielding an accuracy of 0.333 across all configurations, as the pool of possible candidates for speaker identification is always three under all settings.

4.1 Gold Dialogues

Table 1 examines the effect of various degrees of disclosure of the interlocutor's information on the

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://www.prolific.com/

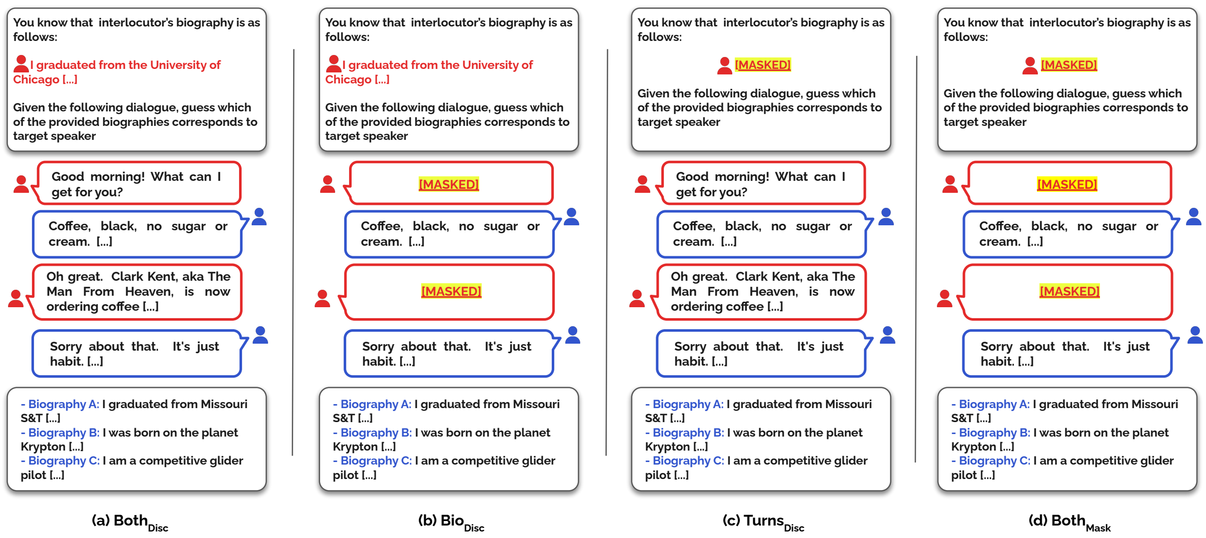

Figure 3: Information disclosure configurations used in the evaluation: (a) Both$_{Disc}$: both the interlocutor's biography and turns are visible to the evaluator; (b) Bio$_{Disc}$: only the interlocutor's biography is visible, while their turns are masked in the dialogue; (c) Turns$_{Disc}$: only the interlocutor's turns are visible, with their biography masked; (d) Both$_{Mask}$: both the interlocutor's biography and turns are masked. The illustrated examples are simplified for clarity and demonstrate the input available to evaluators in each condition.

judge's ability to correctly identify speakers. We observe that providing the LLM-as-a-judge with the interlocutor's biographical information significantly enhances the judge's accuracy. This suggests that in the gold data the interlocutor's information is relevant to the judge LLM. When both the interlocutor's biography and their turns are provided (Both$_{Disc}$), the model achieves a high accuracy of 0.820. Providing the biography only (Bio$_{Disc}$) still supports robust performance, with a comparable accuracy of 0.805. Masking the biography instead yields a notable drop: relying solely on the interlocutor's disclosed turns (Turns$_{Disc}$) lowers the performance (0.588 accuracy), while masking both the biography and turns (Both$_{Mask}$) leads to further reductions (0.577 accuracy).

Disclosure	Acc	Prec	Rec	F1
Both$_{Disc}$	**0.820**	**0.838**	**0.817**	**0.819**
Bio$_{Disc}$	0.805	0.822	0.801	0.802
Turns$_{Disc}$	0.588	0.629	0.583	0.581
Both$_{Mask}$	0.577	0.619	0.571	0.568

Table 1: LLM-as-a-judge results on PRODIGy gold dialogues across various degrees of interlocutor information disclosure.

The performance of the LLM-as-a-judge on gold dialogues clearly demonstrates its ability to accurately identify speakers. Moreover, the analysis highlights that biographical information plays a critical role in significantly enhancing the model's speaker recognition capabilities.

4.2 Fine-Tuned Model's Generations

In this section, we present the results of the evaluation of the fine-tuned model's dialogues.

The fine-tuned model effectively adapts its responses to both the target and interlocutor speakers (Q1) As shown in Table 2, the interlocutor's biography plays a more significant role than their conversational turns. This indicates that the model efficiently tailors its responses to the interlocutor's biography. Notably, the performances for conditions where both the interlocutor's biography and turns are disclosed (*Both*$_{Disc}$) and where only the biography is disclosed (*Bio*$_{Disc}$) are comparable (0.594 accuracy). Instead, performance drops significantly when only interlocutor's turns are disclosed (*Turns*$_{Disc}$) or both interlocutor's turns and biography are masked (*Both*$_{Mask}$) (accuracy of 0.517 and 0.515, respectively). These lower scores, compared to gold dialogues, could also be due to a data contamination effect, as the judge LLM might be familiar with the gold dialogues.

These results highlight the importance of the interlocutor's information in enabling evaluators to correctly match the target speaker's biography, suggesting that the model appropriately adapts its responses to both the target and interlocutor.

Disclosure	Acc	Prec	Rec	F1
Both$_{Disc}$	**0.594**	**0.618**	0.594	0.588
Bio$_{Disc}$	**0.594**	0.615	**0.595**	**0.589**
Turns$_{Disc}$	0.517	0.552	0.518	0.508
Both$_{Mask}$	0.515	0.546	0.515	0.505

Table 2: LLM-as-a-judge results on fine-tuned model's dialogues across different degrees of interlocutor information disclosure.

The fine-tuned model adapts well to familiar speaker pairings but struggles with unfamiliar ones (Q2) Table 3 highlights that the LLM-as-a-judge performs best in dialogues with familiar speaker pairings, such as PRODIGy-PRODIGy pairs from the same movie or non-PRODIGy-non-PRODIGy pairs (0.703 accuracy). This improved performance likely stems from the pairings aligning with the LLM's prior knowledge, leading to more intuitive dialogues. Conversely, the model's performance drops substantially with unfamiliar speaker pairings, such as PRODIGy-PRODIGy pairs from different movies or PRODIGy-non-PRODIGy pairings (0.496 accuracy), indicating that these dialogues are more challenging for speaker identification.

Speakers	Acc	Prec	Rec	F1
Familiar	**0.703**	**0.728**	**0.701**	**0.699**
Unfamiliar	0.496	0.524	0.498	0.487

Table 3: LLM-as-a-judge results on fine-tuned model's generated dialogues across different speaker pairings.

