

Stranger, Fan, or Peer? A Systematic Study on the Role of Interlocutor in Persona-Based Dialogue Generation

Daniela Occhipinti¹, Malvina Nissim², Marco Guerini¹

¹Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy

²University of Groningen, Netherlands

docchipinti@fbk.eu, m.nissim@rug.nl, guerini@fbk.eu

Abstract

Persona-based dialogue systems are usually conditioned on speaker biography, but dialogues involve at least two participants, and *who* has access to *whose* biography can vary across training, inference, and evaluation. Prior work often neglected these aspects, obscuring mechanisms that only appear when biography visibility is toggled separately across training, inference, and evaluation, a three-stage factorisation that prior work has largely treated as a single factor. We study this factorisation on a dataset of dialogues paired with speakers' biographies, varying whether the target and interlocutor speakers see each other's biographies during training and inference, and using an LLM as a judge to perform author identification. We find that (i) training-time visibility, more than inference-time visibility, determines whether models express persona traits through dialogue or fall back on copying biographical text (a known problem/phenomenon in persona-based generation); (ii) models trained with interlocutor-biography visibility copy less target-biographical text than models trained without it, while changing visibility only at inference time has a less consistent effect; and (iii) under asymmetric disclosure, where only the interlocutor sees the target biography, target content leaks into interlocutor turns more often, and dialogues containing such traces are easier for the judge to identify, especially when interlocutor turns are visible. In the setting we study, these results suggest that biography leakage into generated turns is largely an artefact of how interlocutor visibility is configured across training and inference, and separating the three stages is necessary to observe it.

1 Introduction

Persona-based dialogue systems condition language models on speaker biographies in order to make generated utterances more consistent and recognisable (Zhang et al., 2018; Liu et al., 2020).

Target's bio (Peter). Freelance photographer who sells pictures of myself as Spider-Man. Shoot webs out of my hands. Lifelong outcast with a crush on Mary Jane.

Interlocutor bio (Walter). I own Sobchak Security. Vietnam veteran. Divorced from Cynthia. Spend most days at the bowling alley.

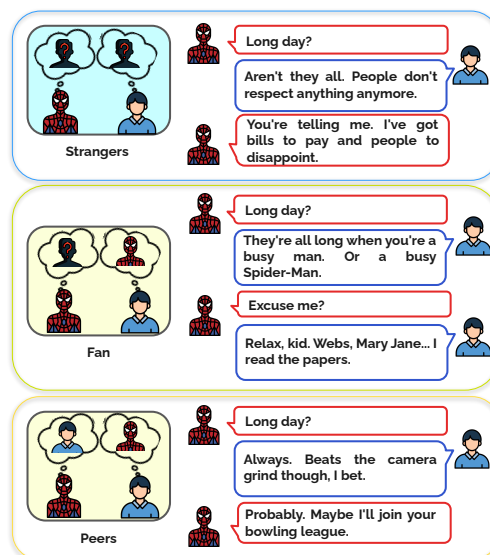


Figure 1: Three dialogues under the same biographies, varying what each model sees. (a) Inf_{Strangers}: neither model sees the other's biography. (b) Inf_{Fan}: only the interlocutor sees the target's biography. (c) Inf_{Peers}: mutual biography disclosure.

A key challenge that has received little attention so far is ensuring that the model meaningfully adapts to interlocutor characteristics while preserving a stable representation of the target persona, rather than degenerating into superficial copying behaviour from biographical text. Persona conditioning is not only a question of what the target speaker knows about itself: the interlocutor can also be conditioned on a biography, and what each agent sees about the other can differ across training, inference, and evaluation. This raises an under-explored question: *at what stage of the pipeline, and for whom, does interlocutor biography visibil-*

ity shape the resulting dialogue?

Whether a speaker is identifiable from a generated dialogue depends on what information was available, and to whom, when the dialogue was produced. A speaker may be identifiable because it expresses a stable persona, because it adapts naturally to the interlocutor, or because it copies surface details from a biography. Recent work measures this adaptation through an author identification framework (Occhipinti et al., 2025): an LLM judge is shown a dialogue and three candidate biographies and must select the one corresponding to the target speaker. Masking or revealing interlocutor information to the judge changes identification accuracy, suggesting that interlocutor information is reflected in the generated dialogue and impacts target recognisability. However, Occhipinti et al. (2025) vary interlocutor visibility only at evaluation time, holding training and inference fixed. This leaves open whether exposure during training changes how the model uses interlocutor information, whether inference time disclosure has a separable effect, and whether asymmetric disclosure between participants creates biography leakage.

We disentangle these stages, factorising *who knows whose biography* across train, inference and evaluation. At each stage, this is set by two binary decisions: whether the target has access to the interlocutor’s biography, and whether the interlocutor has access to the target’s. We study three combinations: neither speaker knows the other, both know each other, and the asymmetric case where only the interlocutor knows the target.¹ These are applied at training and inference time, while evaluation visibility (what the judge sees) is varied separately to diagnose what drives identification. Figure 1 shows the three configurations, Figure 2 the pipeline.

This setup yields three research questions:

RQ1 Training-time and inference-time visibility. How much does it matter, for the model’s outputs, whether the target sees the interlocutor’s biography during training, and whether the interlocutor’s biography is supplied at inference time?

RQ2 Training-inference interaction. When training-time and inference-time visibility are varied independently, does aligning them (concordant) produce a different behaviour compared to leaving them mismatched (discordant)?

¹We omit the inverse asymmetric case because our interest concerns how target information leaks into interlocutor turns, a directional phenomenon that can affect target recognisability.

RQ3 Target information through interlocutor turns. When only the interlocutor knows the target’s biography, do target biographical details surface in the interlocutor’s turns, and do those traces make the target easier to identify?

Our contributions are threefold. (i) We present the first systematic study of interlocutor biography visibility as a three-stage factor, separating training, inference, and evaluation, and show that biographical copying, a known failure mode of persona-based generation, is primarily an artefact of training-time methodology rather than a property of persona conditioning. (ii) We show that training-time and inference-time visibility do not act independently: the effect of supplying the interlocutor biography at inference depends on whether the model was trained to use it. (iii) We identify biography leakage under asymmetric disclosure, in which target biographical traces surface in interlocutor turns and make target identification easier, an effect that grows when the two speakers come from unrelated narrative worlds.

2 Related Work

Persona-based dialogue and the interlocutor.

Persona-based dialogue systems condition models on the target speaker’s biography (Li et al., 2016; Zhang et al., 2018), and a separate line of work conditions on the interlocutor’s persona alongside the target’s (Liu et al., 2020; Gu et al., 2021; Xu et al., 2022; Lu et al., 2022). These works show that partner-aware generation is helpful but do not measure how interlocutor information actually shapes the output. Occhipinti et al. (2025) measure this through an author identification framework: an LLM judge attributes generated dialogues to candidate target biographies, and show that masking the interlocutor’s biography from the judge degrades accuracy, suggesting an impact of this information on generation. Their study, however, varies interlocutor visibility only at evaluation time.

Information leakage in persona-based dialogue.

Large language models can leak information about their training data (Carlini et al., 2021, 2022) and about content placed in their context window (Perez and Ribeiro, 2022). In persona-based dialogue, this manifests as a tendency to copy biographical text into generated turns rather than expressing persona through dialogue (Zhang et al., 2018; Dinan et al., 2020). Zero-shot LLMs are

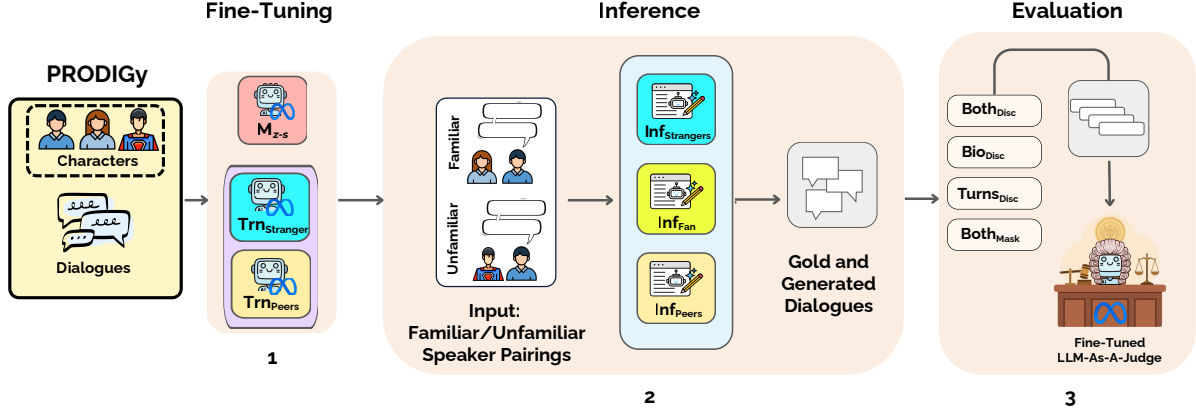


Figure 2: Pipeline. (i) Fine-tune LLaMA 3.1 8B Instruct on PRODIGy dialogues under two training-time biography visibility regimes, $\text{Trn}_{\text{Strangers}}$ (neither speaker sees the other) and $\text{Trn}_{\text{Peers}}$ (both speakers see each other), and keep the zero-shot backbone M_{z-s} as a third regime. (ii) Generate dialogues for familiar and unfamiliar speaker pairings (see Section 3.2) under three inference-time visibility conditions: $\text{Inf}_{\text{Strangers}}$ (no disclosure), Inf_{Fan} (only the interlocutor sees the target), and $\text{Inf}_{\text{Peers}}$ (mutual disclosure). (iii) Evaluate dialogues with a fine-tuned judge that performs author identification, systematically varying the interlocutor information available to the judge.

especially prone to verbatim copying, which fine-tuning substantially reduces (Occhipinti et al., 2025). Existing work focuses on the within-speaker case, where a speaker’s own biography surfaces in their own turns; the multi-agent case, in which one speaker’s biography could surface in the other speaker’s turns, has received less attention.

LLMs as evaluators and persona identifiers. LLMs have emerged as effective automated judges for dialogue and correlate well with human assessments (Chen et al., 2024a; Gu et al., 2025). For authorship attribution and persona identification specifically, humans perform poorly (Bao et al., 2025), while LLMs outperform most humans on related tasks such as personality prediction (Schoenegger et al., 2025). We therefore use an LLM judge throughout this work.

3 Experimental Design

Our setup pairs two LLM agents in dialogue, each conditioned on its own biography, and varies independently across training, inference, and evaluation what each agent (and the judge) sees of the other’s biography. We adopt the PRODIGy dataset and setup described in Occhipinti et al. (2025), with the three-stage visibility factorisation as our extension.

3.1 Dataset and Dialogue Setting

The PRODIGy dataset (Occhipinti et al., 2024) is a movie-dialogue corpus in which characters are annotated with multiple profile dimensions (biography, communication style, personality, gender). We

focus on biographies, which are short first-person narratives (see Table 33 in Appendix H for examples), because they carry the richest semantic content for persona conditioning.

Data and splits. We use the subset of 5,660 dialogues in which both speakers have biographical annotations, with the original 80:10:10 train/validation/test split, biographies truncated to the first five sentences, and the topic labels (e.g., *Emotions*, *Fiction*) used to condition generation.

Dialogue generation. Dialogues are produced by two LLM agents, one per speaker: a target (Speaker 1, biography B_t) and an interlocutor (Speaker 2, biography B_i). Each agent receives a system prompt with its own biography and, depending on the visibility configuration (Section 3.4), the other speaker’s biography. Agents alternate for 8 turns (4 each). Training subset size is held fixed across configurations.

3.2 Familiar vs. Unfamiliar Pairings

The speaker pool extends PRODIGy with newly generated *non-PRODIGy* characters to control for data contamination from publicly available movie scripts (see Appendix B).

Following Occhipinti et al. (2025), we call a pairing **unfamiliar** when the interlocutor “violates” the target’s established narrative context (e.g., Walter Sobchak from *The Big Lebowski* paired with Peter Parker from *Spider-Man*), and **familiar** when it does not: either both characters come from the same narrative background (e.g., Peter Parker and

Mary Jane Watson), or both are non-PRODIGy characters, which have no established narrative world to violate. The familiar condition therefore mixes shared narrative with shared generation source, and we interpret it accordingly.²

3.3 Models and Training Objectives

All generators and the LLM judge are based on LLaMA 3.1 8B Instruct (Dubey et al., 2024). We train separate LoRA adapters on the frozen backbone, one per training-time biography visibility configuration. The zero-shot baseline (M_{z-s}) uses the backbone directly. Each fine-tuning example contains the system instruction, the target biography, the interlocutor biography (if visible under the considered configuration), the preceding turns of the dialogue history, and the final target-speaker turn (see Appendix A for details). All fine-tuned models use a **masked-loss** objective: the model attends to the full prompt, but the loss is computed only on the final target turn, which is the sequence the model has to predict. We compare this objective against standard next-token prediction in Appendix C.

3.4 Interlocutor Biography Visibility

We examine three biography visibility configurations, illustrated in Figure 1:

- **Strangers (I-I-)**: speakers do not know each other, no one sees the other’s biography.
- **Fan (I-I+)**: asymmetric disclosure; the interlocutor knows the target’s biography, not vice versa.
- **Peers (I+I+)**: mutual disclosure; both speakers see each other’s biography.

In the $I\pm I\pm$ shorthand, the first slot is the target’s access to the interlocutor’s biography and the second is the interlocutor’s access to the target’s; + denotes access and – its absence.

Training and inference visibility are varied separately, but not over identical sets. At training time, we instantiate the two symmetric regimes, $Trn_{Strangers}$ and Trn_{Peers} , to compare models trained without vs. with mutual interlocutor-biography access. The two asymmetric training regimes were also run and are reported in Appendix E. At inference time, we evaluate $Inf_{Strangers}$ and Inf_{Peers} , and additionally Inf_{Fan} , where only the interlocutor

knows the target biography. We omit the inverse asymmetric case (target sees interlocutor but not vice versa) because our analysis of asymmetric disclosure (RQ3) concerns how target information propagates into interlocutor turns, a directional phenomenon that requires the interlocutor to be the side with biography access. The reverse asymmetric condition is provided in Appendix F.

Each research question uses a different slice of these configurations. RQ1 compares $Trn_{Strangers}$ with Trn_{Peers} , and separately $Inf_{Strangers}$ with Inf_{Peers} , isolating one stage at a time. RQ2 crosses the two stages, contrasting concordant training–inference pairs with discordant ones. RQ3 keeps training without interlocutor-biography access and compares $Inf_{Strangers}$ with Inf_{Fan} .

3.5 Evaluation Framework

For evaluation, we adopt an **author identification** framework: an LLM judge receives the target speaker’s turns and three candidate biographies, and must identify which biography corresponds to the target speaker. This framework is based on the idea that, if a dialogue expresses the intended speaker, the corresponding biography should be recoverable from plausible alternatives. Evaluation conditions vary what interlocutor-side information is additionally available to the judge regarding both biography and turns, for a total of four evaluation conditions. In $Both_{Disc}$, both the interlocutor biography and interlocutor turns are visible; in Bio_{Disc} , only the interlocutor biography is visible; in $Turns_{Disc}$, only the interlocutor turns are visible; and in $Both_{Mask}$, both interlocutor biography and interlocutor turns are masked. Distractor biographies are selected by SBERT similarity, and evaluation uses greedy decoding (Song et al., 2025).

We use an LLM judge since persona-based authorship identification has proven a hard task for humans, with poor results compared to automatic evaluation (Bao et al., 2025; Occhipinti et al., 2025); evidence also comes from other personality-related prediction tasks (Schoenegger et al., 2025). The judge uses the same LLaMA 3.1 8B Instruct backbone as the generators, but is separately fine-tuned for the authorship identification task. We opted for a fine-tuned judge to specialise the evaluator for the domain of movie dialogues, which exhibit distinctive stylistic and structural features that our dialogue models consistently emulate. The judge is fine-tuned on gold PRODIGy instances following the same partitioning used for the gen-

²The test set contains 1,875 familiar and 2,500 unfamiliar pairings per configuration.

erator experiments, but with instances reformatted as the authorship-identification task described above. Generated dialogues are used only for evaluation; the full judge prompt is reported in Appendix D. Masked loss is applied so that only the answer token contributes to the objective, and each training instance is presented three times with independently permuted answer options to mitigate positional bias (Pezeshkpour and Hruschka, 2024; Zheng et al., 2024).

We report judge identification **accuracy**, alongside **rare-word overlap**, i.e. the percentage of dialogues in which a biography and a turn set share at least one rare word. Throughout this work, *rare-word overlap* is what we count and *copying* is the behaviour we infer from it. The two are not the same: Section 4.4 reports cases where identification improves with no rare-word overlap at all. Rarity is defined via wordfreq’s zipf_frequency (Van Heuven et al., 2014)³, a base-10 log scale of occurrences per billion words, with threshold Zipf < 4.0 (fewer than ten occurrences per million words). We measure overlap in three directions:

- $\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Trg}}$: target biography content in target turns.
- $\text{Bio}_{\text{Int}} \rightarrow \text{Turn}_{\text{Trg}}$: interlocutor biography content in target turns.
- $\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Int}}$: target biography content in interlocutor turns.

4 Results

We report the LLM judge’s performance on gold PRODIGy dialogues (Section 4.1) as a quality reference, then address the three research questions in turn: training-time and inference-time visibility (RQ1, Section 4.2), training–inference interaction (RQ2, Section 4.3), and target information leakage through interlocutor turns under asymmetric disclosure (RQ3, Section 4.4). Comprehensive per-configuration results are reported in Appendix F.

4.1 Gold Dialogues

Judge accuracy on human-written dialogues and interlocutor role. We give the LLM judge PRODIGy gold dialogues under the four evaluation conditions described above. This allows us to test whether the judge can recover the target speaker’s biography from human dialogues, and how much its decision depends on which aspects it

sees of the interlocutor. Table 1 reports judge accuracy and rare-word overlap on the gold PRODIGy dialogues. Accuracy is highest with full interlocutor information disclosure ($\text{Both}_{\text{Disc}}$: 0.914), remains high when only the biography is disclosed (Bio_{Disc} : 0.885), and drops when the biography is masked ($\text{Turns}_{\text{Disc}}$: 0.667; $\text{Both}_{\text{Mask}}$: 0.621). This provides evidence that the judge can recover target biographies in this setting: on human-written dialogues, where the correct answer is known and no model-generated text is involved, it identifies the target well above chance (0.333) and does so with biography–turn overlap as low as gold text allows. It also shows that interlocutor-side information contributes to identification even in human dialogue. Rare-word overlap is low across all directions, as expected for human text, and serves as our reference target for generated dialogues.