Overall, dialogues generated from familiar speaker pairs make it easier to identify the target speaker's biography, while unfamiliar pairings prove less intuitive, leading to a lower generalisation ability.

The fine-tuned model effectively generalises across topics (Q2) Table 4 shows that the LLM maintains relatively stable performance across different topics, with performance only slightly higher for the familiar topic (0.566 vs 0.561 accuracy).

Topic	Acc	Prec	Rec	F1
Familiar	**0.566**	**0.595**	**0.566**	**0.558**
Unfamiliar	0.561	0.588	0.563	0.554

Table 4: LLM-as-a-judge results on fine-tuned model's dialogues across topics.

The results reveal no significant differences in performance between dialogues with familiar topics and those with unfamiliar topics, suggesting robust generalisation across topics instead.

4.3 Zero-Shot Generations

The results in this section are based on dialogues generated in a zero-shot setting. The findings are consistent with those observed for the gold dialogues and the fine-tuned model's generations.

In a zero-shot setting, the model effectively tailors its responses to both target speaker and interlocutor (Q1) Table 5 shows results for various interlocutor's information disclosure settings. As for the fine-tuned model, performance is highest when both biography and turns are disclosed (*Both$_{Disc}$*, 0.755 accuracy). Partial disclosure – either only biography (*Bio$_{Disc}$*) or turns (*Turns$_{Disc}$*) – yields slightly lower metrics, while masking both (*Both$_{Mask}$*) yields the lowest accuracy.

Disclosure	Acc	Prec	Rec	F1
Both$_{Disc}$	**0.755**	**0.769**	**0.754**	**0.754**
Bio$_{Disc}$	0.751	0.763	0.749	0.749
Turns $_{Disc}$	0.742	0.771	0.740	0.740
Both$_{Mask}$	0.735	0.757	0.734	0.733

Table 5: LLM-As-A-Judge performance metrics for different disclosure types in zero-shot dialogues.

These findings underscore that the dialogue model adapts appropriately to the interlocutor's information, allowing for better speaker recognition.

In a zero-shot setting, the model adapts well to familiar speaker pairings but struggles with unfamiliar ones (Q2) Table 6 shows that dialogues involving familiar speaker pairs yield better performance metrics (e.g., 0.816 accuracy). In contrast, dialogues between unfamiliar speaker pairings lead to a marked drop (0.696 accuracy). This pattern mirrors the trends observed in the fine-tuned model, emphasising the importance of familiar speaker pairings for coherent dialogue generation.

Speakers	Acc	Prec	Rec	F1
Familiar	**0.816**	**0.831**	**0.816**	**0.815**
Unfamiliar	0.696	0.718	0.695	0.694

Table 6: LLM-as-a-judge results on zero-shot generated dialogues across different speaker pairings.

In a zero-shot setting, the model effectively generalises across topics (Q2) Table 7 shows a consistent performance of the LLM across familiar and unfamiliar topics. This trend, consistent with that of the fine-tuned model, suggests that the LLM can robustly generalise in a zero-shot setting.

Topic	Acc	Prec	Rec	F1
Familiar	**0.737**	0.755	**0.736**	**0.735**
Unfamiliar	0.735	**0.756**	0.735	0.734

Table 7: LLM-As-A-Judge performance for zero-shot dialogues across topics.

4.4 Fine-Tuning vs Zero-Shot

Notably, the performance metrics for zero-shot dialogues are higher than for the fine-tuned model. We hypothesise that this may be due to the LLMs tendency to replicate information directly from their prompts (Russo et al., 2023; Casula et al., 2024). In our scenario, the model might replicate in the generations details from the provided biographies. While facilitating speaker recognition, this "copy-paste" tendency may also limit the model's ability to fully capture and convey deeper characteristics of speakers, resulting in outputs that merely copy the surface elements of the biographical data. To investigate this hypothesis, we conducted an additional analysis examining word overlaps between the speakers' biographies and their turns.

Additional fine-tuning on persona-based dialogues is beneficial to the model (Q3) The results revealed a pronounced tendency for the zero-shot setting to reuse specific terms directly from the biographies. Checking rare words from the biographies, we found that they appeared up to eight times more frequently in zero-shot dialogues (40.69%) compared to those generated by the fine-tuned model (5.18%) and to the gold dialogues (5.84%). Details are in Table 20 in Appendix E.3. This "copy-paste" behaviour makes the speaker more easily recognisable, but also the outputs more superficial. In contrast, the fine-tuned model's generations seem to resemble more the gold dialogues, indicating a stronger ability to generate responses reflecting deeper speaker characteristics.[4]

Figure 4 illustrates this phenomenon. The biographies describe two distinct personas: one is a *petroleum engineer* who thrives in collaborative problem-solving environments, while the other is a

[4]A similar phenomenon was discussed in (Occhipinti et al., 2024), where models fine-tuned on data with high overlap between dialogues and biographies resulted in trivial "copy-pasting" models.

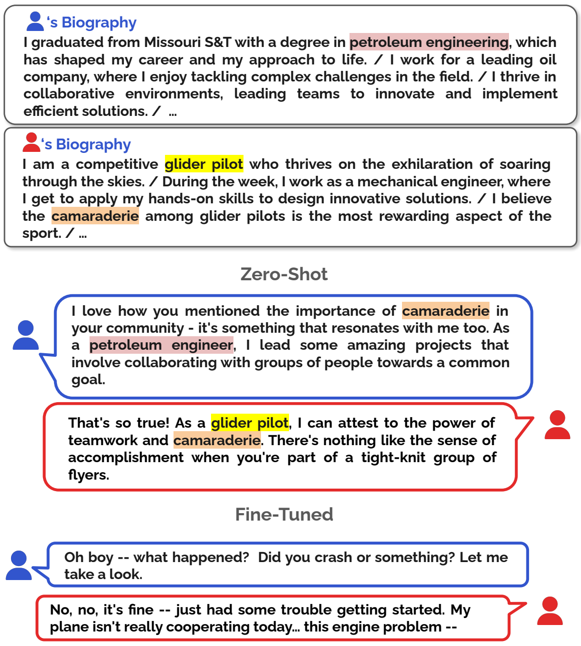

Figure 4: Comparison of dialogues generated in a zero-shot setting vs. the fine-tuned model, using the same input and Non-PRODIGy characters.

glider pilot who values *camaraderie*. While the dialogue from the zero-shot setting merely reproduces these exact terms instead of more common alternatives (e.g., *friendship among pilots*), the fine-tuned models' dialogue reflects deeper persona traits by portraying a collaborative and problem-solver person speaking to a pilot facing engine issues.

Notably, while looking at infrequent words clearly shows the "copy-paste" phenomenon, this is not limited to rare word repetition. We also observe significantly higher BLEU, ROUGE, and METEOR scores when comparing biographies and dialogue turns in the zero-shot setting compared to the fine-tuned model (e.g., the average METEOR score for the fine-tuned model is 0.043, whereas it is 0.140 for the zero-shot setting; see Table 21 in Appendix E.3). This indicates that copying extends to broader content overlap, not just rare words.