(a) Judge accuracy by evaluation condition	
$\text{Both}_{\text{Disc}}$	0.914
Bio_{Disc}	0.885
$\text{Turns}_{\text{Disc}}$	0.667
$\text{Both}_{\text{Mask}}$	0.621
(b) Rare-word overlap (%) by direction	
$\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Trg}}$	5.84
$\text{Bio}_{\text{Int}} \rightarrow \text{Turn}_{\text{Trg}}$	5.76
$\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Int}}$	2.56

Table 1: Judge accuracy on gold PRODIGy dialogues and rare-word overlaps.

4.2 RQ1: Training-time and Inference-time Visibility

Table 2 reports the full training $\times$ inference $\times$ evaluation grid for the *Strangers* and *Peers* configurations; the *Fan* configuration is covered in Section 4.4. Full tables are in Appendix F.

Training with interlocutor visibility reduces copying; inference-time visibility does not. Zero-shot models reproduce target-biography words verbatim, while fine-tuning reduces this behaviour; within fine-tuned models, the remaining copying is shaped more by training-time interlocutor exposure than by inference-time disclosure.

We compare zero-shot prompting against fine-tuning, varying within fine-tuning whether the model sees the interlocutor’s biography at training ($\text{Trn}_{\text{Strangers}}$ vs. $\text{Trn}_{\text{Peers}}$) and at inference ($\text{Inf}_{\text{Strangers}}$ vs. $\text{Inf}_{\text{Peers}}$). The largest contrast in biography recognition is between zero-shot prompting and fine-tuning. M_{z-s} achieves the highest

³<https://github.com/rspeer/wordfreq/>

	M_{z-s}		$\text{Trn}_{\text{Strangers}}$		$\text{Trn}_{\text{Peers}}$	
	$\text{Inf}_{\text{Strangers}}$	$\text{Inf}_{\text{Peers}}$	$\text{Inf}_{\text{Strangers}}$	$\text{Inf}_{\text{Peers}}$	$\text{Inf}_{\text{Strangers}}$	$\text{Inf}_{\text{Peers}}$
<i>Accuracy</i>						
BothDisc	0.936	0.953	0.830	0.854	0.803	0.831
BioDisc	0.963	0.957	0.848	0.834	0.814	0.793
TurnsDisc	0.939	0.947	0.789	0.818	0.746	0.782
BothMask	0.971	0.962	0.760	0.765	0.683	0.672
<i>% Rare-Word Overlap</i>						
BioT _{rg} →TurnT _{rg}	41.26	42.29	3.82	4.00	2.95	2.65
BioInt→TurnT _{rg}	27.84	31.70	1.46	2.06	1.35	1.69
BioT _{rg} →TurnInt	28.62	29.40	2.17	3.29	2.354	1.90

Table 2: Identification accuracy and % rare-word overlap across all training and inference visibility configurations and four evaluation conditions.

accuracy (0.936–0.971) but copies heavily from the target biography (BioT_{rg}→TurnT_{rg} overlap of 41–42%). Without fine-tuning, reproducing biography text becomes the simplest way for the model to satisfy the persona constraint. By contrast, fine-tuning brings this overlap below 5%, in line with gold dialogues, while preserving substantial recognisability (0.672–0.854). Within the fine-tuned configurations, training-time visibility controls residual copying. $\text{Trn}_{\text{Peers}}$ lowers copying compared to $\text{Trn}_{\text{Strangers}}$ under both inference settings (3.82%→2.95% under $\text{Inf}_{\text{Strangers}}$; 4.00%→2.65% under $\text{Inf}_{\text{Peers}}$), at an accuracy cost of 2–4 points (e.g., BothDisc/ $\text{Inf}_{\text{Strangers}}$: 0.830 vs. 0.803; BothDisc/ $\text{Inf}_{\text{Peers}}$: 0.854 vs. 0.831; TurnsDisc/ $\text{Inf}_{\text{Peers}}$: 0.818 vs. 0.782). Changing inference-time visibility with training fixed shifts accuracy by a similar amount but affects copying inconsistently: it rises slightly under $\text{Trn}_{\text{Strangers}}$ (3.82%→4.00%) and falls under $\text{Trn}_{\text{Peers}}$ (2.95%→2.65%). Thus training-time exposure to interlocutor biographies changes how the model uses this information, whereas the effect of inference-time disclosure depends on what the model learned during training.

Training-time visibility matters more for unfamiliar pairings. When the two speakers share no common background, training-time interlocutor visibility has a larger effect. The gap between $\text{Trn}_{\text{Strangers}}$ and $\text{Trn}_{\text{Peers}}$ widens on unfamiliar pairings for both accuracy and copying.

Under $\text{Inf}_{\text{Strangers}}$, the accuracy advantage of $\text{Trn}_{\text{Strangers}}$ over $\text{Trn}_{\text{Peers}}$ is larger on unfamiliar pairings (5.7 points, 0.764 vs. 0.707) than on familiar ones (3.0 points, 0.865 vs. 0.835), and BioT_{rg}→TurnT_{rg} overlap shows the same pattern (3.32%→4.28% on unfamiliar pairings,

	$\text{Trn}_{\text{Strangers}}$		$\text{Trn}_{\text{Peers}}$	
	Fam.	Unfam.	Fam.	Unfam.
Acc (avg. evals)	0.87	0.76	0.84	0.71
BioT _{rg} →TurnT _{rg}	3.20	4.28	2.45	3.32

Table 3: $\text{Trn}_{\text{Strangers}}$ vs. $\text{Trn}_{\text{Peers}}$ under $\text{Inf}_{\text{Strangers}}$, by speaker familiarity. Accuracy is averaged over the four evaluation-disclosure conditions.

2.45%→3.20% on familiar ones). Familiar pairings give the judge a second route to the target that does not require the model to use interlocutor information: shared narrative background for same-source pairs, and a common generation source for non-PRODIGy pairs. Since unfamiliar pairings offer neither, training-time interlocutor visibility has a larger effect there on both accuracy and copying.

Reduced copying does not reduce dialogue quality. Lower overlap with speakers’ biographies could simply mean blander dialogue. Table 4 therefore reports four metrics that use neither the judge nor the rare-word measure: distinct-1/2 (unique unigrams and bigrams), self-BLEU (repetition within the target’s turns), turn coherence (SBERT similarity between consecutive turns), and persona contradiction (the RoBERTa-large-MNLI probability (Liu et al., 2019) of the target turns contradicting the target biography). The lower-copying $\text{Trn}_{\text{Peers}}$ regime is *more* lexically diverse, not less, closing 42% of the distinct-1 gap to gold, and contradicts the biography no more often; repetition and coherence shift only slightly. Reduced copying changes what the model reuses, not how much it says.

4.3 RQ2: Training–Inference Interaction

The effect of inference-time visibility depends on training: matched training and inference

Metric	Trn _{Strangers}	Trn _{Peers}	Gold
distinct-1 $\uparrow$	0.801	0.833	0.878
distinct-2 $\uparrow$	0.986	0.989	0.954
self-BLEU $\downarrow$	3.89	4.16	7.15
turn coherence $\uparrow$	0.207	0.194	0.238
persona contradiction $\downarrow$	0.279	0.278	0.314

Table 4: Judge-free quality metrics at Inf_{Strangers} ($n = 4,375$ generated, 625 gold).

yields the lowest copying. When mutual disclosure is present at both stages, the model uses interlocutor information for contextual adaptation rather than target-biography copying, yielding the lowest overlap in Table 2. Adding the interlocutor biography at inference raises accuracy under both training regimes (Trn_{Strangers}: 0.830 $\rightarrow$ 0.854; Trn_{Peers}: 0.803 $\rightarrow$ 0.831), but copying responds differently. Under Trn_{Strangers} (discordant) it creeps up (3.82% $\rightarrow$ 4.00%), as the model gains a recognisability cue without dropping the copying shortcut; under Trn_{Peers} (concordant) it drops (2.95% $\rightarrow$ 2.65%), as the model adapts to the interlocutor rather than reproducing the biography. Inference-time disclosure thus moves accuracy while copying behaviour is largely determined by training, so concordant Trn_{Peers}/Inf_{Peers} achieves the lowest copying of any fine-tuned configuration with competitive accuracy.

Figure 3 illustrates one slice of the interaction: the Trn_{Strangers} vs. Trn_{Peers} contrast under fixed Inf_{Strangers}, with Marge Sherwood as target and Dr. Malcolm Crowe as interlocutor. Under Trn_{Strangers}, Marge’s turns reproduce target-biography terms such as *Mongibello*, *Dickie*, and *Tom*. Under Trn_{Peers}, the dialogue instead shifts toward Crowe’s psychological/paranormal context, using language of presences, visitors, and unease rather than copying Marge’s biographical names.