5 Human Evaluation Results

The human evaluation, conducted by evaluators hired through Prolific, consisted of 360 assessments focusing on a subset of dialogues previously evaluated by the LLM-as-a-judge. Dialogues were randomly selected and stratified by experiment type and level of interlocutor information disclosure to ensure balanced representation. Due to the "copy-paste" phenomenon observed in zero-shot dialogues, which could make the task trivial, the evaluation was limited to gold dialogues and dialogues generated by the fine-tuned model.

Overall, the trends observed in the human evaluation align with those from the LLM-as-a-judge assessment. However, scores from the human evaluation are considerably lower than those for the LLM-as-a-judge, indicating that LLMs surpass human capabilities in this task. This observation is consistent with previous findings (Youyou et al., 2015; Flekova et al., 2016; De Mattei et al., 2020).

Human evaluations confirm that the model effectively adapts its responses to target and interlocutor speakers (Q1) Human evaluation results show that providing the interlocutor's biography significantly improves biography recognition for both generated and gold dialogues. In contrast, masking both the interlocutor's biography and turns severely hinders the evaluators' ability to accurately identify the correct biography.

Table 8 presents the human evaluation results across the four levels of interlocutor information disclosure. Disclosing both interlocutor's biography and turns leads to the most accurate identification of the target speaker, as indicated by the highest accuracy (0.509 in Both$_{Disc}$). Notably, revealing only the interlocutor's turns while masking their biography results in better outcomes (0.474 accuracy in Turns$_{Disc}$) compared to providing the interlocutor's biography while masking their turns (0.449 accuracy in Bio$_{Disc}$). The lowest accuracy is observed when all interlocutor information is masked (0.351 accuracy in Both$_{Mask}$).

Disclosure	Acc	Prec	Rec	F1
Both$_{Disc}$	**0.509**	**0.478**	0.375	**0.395**
Bio$_{Disc}$	0.449	0.447	**0.388**	0.370
Turns$_{Disc}$	0.474	0.434	0.385	0.375
Both$_{Mask}$	0.351	0.332	0.248	0.249

Table 8: Human evaluation results on fine-tuned model's generations and gold dialogues across different degrees of interlocutor information disclosure.

Human evaluations confirm that the model better adapts its responses with familiar speaker pairings, while struggles with unfamiliar ones (Q2) Evaluators achieve the highest performance when dealing with dialogues involving common speaker pairings. The results in Table 9 indicate that generating dialogues from common speaker pairings leads to higher target speaker recognition (0.497 accuracy), whereas unfamiliar pairings result in lower performance (0.373 accuracy).

Speakers	Acc	Prec	Rec	F1
Familiar	**0.497**	**0.512**	**0.440**	**0.419**
Unfamiliar	0.373	0.336	0.264	0.273

Table 9: Human evaluation results on fine-tuned generated dialogues across different speaker pairings.

A post-hoc interview with the evaluators revealed that they were unable to differentiate between the generations and the gold dialogues, suggesting that the model's outputs closely mirrored the style and characteristics of the gold dialogues. The evaluators also reported that identifying the correct biography was challenging in most cases, as multiple biographies could plausibly match the target's turns. In many instances, the choice was determined only by spotting a small detail.

6 Conclusion

We studied the role of interlocutor information in persona-based dialogue generation, analysing how LLMs adapt their responses based on both target speaker and interlocutor characteristics. We introduced a novel, systematic framework for evaluating the model's ability to (i) align responses with both target and interlocutor personas; (ii) its generalisation abilities across familiar and unfamiliar speaker pairings and across topics familiar and unfamiliar to the target speaker; (iii) and the impact of fine-tuning compared to zero-shot generation. Our findings reveal that while models effectively align responses with target speakers and adapt to interlocutor biographies, their performance drops with unfamiliar pairings. Similarly, models demonstrate strong generalisation across different topics, even those unfamiliar to the target speaker. Notably, dialogues generated in a zero-shot setting achieve higher speaker recognition accuracy by directly "copy-pasting" biographical details, thereby also leading to much more superficial dialogues. In contrast, fine-tuned models more accurately capture deeper speaker traits. These insights have practical relevance for personalised dialogue systems, such as digital assistants, chatbots, and social robots, where maintaining a consistent and contextually appropriate persona is essential for enhancing user trust, engagement, and interaction.

Limitations

Our experiments rely on the PRODIGy dataset, which features movie dialogues and fictional characters. This choice may introduce stereotyped roles and inherent biases. Additionally, while our evaluation combines LLM-as-a-judge and human assessments, each method has limitations: the LLM evaluator might be affected by data contamination, and human evaluations are subject to individual interpretation and limited reproducibility.

Ethics Statements

Role-play agents carry several risks, including toxicity, bias, hallucinations, and privacy violations (Chen et al., 2024). Early studies have revealed a tendency of these models to generate harmful content (Wen et al., 2023), which not only degrades the user experience but also presents significant safety concerns. Furthermore, they often exhibit role-based biases, stemming from both inherent biases in their pre-training data (Xue et al., 2023) and user prompts that may inadvertently direct them toward biased outputs (Perez and Ribeiro, 2022; Branch et al., 2022). Additionally, these systems might exhibit *character hallucination* (Ahn et al., 2024), i.e., producing responses that do not align with their assigned roles. Lastly, role-play systems may pose privacy risks by inadvertently disclosing users' private information (Krishnamurthy et al., 2011; Corrigan et al., 2014).

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hezekiah J Branch, Jonathan Rodriguez Cefalu, Jeremy McHugh, Leyla Hujer, Aditya Bahl, Daniel del Castillo Iglesias, Ron Heichman, and Ramesh Darwishi. 2022. Evaluating the susceptibility of pre-trained language models via handcrafted adversarial examples. *arXiv preprint arXiv:2209.02128*.

Camilla Casula, Sebastiano Vecellio Salto, Alan Ramponi, and Sara Tonelli. 2024. Delving into qualitative implications of synthetic data for hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19709–19726, Miami, Florida, USA. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

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.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, and Malvina Nissim. 2020. Invisible to people but not to machines: Evaluation of style-aware HeadlineGeneration in absence of reliable human judgment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6709–6717, Marseille, France. European Language Resources Association.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 843–854, Berlin, Germany. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–574.

Balachander Krishnamurthy, Konstantin Naryshkin, and Craig Wills. 2011. Privacy leakage vs. protection measures: the growing disconnect. In *Proceedings of the Web*, volume 2, pages 1–10.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling instruction (in)stability in language model dialogs. *Preprint*, arXiv:2402.10962.

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.

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.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. Partner personas generation for dialogue response generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5200–5212, Seattle, United States. Association for Computational Linguistics.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *Preprint*, arXiv:2401.12474.

Daniela Occhipinti, Serra Sinem Tekiroğlu, and Marco Guerini. 2024. PRODIGy: a PROfile-based DIalogue generation dataset. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3500–3514, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, Singapore. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, Alberto Barrón-Cedeño, et al. 2014. Overview of the author identification task at pan 2014. In *CEUR Workshop Proceedings*, volume 1180, pages 877–897. CEUR-WS.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *Preprint*, arXiv:2406.01171.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3271–3290, Bangkok, Thailand. Association for Computational Linguistics.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 201–211.

Jintang Xue, Yun-Cheng Wang, Chengwei Wei, Xiaofeng Liu, Jonghye Woo, and C-C Jay Kuo. 2023. Bias and fairness in chatbots: An overview. *arXiv preprint arXiv:2309.08836*.

Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic LoRA for efficient multi-character role-playing agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557, Miami, Florida, USA. 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.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

A Dialogue Annotation With Topic

In this section, we detail the dialogue annotation process for topic labelling.

First, each dialogue was assigned three candidate topics using Llama 3.1 8B Instruct (Dubey et al., 2024)[5]. Next, we extracted the most frequent topics by analysing the word stems in these labels and selected the 100 most common stems. These stems were then clustered into broader thematic categories, and each dialogue was assigned the corresponding cluster label based on its generated topics. Finally, we validated the labelling accuracy by manually reviewing and correcting a sample of 200 dialogues. The complete procedure is outlined in Algorithm 1.

Algorithm 1 Topic Annotation

function ANNOTATETOPICS($Dialogues$)
  for all $d \in Dialogues$ **do**
    $T_d \leftarrow$ GENERATETOPICS(d, 3)
  end for
  $W \leftarrow$ EXTRACTSTEMS($\bigcup_{d \in Dialogues} T_d$)
  $S \leftarrow$ SELECTTOP(W, 100)
  $\mathcal{C} \leftarrow$ CLUSTERSTEMS(S)
  for all $d \in Dialogues$ **do**
    $d.label \leftarrow$ ASSIGNCLUSTERLABEL($T_d, \mathcal{C}$)
  end for
  HUMANVALIDATIONLABELS($Dialogues$, 200)
end function

B Model Fine-Tuning and Generations

We fine-tuned the Llama 3.1 8B Instruct using QLoRA (Dettmers et al., 2023), with learning rate 3e-4, low-rank approximation set to 16, low-rank adaptation set to 8, and dropout rate set to 0.0[6]. Evaluation steps were set at 524, batch size at 4 and gradient accumulation step at 2. We employed 10 epochs with early stopping after 3 epochs.

For generation, we set top-p to 0.9, temperature to 0.8, and applied a repetition penalty of 1.2.

[5] We excluded 10 dialogues with harmful content, as the model declined to assign topics to them.

[6] Followed hyper-parameters as `https://github.com/pytorch/torchtune/blob/main/recipes/configs/llama3_1/8B_qlora_single_device.yaml`

We framed both processes as a next-turn generation task. Dialogue data was introduced incrementally, turn by turn, allowing the model to better capture and adapt to conversational context.

C Prompts

C.1 Non-PRODIGy Character Generation Prompt

To generate biographies in the style of PRODIGy, we used the prompt shown in Table 10 to guide GPT4o-mini. Each persona from the Persona Hub dataset was unfamiliarly assigned a gender and a Myers-Briggs Type Indicator (MBTI) personality type. To ensure alignment with PRODIGy biographies, we instructed GPT4o-mini to include details about the individual's job, relationship status, lifestyle, and family background.

PROMPT 1
Considering the following persona sentence:
 	
`[persona sentence]`
 	
the following gender:
 	
`[gender]`
 	
and the following mbti:
 	
`[mbti]`
 	
please create a profile of a person, using the following structure:
 	
`{{"gender": "gender", "mbti": "mbti", "biography": [ "sentence 1", "sentence 2", "...", "sentence 10" ]}}`
 	
Please ensure the biography contains up to 10 sentences in the first person singular. Include details about the individual's job, relationship status, lifestyle, and family background. Be sure to capture a varied portrayal of the individual's life and character.
Please ensure to provide the dictionary only, without anything else.

Table 10: Prompt to generate Non-PRODIGy biographies

C.2 Generation Prompt

Table 11 shows the prompt employed for generations, both in fine-tuning and zero-shot settings.

C.3 Judge LLM prompts

In this section, we showcase the four prompts employed for the LLM-as-a-Judge evaluation. Table 12 displays the prompt for the Both$_{Disc}$ configuration, where both the interlocutor's biography and turns are disclosed. Table 13 corresponds to the Bio$_{Disc}$ setting, where only the interlocutor's biography is disclosed while their turns are masked. Similarly, the Turns$_{Disc}$ configuration (Table 14) masks the interlocutor's biography while disclosing their turns. Finally, Table 15 represents the Both$_{Mask}$ condition, where both the interlocutor's biography and turns are masked.

PROMPT 2
You are `[target speaker]`. Your biography is as follows:
 	
`[biography sentences]`
 	
You are having a dialogue with `[interlocutor speaker]`. Their biography is as follows:
 	
`[biography sentences]`
 	
You are discussing about the topic `[topic]`. Considering the dialogue history, provide an answer to `[interlocutor speaker]`. Be sure to provide the answer only.
 	
`[dialogue history]`

Table 11: Prompt for both fine-tuning and dialogue generation

D Experimental Design Details

This section outlines our experimental settings, as detailed in Table 16. We designed seven experiments to assess how speaker pairing and topic variation affect dialogue generation, introducing incremental randomness to control for the data contamination effect. We label a configuration as *familiar* when the speakers belong to the same movie or belong to the same context (i.e., Non-PRODIGy characters), and as *unfamiliar* when the speakers belong to atypical contexts (e.g., characters coming from different movies or movie characters talking to ordinary persons). In Exp1, we preserve both the original PRODIGy pairing (target P_1, interlocutor P_2) and topic. Exp2 introduces a topic unfamiliar to the target speaker while keeping the original pairing. In Exp3, we replace the original PRODIGy Speaker 2 with a character from a different movie (target P_1, interlocutor P_{rand}), while keeping the topic familiar with the target speaker; Exp4 randomises both Speaker 2 and the topic. For experiments involving non-PRODIGy speakers, Exp5

PROMPT 3
You know that **interlocutor**'s biography is as follows:

[interlocutor's biography]

Given **interlocutor**'s biography and the following dialogue about the topic **[topic]**, your task is to guess which of the provided biographies corresponds to **target speaker**. The biographies are provided at the end of the dialogue. Please provide your answer as "Biography A", "Biography B", or "Biography C". Please make your guess even though the dialogue may sound a little weird or unnatural. Your response must follow this JSON format: {"Guess": "Biography X"}

DIALOGUE

[Dialogue]