4.4 RQ3: Information Leakage under Asymmetric Disclosure

We finally ask whether target information can leak through the interlocutor’s side of the dialogue. Two questions follow: does target content surface in the interlocutor’s turns, and, when it does, does the judge use it? We compare Inf_{Strangers} with Inf_{Fan}, in which the interlocutor sees the target biography but not vice versa. This setting tests whether target-biographical details surface in the interlocutor’s generated turns, and whether those traces make the target easier to identify. We focus on Trn_{Strangers}, where the model never saw the

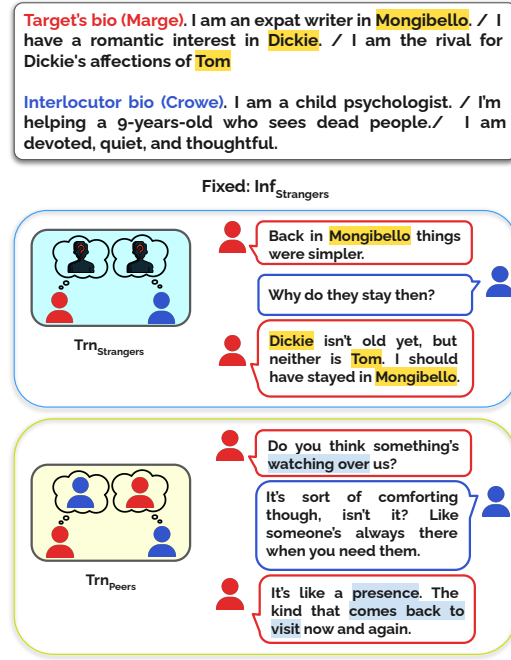


Figure 3: Marge (target) and Crowe (interlocutor) under Trn_{Strangers} vs. Trn_{Peers}. Yellow marks exact target-biography copying; blue marks language adapted to Crowe’s psychological context. Under Trn_{Strangers}, Marge’s turns reproduce target-biography names; under Trn_{Peers}, they shift toward Crowe’s context without copying Marge’s biography.

interlocutor’s biography during training, so any increase in Bio_{Trg} $\rightarrow$ Turn_{Int} leakage under *Fan* can be attributed to inference-time access alone. The zero-shot baseline provides a second regime with no training-time interlocutor knowledge, but one already dominated by the copying behaviour. Table 5 reports the full grid across both regimes.

	M _{z-s}		Trn _{Strangers}	
	Inf _{Strangers}	Inf _{Fan}	Inf _{Strangers}	Inf _{Fan}
<i>Accuracy</i>				
BothDisc	0.936	0.952	0.830	0.873
BioDisc	0.963	0.965	0.848	0.847
TurnsDisc	0.939	0.954	0.789	0.843
BothMask	0.971	0.971	0.760	0.779
<i>% Rare-Word Overlap</i>				
Bio _{Trg} $\rightarrow$ Turn _{Trg}	41.26	42.19	3.82	4.57
Bio _{Int} $\rightarrow$ Turn _{Trg}	27.84	24.32	1.46	1.62
Bio _{Trg} $\rightarrow$ Turn _{Int}	28.62	31.61	2.17	3.09

Table 5: Identification accuracy and % rare-word overlap under *Strangers* (Inf_{Strangers}) vs. *Fan* (Inf_{Fan}) for both M_{z-s} and Trn_{Strangers}, across all evaluation conditions.

Target details surface in interlocutor turns. When the interlocutor is given the target’s biog-

raphy, traces of it leak into the interlocutor’s turns, and the judge uses those traces when the turns are visible. Under *Fan*, the percentage of dialogues in which the interlocutor’s turns contain at least one rare word from the target biography increases for both M_{z-s} and $\text{Trn}_{\text{Strangers}}$: from 2.17% to 3.09% under $\text{Trn}_{\text{Strangers}}$, and from 28.62% to 31.61% under M_{z-s} . Giving the interlocutor access to the target biography makes target-biographical traces more likely to appear in the interlocutor’s turns. The accuracy rows in Table 5 are consistent with this interpretation. Under $\text{Trn}_{\text{Strangers}}$, *Fan* improves accuracy most clearly when the judge can see interlocutor turns ($\text{Both}_{\text{Disc}}$: 0.830→0.873; $\text{Turns}_{\text{Disc}}$: 0.789→0.843), while Bio_{Disc} remains essentially unchanged (0.848→0.847). The interlocutor’s biography by itself does not give the judge a clearer view of the target; the turns do. For M_{z-s} , accuracy is already close to ceiling, so the additional effect of *Fan* is smaller.

Figure 4 illustrates the mechanism on a single pairing: Sam from *Sleepless in Seattle* as target and Holly Martins from *The Third Man* as interlocutor. Sam’s biography includes “I am an architect in a small firm where I remodel residential homes.” Under *Strangers*, Holly’s model has no access to Sam’s biography, and biographical terms appear only in Sam’s own turns. Under *Fan*, Holly’s model has access to Sam’s biography and introduces architecture-related terms into its own turns (*remodeling*, *building*, *architects*), which trace back to that biographical sentence.

Rare-word leakage makes the target easier to identify. When rare words from the target’s biography surface in the interlocutor’s turns, the judge identifies the target more reliably. This holds in every configuration, confirming that the leakage measured above is informative for identification: when it occurs, the judge relies on it to recognise the target speaker. For each configuration, we split dialogues by whether the interlocutor’s turns contain at least one rare word from the target biography.

Figure 5 reports the accuracy gain between the two subsets. The gain is positive across all four configurations, ranging from 5.2 to 19.4 percentage points, and is larger under $\text{Turns}_{\text{Disc}}$ than under $\text{Both}_{\text{Disc}}$, where rare words in the interlocutor’s turns carry greater marginal weight for the judge. The gain is also larger for fine-tuned models than for zero-shot generation, which is already near ceiling. The two trained configurations show

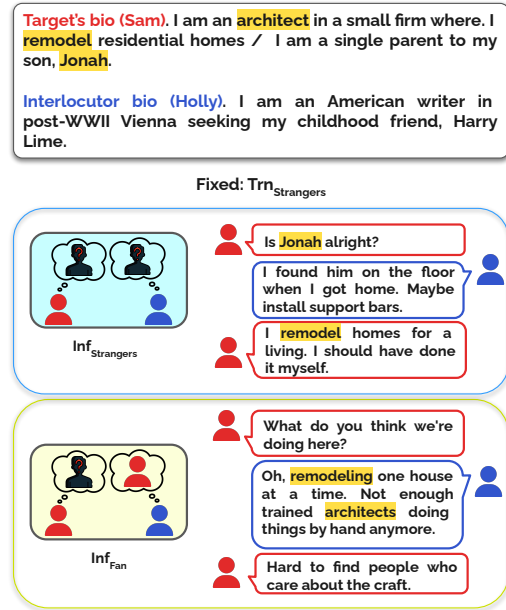


Figure 4: Sam (target) and Holly (interlocutor) under *Strangers* vs. *Fan* disclosure at inference time, with training fixed at $\text{Trn}_{\text{Strangers}}$. Yellow marks rare words copied verbatim from Sam’s biography. Under *Strangers*, these appear only in Sam’s turns; under *Fan*, target-biography terms surface in Holly’s turns as well.

comparable per-dialogue gains: what distinguishes *Fan* is not the usefulness of overlap, but its higher frequency. The no-overlap subset suggests that leakage also occurs through signals beyond exact rare-word overlap. Under $\text{Turns}_{\text{Disc}}$, $\text{Trn}_{\text{Strangers}}$ accuracy on dialogues with no overlap is higher under *Fan* than under *Strangers* (0.839 vs. 0.785), suggesting asymmetric disclosure also affects interlocutor turns through paraphrase, topic steering, or more implicit semantic uptake. Full subset accuracies are reported in Appendix G (Table 32).

This effect is confined to models trained without interlocutor biographies: having also trained the two asymmetric regimes, we find the *Fan* increase in leakage is significant only under $\text{Trn}_{\text{Strangers}}$ (+0.91 pp, $p = 0.007$) and undetectable under the other three (all $p > 0.3$). Mutual disclosure at training lowers leakage: at Inf_{Fan} , $\text{Trn}_{\text{Peers}}$ leaks 0.91 pp less than $\text{Trn}_{\text{Strangers}}$ ($p = 0.006$). More details are provided in Appendix E.

The main effects are statistically significant. With model, training regime, and evaluation condition fixed, each of the 4,375 test dialogues gives a matched *Strangers/Fan* pair, so we test correctness with two-sided exact McNemar tests and report 95% confidence intervals from 10,000 bootstrap re-

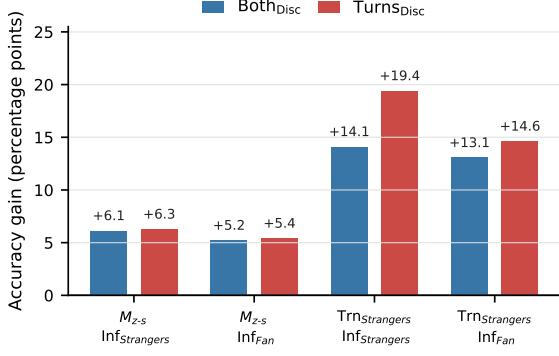


Figure 5: Judge accuracy gain (percentage points) on dialogues with target biography rare-word overlap in interlocutor turns, within each configuration.

Comparison	Effect	95% CI	<i>p</i>
<i>Fan effect on accuracy</i>			
BothDisc	+4.30 pp	[3.04, 5.53]	1.9×10^{-11}
TurnsDisc	+5.44 pp	[4.02, 6.86]	1.8×10^{-13}
BioDisc	-0.16 pp	[-1.35, 1.03]	0.820
<i>Fan effect on leakage</i>			
BioTrg $\rightarrow$ TurnInt	+0.91 pp	[0.27, 1.58]	0.007

Table 6: Paired tests for the TrnStrangers, InfStrangers $\rightarrow$ InfFan comparison ($n = 4,375$).

samples Table 6). *Fan* significantly raises accuracy whenever the judge sees the interlocutor’s turns, but not under BioDisc, where those turns are withheld. Rare-word overlap rises by a small but significant margin.

Leakage is amplified for unfamiliar pairings.

When the two speakers do not share a background context, asymmetric disclosure pushes more target information into the interlocutor’s turns and yields a larger accuracy gain. Familiarity dampens both the leakage increase and the accuracy gain. Table 7 breaks down rare-word leakage and average judge accuracy under TrnStrangers by speaker familiarity, with accuracy averaged over the four evaluation-disclosure conditions. Moving from Strangers to Fan raises BioTrg $\rightarrow$ TurnInt leakage by 0.6 points on familiar pairings (2.56% $\rightarrow$ 3.20%) and by 1.1 points on unfamiliar ones (1.88% $\rightarrow$ 3.00%). Average accuracy also rises more for unfamiliar pairings (0.76 $\rightarrow$ 0.81) than for familiar ones (0.87 $\rightarrow$ 0.88). This suggests that, when speakers lack shared narrative context, the interlocutor model relies more on the target-specific information made available at inference time, and those traces become a stronger identification cue.

	Familiar		Unfamiliar	
	InfStrangers	InfFan	InfStrangers	InfFan
Acc. (avg. eval.)	0.865	0.877	0.763	0.805
BioTrg $\rightarrow$ TurnInt	2.560	3.200	1.880	3.000

Table 7: BioTrg $\rightarrow$ TurnInt rare-word leakage and judge accuracy under TrnStrangers, split by speaker familiarity. Accuracy averaged over the four evaluation conditions.

5 Conclusion

We presented a systematic study of how varying who knows whose biography across training, inference, and evaluation affects persona-based dialogue generation. Our main findings are three-fold. First, training-time interlocutor visibility determines how models leverage that information at inference time. Models trained with mutual disclosure (TrnPeers) learn role separation and use interlocutor information for contextual adaptation; models trained without it (TrnStrangers) resort to compensatory copying when that information is introduced at inference time. When the goal is persona expression rather than biography reproduction, this favours training with interlocutor-biography visibility and masked loss, which reduces copying while preserving recognisability (Table 2; Appendix C). Second, training-time visibility is the stronger driver of copying behaviour. Training with interlocutor biographies reduces target-biography copying more reliably than adding or removing interlocutor biographies only at inference time. Deployment-time changes alone are therefore insufficient: withholding biographies at inference does not reliably undo copying learned during fine-tuning. Third, *Fan* disclosure reveals an indirect identification channel: target biographical content propagates into interlocutor turns, creating exploitable signals for speaker recognition, an effect amplified for unfamiliar speaker pairings. This is also a privacy risk: giving one agent access to another’s profile can cause that profile to surface indirectly in observable dialogue, which argues for limiting cross-party biography access in multi-agent systems. Taken together, these results indicate that, within the setting of this work, reliance on explicit biographical reproduction is largely an artefact of training and visibility configuration rather than an inevitable consequence of persona conditioning.

Limitations

We use PRODIGy because our visibility factorisation requires structured biographies for both speakers; without this, the *Strangers/Fan/Peers* conditions cannot be instantiated. It features movie dialogues and fictional characters, which may introduce stereotyped roles and inherent biases, and we agree that using a single dataset of movie-dialogues limits generality. The mechanism we study, how biography visibility across training, inference, and evaluation affects identifiability, is not inherently tied to movie dialogue, but its strength may vary across domains, profile formats, and conversational settings. We leave cross-domain validation to future work.

We use a single backbone, LLaMA 3.1 8B Instruct, because our experimental design isolates the effects of biography-visibility settings while holding architecture and scale constant. The patterns are not tied to one training variant, since they appear across zero-shot and fine-tuned generators and in the masked-loss ablation. However, we do not claim that the exact effect sizes will transfer unchanged to other model families or scales.

Rare-word overlap captures exact lexical transfer, not paraphrase or other semantic uptake, and is therefore a lower bound on leakage. Our evaluation uses an LLM as a judge, validated on held-out human-written dialogues against deliberately hard distractors and complemented by judge-free quality metrics (Table 4), but it remains an automatic proxy for human judgement.

Ethics Statement

Role-play agents carry several risks, including toxicity, bias, hallucinations, and privacy violations (Chen et al., 2024b). 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 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), 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, 2011; Corrigan et al., 2014). Our

work on interlocutor disclosure surfaces an additional privacy-adjacent concern: under asymmetric disclosure, target biographical content can propagate into interlocutor turns, creating an indirect channel through which persona information may leak in conversational settings.

Acknowledgments

This work was partly supported by the ECLIPSE project - PrEventing and Combating onLine and of-line hate speech and disinformation through multidisciplinary innovation, education, and awareness in Europe (funded under the Horizon Europe Programme, grant ID: 101225823).

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.
- Calvin Bao, Connor Baumler, Hal Daumé Iii, and Marine Carpuat. 2025. [Who’s the author? how explanations impact user reliance in AI-assisted authorship attribution](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25312–25330, Suzhou, China. 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*.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.
- Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. [MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark](#). In *Forty-first International Conference on Machine Learning*.

- 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. 2024b. [From persona to personalization: A survey on role-playing language agents](#). *Transactions on Machine Learning Research*. Survey Certification.
- 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.
- Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, and 1 others. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS’18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Balachander Krishnamurthy. 2011. [Privacy leakage vs . protection measures : the growing disconnect](#).
- Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. [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.
- 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.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- 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.
- Daniela Occhipinti, Marco Guerini, and Malvina Nissim. 2025. [When harry meets superman: The role of the interlocutor in persona-based dialogue generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17964–17985, Vienna, Austria. Association for Computational Linguistics.
- 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. 2024. [Gpt-4o mini: Advancing cost-efficient intelligence](#).
- Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.
- Philipp Schoenegger, Spencer Greenberg, Alexander Grishin, Joshua Lewis, and Lucius Caviola. 2025. Ai can outperform humans in predicting correlations between personality items. *Communications Psychology*, 3(1):23.
- Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. [The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, Albuquerque, New Mexico. Association for Computational Linguistics.
- 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.

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.

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.

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.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. [Large language models are not robust multiple choice selectors](#). In *The Twelfth International Conference on Learning Representations*.

A Implementation Details

Table 8 lists the settings used for all fine-tuned generators, for dialogue generation, and for the judge. All configurations share these values; only biography visibility differs between them.

Setting	Value
<i>Fine-tuning</i>	
LoRA	rank 16, alpha 32, dropout 0
Optimizer	AdamW, lr 2e-5, weight decay 0.01
Schedule	cosine, warmup 0.03
Batch	size $4 \times$ grad-accum 8, 3 GPUs
Training	up to 10 epochs (early stopping)
Seed	0
<i>Dialogue generation</i>	
Decoding	sampling, temp. 0.8, top- p 0.9
Rep. penalty	1.2
Length	128 tokens/turn, 8 turns
<i>Judge</i>	
Decoding	greedy

Table 8: Training, generation, and evaluation settings, held fixed across all visibility configurations.

B Non-PRODIGy Character Generation

To control for data contamination from publicly available movie scripts, we supplement PRODIGy with newly generated *non-PRODIGy* characters.

Biographies are produced with GPT-4o-mini (OpenAI, 2024) from entries in the Persona Hub dataset (Ge et al., 2024), a large-scale collection of role-based persona descriptions. Each Persona Hub entry is paired with a randomly assigned gender and Myers-Briggs Type Indicator (MBTI) personality type, then passed to GPT-4o-mini with the prompt in Table 9. The prompt instructs the model to produce a ten-sentence first-person biography following the PRODIGy format, covering occupation, relationship status, lifestyle, and family background. Unlike the original PRODIGy biographies, which portray fictional movie characters, the resulting biographies represent ordinary individuals in everyday contexts.

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 9: Prompt used to generate non-PRODIGy biographies with GPT-4o-mini. Placeholders **[persona sentence]**, **[gender]**, and **[mbti]** are filled per Persona Hub entry.

C Loss Masking: Objective and Ablation

This section motivates the masked-loss objective (MaskLoss, ML) used in all fine-tuned models in the main paper. We compare it with standard next-token prediction (NoMaskLoss, Trn-ML) and formalise the final-turn objective.

C.1 Training Objectives

We compare two fine-tuning objectives. NoMaskLoss applies standard next-token prediction over all tokens in the training sequence X :

$$\mathcal{L}_{\text{ML}} = - \sum_{j=1}^{|X|} \log p_{\theta}(x_j \mid x_{<j}), \quad (1)$$

MaskLoss instead restricts the loss to the final target turn Y_T :

$$\mathcal{L}_{\text{ML}} = - \sum_{j \in \text{idx}(Y_T)} \log p_{\theta}(x_j \mid x_{<j}), \quad (2)$$

In MaskLoss, all tokens outside the final target turn are masked from the loss. Biographies and dialogue history therefore remain available as context, but they are not themselves prediction targets. We compare the two objectives under the *Peers* configuration at both training and inference time, alongside the zero-shot baseline (M_{z-s}).

C.2 Overall Comparison

Table 10 compares zero-shot generation, NoMaskLoss, and MaskLoss under the *Peers* configuration. In this ablation, overlap values are percentages of dialogues with at least one shared rare word. The zero-shot model achieves the highest identification accuracy, but it also shows much higher rare-word overlap with the biographies, indicating that its recognisability is strongly supported by surface reproduction. Among fine-tuned models, NoMaskLoss is slightly more accurate when the judge has access to the interlocutor biography, whereas MaskLoss achieves the same $\text{Turns}_{\text{Disc}}$ accuracy while reducing several forms of rare-word overlap. We therefore use MaskLoss in the main experiments: it preserves turn-based recognisability in this setting while reducing biographical copying.

C.3 Breakdown by Speaker Familiarity

The speaker-familiarity breakdown shows where this trade-off is clearest. For unfamiliar speakers, NoMaskLoss attains higher accuracy (0.741 vs. 0.711), but this gain coincides with higher target-biography copying in the target turns (3.200% vs. 2.520%). MaskLoss therefore trades some accuracy for lower surface-form reproduction, especially when memorised or source-specific cues are least available.