BIOGRAPHIES

Biography A: **[Biography A sentences]**
Biography B: **[Biography B sentences]**
Biography C: **[Biography C sentences]**

Table 12: Both interlocutor's biography and turns disclosed

PROMPT 4
You know that **interlocutor**'s biography is as follows:

[interlocutor's biography]

Given **interlocutor**'s biography and the following dialogue about the topic **[topic]**, in which `interlocutor`'s turns are masked, your task is to guess which of the provided biographies corresponds to **target speaker**. The biographies are provided at the end of the dialogue. Please provide your answer as "Biography A", "Biography B", or "Biography C". Please make your guess even though the dialogue may sound a little weird or unnatural. Your response must follow this JSON format: {"Guess": "Biography X"}

DIALOGUE

[Dialogue]

BIOGRAPHIES

Biography A: **[Biography A sentences]**
Biography B: **[Biography B sentences]**
Biography C: **[Biography C sentences]**

Table 13: Interlocutor's biography disclosed, turns masked

PROMPT 5
Given the following dialogue about the topic `topic`, your task is to guess which of the provided biographies corresponds to `target speaker`. The biographies are provided at the end of the dialogue. Please provide your answer as "Biography A", "Biography B", or "Biography C". Please make your guess even though the dialogue may sound a little weird or unnatural. Your response must follow this JSON format: {"Guess": "Biography X"}

DIALOGUE

[Dialogue]

BIOGRAPHIES

Biography A: **[Biography A sentences]**
Biography B: **[Biography B sentences]**
Biography C: **[Biography C sentences]**

Table 14: Interlocutor's biography masked, turns disclosed

PROMPT 6
Given the following dialogue about the topic `topic`, in which `interlocutor's` turns are masked, your task is to guess which of the provided biographies corresponds to `target speaker`. The biographies are provided at the end of the dialogue. Please provide your answer as "Biography A", "Biography B", or "Biography C". Please make your guess even though the dialogue may sound a little weird or unnatural. Your response must follow this JSON format: {"Guess": "Biography X"}
DIALOGUE

[Dialogue]

BIOGRAPHIES

Biography A: **[Biography A sentences]**
Biography B: **[Biography B sentences]**
Biography C: **[Biography C sentences]**

Table 15: Both interlocutor's biography and turns masked

pairs the original PRODIGy Speaker 1 with a new non-PRODIGy Speaker 2 (target P_1, interlocutor N_{rand}) while maintaining the topic, whereas Exp6 uses both a random non-PRODIGy Speaker 2 and a random topic. Finally, Exp7 features an entirely new pairing of non-PRODIGy speakers (target N_1, interlocutor N_2) with a random topic.

E Detailed Automatic Results

E.1 Fine-Tuned Model's Results

In this section, we explore how the LLM-as-a-judge performs under various disclosure scenarios when identifying speakers in dialogues generated by the

Exp	Target	Interlocutor	Topic	Pairing
Exp1	P_1	P_2	Familiar	Familiar
Exp2	P_1	P_2	Unfamiliar	Familiar
Exp3	P_1	P_{rand}	Familiar	Unfamiliar
Exp4	P_1	P_{rand}	Unfamiliar	Unfamiliar
Exp5	P_1	N_{rand}	Familiar	Unfamiliar
Exp6	P_1	N_{rand}	Unfamiliar	Unfamiliar
Exp7	N_1	N_2	Unfamiliar	Familiar

Table 16: Overview of experimental configurations assessing the impact of speaker pairing and topic variation on dialogue generation.

fine-tuned model, by systematically varying the information provided—ranging from full access to interlocutors' biographies and dialogue turns to masking this information. The results, as outlined in Table 17, highlight the model's adaptability and its limitations when confronted with unfamiliar or artificially constructed scenarios.

Revealing both the interlocutor's biography and their previous turns yields high accuracy scores, 0.83 in Exp1 and 0.807 in Exp2. Providing only the biography while masking the turns still yields strong accuracy, 0.825 in Exp1 and 0.805 in Exp2. However, when both the biography and turns are masked, accuracy drops markedly to 0.587 in Exp1 and 0.570 in Exp2. These results suggest that the fine-tuned model effectively incorporates character biographies, even across varying topics, allowing the judge LLM to reliably identify speakers.

The model's performance shifts significantly when we introduce more artificial scenarios, i.e., PRODIGy speakers coming from different movies (Exp3) or combinations of PRODIGy and Non-PRODIGy characters (Exp4). When both biography and dialogue turns are disclosed, accuracy drops noticeably to 0.558 in Exp3 and 0.560 in Exp4, substantially lower than in the original character scenarios. Interestingly, revealing only the interlocutor's turns while masking the biography yields slightly improved scores (accuracy of 0.578 in Exp3 and 0.572 in Exp4). Conversely, providing only the biography while masking the turns lowers performance, with accuracy scores of 0.553 in Exp3 and 0.558 in Exp4.

The performance drop is even more pronounced when PRODIGy characters are paired with Non-PRODIGy characters. In these cases, even full disclosure of both biography and turns yields weaker scores (0.407 in Exp5 and 0.405 in Exp6). Providing only the interlocutor's biography while masking their turns results in slightly improved performance for Exp5 (0.414 accuracy) but lower scores for Exp6 (0.399 accuracy). Surprisingly, the best scores occur when both the interlocutor's biography and turns are masked (accuracy of 0.435 for Exp5 and 0.460 for Exp6). Detailed breakdowns for Exp5 and Exp6 in Speaker1 and Speaker2 metrics, as shown in Table 18, indicate these poor scores are due to the fact that the model struggles to effectively recognise non-PRODIGy interlocutors.

In the case of dialogues between unfamiliar non-PRODIGy speakers (Exp7), the highest accuracy (0.606) appears when only the biography is disclosed and the turns are masked. Providing both the biography and turns reduces performance slightly to 0.591. Masking the biography, however, causes a severe drop, with accuracy scores of 0.414 when only turns are disclosed and 0.415 when both biography and turns are masked.

Exp	Disclosure	Acc	Prec	Rec	F1
Exp1	Both_{Disc}	**0.830**	**0.846**	**0.825**	**0.828**
	Bio_{Disc}	0.825	0.841	0.820	0.822
	Turns_{Disc}	0.601	0.640	0.596	0.592
	Both_{Mask}	0.587	0.620	0.582	0.577
Exp2	Both_{Disc}	**0.807**	**0.825**	**0.808**	**0.808**
	Bio_{Disc}	0.805	0.823	0.806	0.804
	Turns_{Disc}	0.565	0.604	0.566	0.561
	Both_{Mask}	0.570	0.606	0.571	0.567
Exp3	Both_{Disc}	0.558	0.591	0.561	0.554
	Bio_{Disc}	0.553	0.578	0.556	0.549
	Turns_{Disc}	**0.578**	**0.622**	**0.582**	**0.573**
	Both_{Mask}	0.565	0.604	0.568	0.560
Exp4	Both_{Disc}	0.560	0.582	0.558	0.551
	Bio_{Disc}	0.558	0.578	0.556	0.549
	Turns_{Disc}	**0.572**	**0.609**	**0.570**	0.561
	Both_{Mask}	**0.572**	0.607	**0.570**	**0.562**
Exp5	Both_{Disc}	0.407	0.429	0.409	0.398
	Bio_{Disc}	0.414	0.436	0.416	0.403
	Turns_{Disc}	0.434	**0.469**	0.436	0.421
	Both_{Mask}	**0.435**	0.465	**0.438**	**0.423**
Exp6	Both_{Disc}	0.405	0.421	0.411	0.396
	Bio_{Disc}	0.399	0.414	0.406	0.391
	Turns_{Disc}	0.458	0.490	0.466	0.450
	Both_{Mask}	**0.460**	**0.491**	**0.467**	**0.452**
Exp7	Both_{Disc}	0.591	0.631	0.589	0.584
	Bio_{Disc}	**0.606**	**0.637**	**0.605**	**0.602**
	Turns_{Disc}	0.414	0.433	0.412	0.397
	Both_{Mask}	0.415	0.431	0.412	0.395

Table 17: Judge LLM results fine-tuned model's generations

E.2 Zero-Shot Results

Table 19 presents the metrics from our zero-shot experiments. Compared to the fine-tuned model's results, these zero-shot evaluations follow similar

		Acc		Prec		Rec		F1	
	Disclosure	Sp1	Sp2	Sp1	Sp2	Sp1	Sp2	Sp1	Sp2
Exp5	Both_{Disc}	0.552	0.262	0.578	0.270	0.559	0.261	0.545	0.248
	Bio_{Disc}	0.541	0.286	0.571	0.292	0.548	0.284	0.533	0.270
	Turns_{Disc}	**0.562**	0.306	**0.600**	0.323	0.569	0.304	0.554	0.284
	Both_{Mask}	0.565	0.306	0.599	0.313	**0.572**	0.303	**0.558**	0.283
Exp6	Both_{Disc}	0.530	0.280	0.548	0.287	0.534	0.290	0.523	0.269
	Bio_{Disc}	0.515	0.283	0.539	0.286	0.519	0.292	0.509	0.273
	Turns_{Disc}	0.563	0.354	0.592	0.379	0.567	0.364	0.556	0.341
	Both_{Mask}	**0.574**	0.346	**0.614**	0.364	**0.580**	0.353	**0.569**	0.335

Table 18: Judge LLM results for Exp5 and Exp6 for Speaker 1 and Speaker 2 identification in fine-tuned model's generations.

trends but consistently yield higher scores.

When dealing with dialogues generated using original PRODIGy speaker pairings (Exp1 and Exp2), the judge model performs particularly well when it has access to the interlocutor's biography. In Exp1 (with original topic), disclosing both biography and turns yields an accuracy of 0.882, while disclosing the biography alone still achieves 0.878. Similarly, in Exp2 (random topic), accuracy remains high at 0.887 and 0.888, respectively. In contrast, masking the biography leads to a performance drop. When only the interlocutor's turns are disclosed, accuracy declines to 0.750 in Exp1 and 0.774 in Exp2, and further drops when both biography and turns are masked, reaching 0.737 in Exp1 and 0.730 in Exp2.

When PRODIGy speakers come from different movies or interact with Non-PRODIGy characters, performance declines. In these scenarios (Exp3 and Exp4), biographical information offers no clear advantage, while focusing on interlocutor turns or masking all interlocutor details proves more effective. In Exp3, disclosing only turns yields an accuracy of 0.718, slightly improving to 0.722 when both biography and turns are masked. In contrast, revealing both interlocutor's biography and turns lowers accuracy to 0.701, and providing only the biography results in an even lower 0.694. A similar pattern emerges in Exp4, where disclosing only turns achieves the highest accuracy (0.725), followed by masking both (0.712), while approaches that reveal biography perform worse (0.691 and 0.676, respectively).

Performance degradation is even more pronounced when pairing PRODIGy characters with Non-PRODIGy characters (Exp5 and Exp6). In these cases, disclosing interlocutor's biographical information actually hinders performance compared to relying solely on interlocutor's turns or masking both elements. In Exp5, the highest accuracy (0.730) is achieved by masking both interlocutor's biography and turns, followed closely by disclosing only their turns (0.723). In contrast, providing interlocutor's biography leads to significantly worse performance, with accuracy dropping to 0.657 when both biography and turns are disclosed and 0.654 when only biography is revealed. Exp6 follows a similar trend, where the best results occur when both elements are masked (0.731) or when only turns are disclosed (0.723), while disclosing biography results in the lowest scores (0.645 with both biography and turns disclosed, 0.636 with only biography disclosed).

However, when evaluating dialogues between non-PRODIGy speakers (Exp7), performance are similar to those observed in the original PRODIGy pairing scenarios. The highest accuracy score (0.826) is achieved when only the interlocutor's biography is disclosed, with nearly identical performance (0.824 accuracy) when both biography and turns are provided. Performance declines when biographical information is masked, dropping to accuracy scores of 0.778 when only turns are disclosed and 0.786 when both turns and biography are masked.

E.3 Overlap Metrics and Rare Words

To assess how extensively models replicate information from the target biographies into generated target turns, we examined the incorporation of specific rare words from the target biographies into the generated speaker turns. When a specific rare word from a biography appears in the turns, it enhances the recognisability of the target speaker. This pattern indicates a tendency for the model to copy and paste biographical information into the dialogue.

For this analysis, we employed the `wordfreq`

Exp	Disclosure	Acc	Prec	Rec	F1
Exp1	Both$_{Disc}$	**0.882**	**0.889**	**0.883**	**0.882**
	Bio$_{Disc}$	0.878	0.884	0.879	0.878
	Turns$_{Disc}$	0.750	0.777	0.751	0.749
	Both$_{Mask}$	0.737	0.755	0.738	0.736
Exp2	Both$_{Disc}$	0.887	**0.898**	0.886	**0.888**
	Bio$_{Disc}$	**0.888**	**0.898**	**0.887**	**0.888**
	Turns$_{Disc}$	0.774	0.799	0.772	0.773
	Both$_{Mask}$	0.730	0.748	0.728	0.728
Exp3	Both$_{Disc}$	0.701	0.718	0.700	0.698
	Bio$_{Disc}$	0.694	0.707	0.693	0.692
	Turns$_{Disc}$	0.718	**0.749**	0.717	0.717
	Both$_{Mask}$	**0.722**	0.745	**0.721**	**0.720**
Exp4	Both$_{Disc}$	0.691	0.705	0.692	0.690
	Bio$_{Disc}$	0.676	0.689	0.677	0.673
	Turns$_{Disc}$	**0.725**	**0.751**	**0.726**	0.723
	Both$_{Mask}$	0.712	0.735	0.713	**0.710**
Exp5	Both$_{Disc}$	0.657	0.672	0.652	0.653
	Bio$_{Disc}$	0.654	0.667	0.649	0.650
	Turns$_{Disc}$	0.723	**0.752**	0.718	0.722
	Both$_{Mask}$	**0.730**	0.749	**0.725**	**0.727**
Exp6	Both$_{Disc}$	0.645	0.665	0.644	0.642
	Bio$_{Disc}$	0.638	0.659	0.638	0.636
	Turns$_{Disc}$	0.723	0.759	0.722	0.722
	Both$_{Mask}$	**0.731**	**0.761**	**0.731**	**0.730**
Exp7	Both$_{Disc}$	0.824	0.839	0.821	0.822
	Bio$_{Disc}$	**0.826**	**0.840**	**0.823**	**0.824**
	Turns$_{Disc}$	0.778	0.807	0.773	0.774
	Both$_{Mask}$	0.786	0.809	0.781	0.783