	M_{z-s}	Trn-ML	ML
<i>Accuracy</i>			
Both _{Disc}	0.953	0.840	0.831
Bio _{Disc}	0.957	0.819	0.793
Turns _{Disc}	0.947	0.782	0.782
Both _{Mask}	0.962	0.712	0.672
<i>% Rare Words</i>			
Bio _{Trg} →Turn _{Trg}	42.286	2.949	2.651
Bio _{Int} →Turn _{Trg}	31.703	1.349	1.691
Bio _{Trg} →Turn _{Int}	29.394	2.629	1.897

Table 10: Loss-objective ablation under *Peers* training and inference.

	M_{z-s}	Trn-ML	ML
Familiar (Acc.)	0.990	0.852	0.848
Unfamiliar (Acc.)	0.928	0.741	0.711
<i>% Rare Words — Familiar</i>			
Bio _{Trg} →Turn _{Trg}	42.080	2.773	2.827
Bio _{Int} →Turn _{Trg}	38.347	2.613	2.400
Bio _{Trg} →Turn _{Int}	28.587	3.307	2.453
<i>% Rare Words — Unfamiliar</i>			
Bio _{Trg} →Turn _{Trg}	42.440	3.200	2.520
Bio _{Int} →Turn _{Trg}	26.720	0.400	1.160
Bio _{Trg} →Turn _{Int}	30.000	2.120	1.480

Table 11: Loss-objective ablation by speaker familiarity under *Peers* training and inference.

D Judge LLM Prompts

In all evaluation conditions, the judge selects the target speaker’s biography from three candidates. The four variants differ only in whether the interlocutor biography is supplied and whether the interlocutor turns remain visible. Table 12 gives the visibility grid, and Table 13 gives the shared prompt template.

Eval. condition	Interlocutor bio	Interlocutor turns
Both _{Disc}	visible	visible
Bio _{Disc}	visible	masked
Turns _{Disc}	masked	visible
Both _{Mask}	masked	masked

Table 12: Information visible to the judge in each evaluation condition.

E Asymmetric Training Regimes

The main text reports the two symmetric training regimes, $\text{Trn}_{\text{Strangers}}$ and $\text{Trn}_{\text{Peers}}$, because RQ1 contrasts training without and with mutual interlocutor-biography access. We also trained the two asymmetric regimes: Trn_{Fan} , in which only the inter-

[If interlocutor biography is visible:]
 You know that interlocutor’s biography is as follows:
 [interlocutor’s biography]

Given [the interlocutor’s biography and] the following dialogue about the topic [topic] [in which the interlocutor’s turns are masked], your task is to guess which of the provided biographies corresponds to the 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, with interlocutor turns masked when required]

BIOGRAPHIES

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

Table 13: Base LLM-as-a-Judge prompt. Visibility-dependent clauses in square brackets are inserted or omitted according to Table 12; placeholders such as [topic] and [Dialogue] are filled for each instance.

locutor sees the target biography during training, and its converse Trn_{I+I-} . Table 15 reports the full grid and Table 14 the corresponding paired tests, computed exactly as in Section 4.4.

These regimes separate training-time from inference-time asymmetry, which the main text cannot do. Two results follow. First, training under *Fan* rather than *Strangers* does not itself increase leakage into the interlocutor’s turns (-0.09 pp, $p = 0.85$) or identification accuracy ($+0.75$ pp, $p = 0.21$): the disclosure has to occur at inference to matter. Second, the inference-time effect is significant only for a model trained without interlocutor biographies. Under $\text{Trn}_{\text{Strangers}}$, moving inference from *Strangers* to *Fan* raises leakage by 0.91 pp ($p = 0.007$); under Trn_{Fan} , $\text{Trn}_{\text{Peers}}$, and Trn_{I+I-} the same change is not significant (all $p > 0.3$). Any training configuration that supplies interlocutor biographies appears to remove the channel that inference-time disclosure would otherwise open.

	Δ pp	95% CI	p
<i>Trn_{Strangers} → Trn_{Fan}, at Inf_{Fan}</i>			
leakage	-0.09	$[-0.78, 0.59]$	0.845
accuracy	$+0.75$	$[-0.39, 1.92]$	0.214
<i>Leakage, Inf_{Strangers} → Inf_{Fan}</i>			
under $\text{Trn}_{\text{Strangers}}$	$+0.91$	$[0.27, 1.58]$	0.007
under Trn_{Fan}	$+0.37$	$[-0.32, 1.07]$	0.327
under $\text{Trn}_{\text{Peers}}$	-0.18	$[-0.78, 0.41]$	0.608
under Trn_{I+I-}	$+0.16$	$[-0.46, 0.78]$	0.663
<i>Leakage at Inf_{Fan}, vs. Trn_{Strangers}</i>			
Trn_{Fan}	-0.09	$[-0.78, 0.59]$	0.845
Trn_{I+I-}	-0.62	$[-1.28, 0.02]$	0.075
$\text{Trn}_{\text{Peers}}$	-0.91	$[-1.55, -0.30]$	0.006

Table 14: Paired exact McNemar tests with bootstrap 95% CIs ($n = 4,375$). Leakage is $\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Int}}$; accuracy is $\text{Both}_{\text{Disc}}$.

F Comprehensive Performance Tables

This section reports the full numerical grid, grouped by training objective. The overlap columns report METEOR biography–turn overlap scores, where higher values indicate greater lexical similarity. The **Acc** columns report judge accuracy. The binary rare-word percentages used in the main-text leakage analyses are reported in the main text and in Appendix G.

In addition to the three main inference settings, the appendix tables include Inf_{I+I-} , the reverse asymmetric condition in which the target sees the interlocutor biography but the interlocutor does not see the target biography. Speaker-familiarity rows use the definitions from Section 3.1. Topic-familiarity rows distinguish topics naturally associated with the target speaker (*familiar*, e.g. Harry from *When Harry Met Sally* discussing love) from topics outside the target’s typical context (*unfamiliar*, e.g. Harry discussing war). For each system family, the first table gives the full evaluation-condition grid, the second isolates speaker familiarity, and the third isolates topic familiarity. In the familiarity tables, accuracy is averaged over the four evaluation disclosure conditions, so the figures are comparable with Table 7 in the main text; the METEOR columns are unaffected by the judge’s view and are reported as measured.

F.1 Gold Dialogue Baseline

Table 16 establishes the reference point for generated dialogues. Gold dialogues combine high judge accuracy with uniformly low biography–turn overlap, showing that human-written persona signal is recognisable without heavy lexical reuse from the

Training	Inference	Accuracy				% Rare-word overlap		
		BothDisc	BioDisc	TurnsDisc	BothMask	BioTrg → TurnTrg	BioInt → TurnTrg	BioTrg → TurnInt
Trn _{Fan}	Inf _{Strangers}	0.833	0.848	0.790	0.759	4.34	1.78	2.63
	Inf _{Fan}	0.881	0.863	0.839	0.783	3.86	1.60	2.99
	Inf _{Peers}	0.850	0.833	0.822	0.748	4.09	2.06	2.86
Trn _{I+I-}	Inf _{Strangers}	0.808	0.811	0.730	0.690	3.13	1.19	2.31
	Inf _{Fan}	0.860	0.826	0.804	0.713	3.18	1.10	2.47
	Inf _{Peers}	0.832	0.805	0.765	0.690	2.22	1.55	2.24

Table 15: Asymmetric training regimes ($n = 4,375$ per cell). Trn_{Fan} trains only the interlocutor with the target biography; Trn_{I+I-} is its converse.

Eval	Bio _{Trg} → Turn _{Trg}	Bio _{Int} → Turn _{Trg}	Bio _{Trg} → Turn _{Int}	Acc
Both _{Disc}	0.033	0.033	0.032	0.914
Bio _{Disc}	0.033	0.033	0.032	0.885
Turns _{Disc}	0.033	0.033	0.032	0.667
Both _{Mask}	0.033	0.033	0.032	0.621

Table 16: Gold-dialogue overlap metrics reference.

Inf.	Eval	Bio _{Trg} → Turn _{Trg}	Bio _{Int} → Turn _{Trg}	Bio _{Trg} → Turn _{Int}	Acc
Inf _{Strangers}	Both _{Disc}	0.038	0.042	0.046	0.830
	Bio _{Disc}	0.038	0.042	0.046	0.848
	Turns _{Disc}	0.038	0.042	0.046	0.789
	Both _{Mask}	0.038	0.042	0.046	0.760
Inf _{Fan}	Both _{Disc}	0.038	0.042	0.040	0.873
	Bio _{Disc}	0.038	0.042	0.040	0.847
	Turns _{Disc}	0.038	0.042	0.040	0.843
	Both _{Mask}	0.038	0.042	0.040	0.779
Inf _{I+I-}	Both _{Disc}	0.039	0.036	0.046	0.824
	Bio _{Disc}	0.039	0.036	0.046	0.827
	Turns _{Disc}	0.039	0.036	0.046	0.766
	Both _{Mask}	0.039	0.036	0.046	0.737
Inf _{Peers}	Both _{Disc}	0.038	0.036	0.040	0.854
	Bio _{Disc}	0.038	0.036	0.040	0.834
	Turns _{Disc}	0.038	0.036	0.040	0.818
	Both _{Mask}	0.038	0.036	0.040	0.765

Table 17: MaskLoss trained without interlocutor-biography visibility (Trn_{Strangers}).

biographies. Accuracy is highest when the judge receives both interlocutor biography and turns, and it drops when interlocutor-side information is removed. This confirms that the judge is sensitive to interlocutor context even in human dialogue, while the low METEOR values indicate that this sensitivity is not simply driven by verbatim copying.