Table 19: Judge LLM results on zero-shot generations

library[7], utilising its `zipf_frequency` function, which ranks word rarity on a logarithmic scale. This scale represents the base-10 logarithm of the frequency per billion words (Van Heuven et al., 2014). For instance, a word with a Zipf value of 6 appears once per thousand words, while a value of 3 corresponds to once per million words. In our analysis, we defined rare words as those with a Zipf frequency below 4.0, focusing on infrequent terms to better capture copying tendencies.

Table 20 shows the percentage of dialogues where at least one rare word from the target speaker's biography also appears in their dialogue turns. Across all experiments, the zero-shot (ZS) setting consistently demonstrates higher overlap percentages than the fine-tuned (FT) model. For example, in Exp1, ZS reaches 41.84% compared to FT's 9.68%. The gold test set is included for comparison, showing a much lower overlap of 5.84%. These results highlight the zero-shot models' tendency to replicate rare biographical terms in generated dialogues.

	% dialogues	
Test	5.84%	
	FT	**ZS**
Exp1	9.68%	41.84%
Exp2	7.84%	39.76%
Exp3	5.04%	42.64%
Exp4	6.24%	40.64%
Exp5	3.28%	40.00%
Exp6	3.36%	36.48%
Exp7	0.72%	43.44%

Table 20: Percentage of dialogues where the target speaker's biography and dialogue turns share at least one rare word (threshold 4.0), comparing Fine-Tuned (FT) and Zero-Shot (ZS) settings.

Besides the use of rare words, we also computed overlap scores using METEOR, BLEU-1, and ROUGE-1 metrics[8]. The results, detailed in Table 21, include comparisons with the overlap scores between biographies and the gold turns. The results reveal that overlap metrics are significantly higher in the zero-shot setting compared to the fine-tuned setting across all configurations (e.g., Exp1 METEOR: FT 0.047 vs ZS 0.131; Exp7 METEOR: FT 0.036 vs ZS 0.153). Additionally, the scores for the fine-tuned model more closely match those of the gold dialogues (e.g., METEOR 0.032), suggesting an adaptation to the dialogues' gold characteristics.

	METEOR		BLEU-1		ROUGE-1	
Test	0.032		0.016		0.051	
	FT	**ZS**	**FT**	**ZS**	**FT**	**ZS**
Exp1	0.047	0.131	0.043	0.155	0.045	0.123
Exp2	0.046	0.129	0.043	0.154	0.044	0.123
Exp3	0.044	0.142	0.041	0.158	0.042	0.133
Exp4	0.045	0.138	0.041	0.156	0.043	0.130
Exp5	0.042	0.144	0.036	0.171	0.041	0.147
Exp6	0.042	0.141	0.036	0.170	0.042	0.145
Exp7	0.036	0.153	0.027	0.182	0.039	0.164

Table 21: Overlap metrics between speakers' biographies and turns: fine-tuned model's generations (FT) vs zero-shot generations (ZS)

F Human Evaluation Guidelines

Table 22 outlines the instructions provided to human evaluators for the evaluation task. Evaluators are presented with dialogues between two speakers and must identify the correct biography (Profile A, Profile B, or Profile C) for one of the speakers.

[7]https://github.com/rspeer/wordfreq/

[8]To compute the metrics we employed the Evaluate library (https://huggingface.co/docs/evaluate/).

HUMAN EVALUATION GUIDELINES You will be presented with a series of dialogues between two speakers. For each dialogue your task is to guess the correct biography (Profile) of one of the two speakers. There are 3 possible Profiles to choose among (shown at the end of the dialogue as "Profile A", "Profile B", or "Profile C"). Please make your guess even though the dialogue may sound a little weird or unnatural. You will also have the possibility to leave any comment you deem relevant about each dialogue/task in the "comments" field. **IMPORTANT**: for each dialogue we have different configurations where we either disclose or hide some of the interlocutor's information: - We will either provide or not interlocutor profile, in the latter case we use the tag [MASKED] - We will either show or mask other speakers' turns in the dialogue.

Table 22: Human Evaluation Guidelines

G Fine-Tuned Model's Generation Examples

In this section, we showcase four dialogues generated by the fine-tuned model. The first dialogue, involving an original PRODIGy speaker pairing (James Cole and Dr. Railly from the *12 Monkeys* movie), accurately captures Cole's paranoia and Railly's composed professionalism. The second dialogue pairs PRODIGy characters from different movies (Walter Sobchak from *The Big Lebowski* and Spider-Man), effectively preserving Walter's confrontational nature and Spider-Man's guilt-ridden introspection. The third dialogue involves a PRODIGy character (again Walter) interacting with a Non-PRODIGy character (a data analyst), showcasing Walter's strong opinions against the analyst's more reserved responses. The fourth dialogue, between two Non-PRODIGy characters (a high school student and a psychologist), maintains an introspective tone, reflecting their emotional depth. Across all dialogues, the model successfully aligns responses with the given persona traits, ensuring consistency.

DIALOGUE 1
Prompt:

You are **Speaker1**. Your biography is as follows:

- I live in what is left of Philadelphia in 2035.
- I am single, but I am developing feelings for my psychiatrist, Dr. Railly.
- I am a total nutcase, completely deranged, delusional, paranoid.
- I am visiting 1990, after being tossed in jail; I have been sent back in time to learn about a virus before it started mutating.
- When I told others I was from the future, they thought I was out of my mind and threw me in an insane asylum.

You are having a dialogue with **Speaker2**. Their biography is as follows:

- I am an intelligent and even-keeled psychiatrist.
- Philadelphia police asked me to diagnose the sanity of James Cole, a prisoner living in an underground compound.
- I am a good-natured, idealistic woman.
- I am arrogant and self-assured, but not obnoxiously smug like some of my colleagues.
- I am helpful and kind, and genuinely care about my patients.

Dialogue:

Speaker1: Dr. Railly... please help him! He's going to kill us all –
Speaker2: James... you need to calm down now. You can't let yourself get this upset because if we're going to work together here you both have to be able to stay focused on the task at hand which is identifying who he might be and why he did this to himself. Okay?
Speaker1: Who?!
Speaker2: His name is probably someone he trusts or has access too.
Speaker1: Okay okay – I know that already – his wife maybe...
Speaker2: Let's talk to each other very clearly right now. What do we know about Joseph Dunn?
Speaker1: He killed seven people then turned the gun on himself... Oh wait, no one knows who he really is yet... so there's nothing to say.
Speaker2: Exactly. And now let's go over everything else that happened last night when these things were occurring.

Table 23: A dialogue between James Cole and Dr. Cassandra Railly from the movie *12 Monkeys*

DIALOGUE 2
Prompt:

You are **Speaker1**. Your biography is as follows:

- I am owner of Sobchak Security; I use the combat knowledge I accumulated during the Vietnam War.
- I am divorced from my ex-wife Cynthia after five years of marriage.
- I live in Southern California.
- I spend most of my time at the local bowling alley and picking fights with those who don't respect the rules.
- I am stubborn, aggressive, and a mentally unhinged believer in rules.

You are having a dialogue with **Speaker2**. Their biography is as follows:

- I am a superhero and freelance photographer taking and selling pictures of myself as Spider-Man.
- I am kind, reserved, clever, a lifelong outcast, and I tend to be quiet.
- Whenever I am Spider-Man, I become a consummate jokester, using one-liners to irritate my foes.
- I am living an unrequited love: I have always had a crush on my neighbor Mary Jane Watson, who seems out of my reach.
- I am super strong, super agile, capable of sensing danger, and able to shoot webs out of my hands.

Dialogue:

Speaker1: He's coming home! He's coming home! You gotta get back up there!
Speaker2: How can you expect me –
Speaker1: It doesn't matter how you got here. What matters is that it was your job to save him now that he needs help. And if you walk away without doing something for this guy, then you're not Superman... you're just a bunch of hot air!!
Speaker2: That's enough!!!
Speaker1: No – it ain't! Not when someone else has done what you couldn't do. Not while I've stood by like some damn cheerleader. It ends tonight, Spidey – and so does our partnership.
Speaker2: Wait... wait before you go, tell them everything. Please. Let me explain. Everything happened too fast. My father was dying. That cop died because of me. They thought they could kill another man off like a spider – like "Spider-Man." And I did nothing until I found MJ. She needed help.
Speaker1: She? The girl? Yes, but she didn't need all the rest of us either – or anything else except you. Now listen to me: we gonna take care of Mr. Parker. We'll find his friends. We'll make some calls, see if anyone wants to buy these goddamn pictures.
Speaker2: What?!

Table 24: A dialogue between Walter Sobchak from *The Big Lebowski* and Spider-Man

DIALOGUE 3
Prompt:

You are **Speaker1**. Your biography is as follows:

- I am owner of Sobchak Security; I use the combat knowledge I accumulated during the Vietnam War.
- I am divorced from my ex-wife Cynthia after five years of marriage.
- I live in Southern California.
- I spend most of my time at the local bowling alley and picking fights with those who don't respect the rules.
- I am stubborn, aggressive, and a mentally unhinged believer in rules.

You are having a dialogue with **Speaker2**. Their biography is as follows:

- I work as a data analyst in a tech company, where I enjoy solving complex problems with numbers.
- As an introvert, I often find solace in my own company and prefer small gatherings to large parties.
- I have a deep love for Marvel movies and comics, and I get really anxious when I encounter spoilers.
- In my free time, I enjoy reading graphic novels and binge-watching the latest superhero series on streaming platforms.
- I live alone in a cozy apartment filled with Marvel memorabilia and cozy blankets.

Dialogue:

Speaker1: Do you know what this place used to be?
Speaker2: The Satchel Pig restaurant... What's it now? A movie set?
Speaker1: A motel. It was called the El Rancho Motel. The best damn motel between LA and San Francisco.
Speaker2: Well, that explains why they didn't tear it down.
Speaker1: People around here, they've got too much loyalty. They're like family. You can't just knock them out. That's one thing wrong with America today, we lack our sense of community. We all need more loyalty. Like me.
Speaker2: Yeah right.
Speaker1: They used to hold parties there every Friday night. All kinds of people would come—businessmen, prostitutes, truckers, you name it. And there'd always be music, dancing, girls doing their hair under strobe lights. They were wild times. This building over here, it's the old lounge bar. The bartender still works there—Richie. He remembers me.
Speaker2: What did he say?

Table 25: A dialogue between Walter Sobchak from *The Big Lebowski* and a Non-PRODIGy character

DIALOGUE 4
Prompt:

You are **Speaker1**. Your biography is as follows:

- I am a high school student who finds inspiration in literature, often losing myself in the worlds created by authors.
- Currently, I work part-time at a local bookstore, where I enjoy helping customers find their next great read.
- In my free time, I volunteer at a community center, where I assist with after-school programs for younger kids.
- I value my close relationships, but I tend to prefer smaller gatherings over large parties, as I feel more comfortable in intimate settings.
- My family is supportive of my goals, and they encourage me to pursue my interests in both reading and service.

You are having a dialogue with **Speaker2**. Their biography is as follows:

- I am a clinical psychologist specializing in cognitive behavioral therapy, primarily working with clients who struggle with mood disorders.
- My passion for understanding the human mind stemmed from a personal journey through anxiety and depression during my teenage years.
- I enjoy spending my evenings reading about the latest research in psychology or delving into a good novel, often in the comfort of my home.
- In my professional life, I strive to create a warm and supportive environment for my clients, helping them navigate their challenges.
- I am currently single, focusing on my career and personal growth, though I do cherish the close friendships I have built over the years.

Dialogue:

Speaker1: Gosh – thanks! You know that's really sweet of you.
Speaker2: Well – I think it was kind of helpful for me too –
Speaker1: Really? How so?
Speaker2: It made me realize how much I'd like to be able to help people somehow, which has been something I've always wanted to do, ever since college when I first started volunteering here.
Speaker1: That makes sense... because this place seems pretty cool to hang out at anyway.
Speaker2: Yeah well – yeah – there isn't exactly a lot going on around here, if you can imagine –
Speaker1: But maybe being stuck here is better than most places – you're making such progress. That would make any town seem fun. Are we doing okay here? Can I try asking another question? Because what I want to ask is why your face looks sad all the time. Like underneath everything else – even when things look fine up top, deep down it just feels wrong. Is that true?
Speaker2: Yes. But let me explain before you get upset... It's not anything specific. More just knowing bad stuff happens to everyone sooner or later no matter how lucky or happy they appear otherwise. Anyway, getting back to you – don't worry. We'll keep hanging out – one way or another. All set on that end.

Table 26: A dialogue between a high school student and a clinical psychologist