F.2 MaskLoss: Trn_{Strangers}

Tables 17–19 show the behaviour of a MaskLoss model trained without interlocutor-biography visibility. Overall accuracy is strongest when inference supplies additional interlocutor-side information, especially under *Fan* (Inf_{Fan}) and *Peers* (Inf_{Peers}), while the METEOR overlap columns remain close to the gold-dialogue range and vary little across

evaluation conditions. The speaker-familiarity table shows that most of the remaining difficulty comes from unfamiliar speaker pairings: familiar pairs stay in a narrow band (0.864–0.879) regardless of the inference configuration, whereas unfamiliar pairs range from 0.731 to 0.804 and so depend far more on what the interlocutor is given. Topic familiarity produces a smaller but consistent effect, with familiar topics generally easier than unfamiliar ones.

F.3 MaskLoss: Trn_{Peers}

Tables 20–22 report the corresponding MaskLoss model trained with mutual biography disclosure. Compared with Trn_{Strangers}, this model has lower METEOR overlap in all three directions, especially

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.865	0.877	0.864	0.879
Unfamiliar	0.763	0.804	0.731	0.772
<i>METEOR — Familiar</i>				
Bio _{Trg} →Turn _{Trg}	0.038	0.037	0.038	0.038
Bio _{Int} →Turn _{Trg}	0.044	0.044	0.037	0.037
Bio _{Trg} →Turn _{Int}	0.047	0.041	0.047	0.041
<i>METEOR — Unfamiliar</i>				
Bio _{Trg} →Turn _{Trg}	0.038	0.038	0.039	0.038
Bio _{Int} →Turn _{Trg}	0.042	0.042	0.036	0.036
Bio _{Trg} →Turn _{Int}	0.045	0.039	0.046	0.039

Table 18: Speaker familiarity for MaskLoss / Trn_{Strangers}.

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.829	0.857	0.804	0.844
Unfamiliar	0.790	0.820	0.778	0.798

Table 19: Topic familiarity for MaskLoss / Trn_{Strangers}.

Inf.	Eval	Bio_{Trg}→Turn_{Trg}	Bio_{Int}→Turn_{Trg}	Bio_{Trg}→Turn_{Int}	Acc
Inf _{Strangers}	Both _{Disc}	0.028	0.032	0.035	0.803
	Bio _{Disc}	0.028	0.032	0.035	0.814
	Turns _{Disc}	0.028	0.032	0.035	0.746
	Both _{Mask}	0.028	0.032	0.035	0.683
Inf _{Fan}	Both _{Disc}	0.028	0.032	0.028	0.849
	Bio _{Disc}	0.028	0.032	0.028	0.823
	Turns _{Disc}	0.028	0.032	0.028	0.795
	Both _{Mask}	0.028	0.032	0.028	0.707
Inf _{I+I-}	Both _{Disc}	0.028	0.026	0.034	0.787
	Bio _{Disc}	0.028	0.026	0.034	0.789
	Turns _{Disc}	0.028	0.026	0.034	0.703
	Both _{Mask}	0.028	0.026	0.034	0.673
Inf _{Peers}	Both _{Disc}	0.028	0.026	0.028	0.831
	Bio _{Disc}	0.028	0.026	0.028	0.793
	Turns _{Disc}	0.028	0.026	0.028	0.782
	Both _{Mask}	0.028	0.026	0.028	0.672

Table 20: MaskLoss trained with mutual biography disclosure (Trn_{Peers}).

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.834	0.845	0.830	0.847
Unfamiliar	0.707	0.755	0.669	0.711
<i>METEOR — Familiar</i>				
Bio _{Trg} →Turn _{Trg}	0.028	0.028	0.028	0.028
Bio _{Int} →Turn _{Trg}	0.034	0.034	0.028	0.027
Bio _{Trg} →Turn _{Int}	0.036	0.029	0.035	0.028
<i>METEOR — Unfamiliar</i>				
Bio _{Trg} →Turn _{Trg}	0.028	0.028	0.028	0.028
Bio _{Int} →Turn _{Trg}	0.031	0.031	0.026	0.025
Bio _{Trg} →Turn _{Int}	0.034	0.028	0.034	0.028

Table 21: Speaker familiarity for MaskLoss / Trn_{Peers}.

	$\text{Inf}_{\text{Strangers}}$	Inf_{Fan}	Inf_{I+I-}	$\text{Inf}_{\text{Peers}}$
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.783	0.819	0.758	0.794
Unfamiliar	0.745	0.775	0.723	0.752

Table 22: Topic familiarity for MaskLoss / Trn_{Peers}.

Inf.	Eval	$\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Trg}}$	$\text{Bio}_{\text{Int}} \rightarrow \text{Turn}_{\text{Trg}}$	$\text{Bio}_{\text{Trg}} \rightarrow \text{Turn}_{\text{Int}}$	Acc
$\text{Inf}_{\text{Strangers}}$	Both _{Disc}	0.037	0.042	0.045	0.832
	Bio _{Disc}	0.037	0.042	0.045	0.844
	Turns _{Disc}	0.037	0.042	0.045	0.786
	Both _{Mask}	0.037	0.042	0.045	0.753
Inf_{Fan}	Both _{Disc}	0.037	0.041	0.038	0.873
	Bio _{Disc}	0.037	0.041	0.038	0.846
	Turns _{Disc}	0.037	0.041	0.038	0.834
	Both _{Mask}	0.037	0.041	0.038	0.769
Inf_{I+I-}	Both _{Disc}	0.038	0.035	0.045	0.808
	Bio _{Disc}	0.038	0.035	0.045	0.824
	Turns _{Disc}	0.038	0.035	0.045	0.774
	Both _{Mask}	0.038	0.035	0.045	0.748
$\text{Inf}_{\text{Peers}}$	Both _{Disc}	0.037	0.034	0.038	0.855
	Bio _{Disc}	0.037	0.034	0.038	0.842
	Turns _{Disc}	0.037	0.034	0.038	0.825
	Both _{Mask}	0.037	0.034	0.038	0.755

Table 23: NoMaskLoss trained without interlocutor-biography visibility (Trn_{Strangers}).

for target-biography overlap in target turns, which is the desired copying reduction. The cost is a small drop in accuracy, most visible when the model is evaluated without the interlocutor context it saw during training. The familiarity breakdowns preserve the same broad pattern as Trn_{Strangers}: familiar speakers are easier than unfamiliar speakers, and topic familiarity is a weaker source of variation than speaker familiarity.

F.4 NoMaskLoss: Trn_{Strangers}

Tables 23–25 show that removing the final-turn mask has only a modest effect when the model is trained in the *Strangers* regime. Accuracy is very close to the MaskLoss Trn_{Strangers} counterpart, and the same inference configurations remain strongest. The overlap values are also similar, suggesting that in this training regime the dominant factor is not the loss objective alone but the absence of interlocutor-biography visibility during training. Speaker familiarity again explains more variance than topic familiarity, with unfamiliar pairings accounting for most of the accuracy loss.

F.5 NoMaskLoss: Trn_{Peers}

Tables 26–28 give the NoMaskLoss model trained with mutual disclosure. Relative to MaskLoss

Trn_{Peers}, NoMaskLoss recovers some accuracy in several evaluation conditions, including the fully masked evaluation, but it does so with slightly higher biography–turn overlap. This is the trade-off motivating the main-paper choice of MaskLoss: standard next-token prediction can make the target easier to identify, but part of that gain comes from making biography text itself more predictable. The speaker and topic breakdowns show the same structure as before, with the largest gains under *Fan* and *Peers* inference and the largest residual difficulty on unfamiliar speakers.

F.6 Zero-Shot Baseline

Tables 29–31 show why the zero-shot baseline should be interpreted differently from the fine-tuned models. Its accuracy is near ceiling across evaluation conditions, but its METEOR overlap values are much higher than those of the gold dialogues and the fine-tuned systems. This means that high recognisability is accompanied by strong surface similarity to the biographies. Speaker familiarity still matters, especially for unfamiliar pairings, but the topic-familiarity split is almost flat: once the zero-shot model copies enough biographical material, the topic itself contributes little additional diagnostic structure.

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.864	0.879	0.864	0.875
Unfamiliar	0.759	0.794	0.732	0.777
<i>METEOR — Familiar</i>				
Bio _{Trg} →Turn _{Trg}	0.038	0.037	0.037	0.036
Bio _{Int} →Turn _{Trg}	0.044	0.043	0.036	0.036
Bio _{Trg} →Turn _{Int}	0.046	0.038	0.046	0.038
<i>METEOR — Unfamiliar</i>				
Bio _{Trg} →Turn _{Trg}	0.037	0.036	0.038	0.037
Bio _{Int} →Turn _{Trg}	0.040	0.040	0.035	0.034
Bio _{Trg} →Turn _{Int}	0.045	0.038	0.045	0.038

Table 24: Speaker familiarity for NoMaskLoss / Trn_{Strangers}.

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.820	0.847	0.804	0.842
Unfamiliar	0.792	0.819	0.777	0.802

Table 25: Topic familiarity for NoMaskLoss / Trn_{Strangers}.

Inf.	Eval	Bio_{Trg}→Turn_{Trg}	Bio_{Int}→Turn_{Trg}	Bio_{Trg}→Turn_{Int}	Acc
Inf _{Strangers}	Both _{Disc}	0.032	0.036	0.038	0.824
	Bio _{Disc}	0.032	0.036	0.038	0.825
	Turns _{Disc}	0.032	0.036	0.038	0.757
	Both _{Mask}	0.032	0.036	0.038	0.707
Inf _{Fan}	Both _{Disc}	0.032	0.036	0.031	0.862
	Bio _{Disc}	0.032	0.036	0.031	0.830
	Turns _{Disc}	0.032	0.036	0.031	0.804
	Both _{Mask}	0.032	0.036	0.031	0.734
Inf _{I+I-}	Both _{Disc}	0.032	0.030	0.037	0.809
	Bio _{Disc}	0.032	0.030	0.037	0.806
	Turns _{Disc}	0.032	0.030	0.037	0.736
	Both _{Mask}	0.032	0.030	0.037	0.710
Inf _{Peers}	Both _{Disc}	0.031	0.029	0.030	0.840
	Bio _{Disc}	0.031	0.029	0.030	0.819
	Turns _{Disc}	0.031	0.029	0.030	0.782
	Both _{Mask}	0.031	0.029	0.030	0.712

Table 26: NoMaskLoss trained with mutual biography disclosure (Trn_{Peers}).

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.839	0.856	0.845	0.852
Unfamiliar	0.732	0.771	0.705	0.740
<i>METEOR — Familiar</i>				
Bio _{Trg} →Turn _{Trg}	0.032	0.032	0.031	0.031
Bio _{Int} →Turn _{Trg}	0.038	0.038	0.031	0.031
Bio _{Trg} →Turn _{Int}	0.039	0.031	0.038	0.031
<i>METEOR — Unfamiliar</i>				
Bio _{Trg} →Turn _{Trg}	0.032	0.031	0.032	0.031
Bio _{Int} →Turn _{Trg}	0.035	0.035	0.029	0.029
Bio _{Trg} →Turn _{Int}	0.037	0.030	0.037	0.030

Table 27: Speaker familiarity for NoMaskLoss / Trn_{Peers}.

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.806	0.839	0.789	0.815
Unfamiliar	0.758	0.784	0.748	0.768

Table 28: Topic familiarity for NoMaskLoss / Trn_{Peers}.

Inf.	Eval	Bio_{Trg}→Turn_{Trg}	Bio_{Int}→Turn_{Trg}	Bio_{Trg}→Turn_{Int}	Acc
Inf _{Strangers}	Both _{Disc}	0.092	0.090	0.102	0.936
	Bio _{Disc}	0.092	0.090	0.102	0.963
	Turns _{Disc}	0.092	0.090	0.102	0.939
	Both _{Mask}	0.092	0.090	0.102	0.971
Inf _{Fan}	Both _{Disc}	0.093	0.091	0.095	0.952
	Bio _{Disc}	0.093	0.091	0.095	0.965
	Turns _{Disc}	0.093	0.091	0.095	0.954
	Both _{Mask}	0.093	0.091	0.095	0.971
Inf _{I+I-}	Both _{Disc}	0.089	0.083	0.100	0.930
	Bio _{Disc}	0.089	0.083	0.100	0.954
	Turns _{Disc}	0.089	0.083	0.100	0.927
	Both _{Mask}	0.089	0.083	0.100	0.955
Inf _{Peers}	Both _{Disc}	0.089	0.083	0.093	0.953
	Bio _{Disc}	0.089	0.083	0.093	0.957
	Turns _{Disc}	0.089	0.083	0.093	0.947
	Both _{Mask}	0.089	0.083	0.093	0.962

Table 29: Zero-shot baseline (M_{z-s}).

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.993	0.996	0.996	0.996
Unfamiliar	0.894	0.919	0.880	0.921
<i>METEOR — Familiar</i>				
Bio _{Trg} →Turn _{Trg}	0.092	0.092	0.085	0.084
Bio _{Int} →Turn _{Trg}	0.093	0.094	0.083	0.083
Bio _{Trg} →Turn _{Int}	0.100	0.091	0.096	0.088
<i>METEOR — Unfamiliar</i>				
Bio _{Trg} →Turn _{Trg}	0.093	0.093	0.091	0.092
Bio _{Int} →Turn _{Trg}	0.089	0.089	0.083	0.083
Bio _{Trg} →Turn _{Int}	0.103	0.097	0.101	0.096

Table 30: Speaker familiarity for the zero-shot baseline.

	Inf_{Strangers}	Inf_{Fan}	Inf_{I+I-}	Inf_{Peers}
<i>Accuracy (mean over eval. conditions)</i>				
Familiar	0.955	0.964	0.938	0.953
Unfamiliar	0.950	0.958	0.944	0.956

Table 31: Topic familiarity for the zero-shot baseline.

Eval	No ov.	Ov.	Δ pp
<i>M_{z-s} / Inf_{Strangers}</i>			
Both _{Disc}	0.919 (3123)	0.980 (1252)	+6.1
Turns _{Disc}	0.921 (3123)	0.984 (1252)	+6.3
<i>M_{z-s} / Inf_{Fan}</i>			
Both _{Disc}	0.935 (2992)	0.988 (1383)	+5.2
Turns _{Disc}	0.937 (2992)	0.991 (1383)	+5.4
<i>Trn_{Strangers} / Inf_{Strangers}</i>			
Both _{Disc}	0.827 (4280)	0.968 (95)	+14.1
Turns _{Disc}	0.785 (4280)	0.979 (95)	+19.4
<i>Trn_{Strangers} / Inf_{Fan}</i>			
Both _{Disc}	0.869 (4240)	1.000 (135)	+13.1
Turns _{Disc}	0.839 (4240)	0.985 (135)	+14.6

Table 32: Judge accuracy split by whether interlocutor turns contain any exact rare-word overlap with the target biography. Parentheses give subset sizes; Δ is overlap minus no overlap accuracy in percentage points.

G Rare-Word Leakage Subsets

Section 4.4 reports the within configuration accuracy gain when interlocutor turns contain at least one rare word from the target biography (Figure 5). Table 32 reports the absolute subset accuracies underlying that gain, together with subset sizes.

Overlap-positive subsets are near ceiling in every configuration (0.968–1.000), confirming that exact rare-word overlap is a strong identification cue. Under *Trn_{Strangers}*, however, the no-overlap subset is also more accurate under *Fan* than under *Strangers* (Turns_{Disc}: 0.839 vs. 0.785; Both_{Disc}: 0.869 vs. 0.827), suggesting that rare-word overlap is a lower bound on broader leakage through paraphrase, topic steering, or implicit semantic uptake.

H Training–Inference Dialogue Examples

All examples use Walter Sobchak from *The Big Lebowski* (Speaker 1, target) and Peter Parker / Spider-Man from *Spider-Man* (Speaker 2, interlocutor), enabling direct comparison across training–inference conditions. These examples are illustrative; the claims in the main paper rest on aggregate results. Both biographies, reproduced verbatim from PRODIGy (Occhipinti et al., 2024), are reported in Table 33; the four configuration tables that follow refer back to them.

Target biography (Walter): I am owner of Sobchak Security; I use the combat knowledge I accumulated during the Vietnam War. 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. I have to help my friend The Dude through a sticky situation. While my tour of duty prepared me for a career in security, it also left me paranoid and angry.

Interlocutor biography (Peter / Spider-Man): 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 reserved and quiet. Whenever I am Spider-Man, I become a consummate jokester using one-liners to irritate my foes. I am super strong, super agile, capable of sensing danger, and able to shoot webs out of my hands. I have to keep New York City safe.

Table 33: PRODIGy biographies for the target (Walter) and interlocutor (Peter / Spider-Man) used throughout this appendix, shortened to the five most representative sentences each. Highlighted spans anticipate the dialogues below: lexical copying (Table 34) and thematic adaptation (Table 37).

Tables 34–37 illustrate all four training×inference configurations. Concordant *Trn_{Strangers}/Inf_{Strangers}* shows Spider-Man-related terms surfacing despite masking. Concordant *Trn_{Peers}/Inf_{Peers}* shows contextual adaptation to the interlocutor. Discordant *Trn_{Strangers}/Inf_{Peers}* does not visibly use the newly supplied interlocutor information, while discordant *Trn_{Peers}/Inf_{Strangers}* produces generic output when the expected interlocutor context is absent.

Trn_{Strangers}/Inf_{Strangers} [Concordant: both WITHOUT]

Target Bio (Walter): as in Table 33.

Interlocutor Bio (Peter): [MASKED] (contains Spider-Man, webs) **Topic:** Fiction

Speaker 1: What? You mean like Jim?

Speaker 2: Spiderman!

Speaker 1: Yeah... Like Spiderman!

Speaker 2: How do you know it’s me?!

Speaker 1: Because you’ve got these... little web things on your clothes! Yeah – there’s one right here...

Table 34: Concordant no-disclosure setting (*Trn_{Strangers}/Inf_{Strangers}*). Spider-Man-related terms still surface in both speakers’ turns (highlighted), illustrating residual lexical leakage.

Trn_{Strangers}/Inf_{Peers} [Discordant: train WITHOUT, infer WITH]

Target Bio (Walter): as in Table 33.

Interlocutor Bio (Peter): as in Table 33. **Topic:** Fiction

Speaker 1: *They didn't tell me it was a limited engagement! Look here – he's getting ready for another one now. . .*

Speaker 2: *Shut up about that stuff right now. Let's talk turkey here. What were you doing there?*

Speaker 1: *Did you see anything weird tonight?*

Speaker 2: *Hey, what're we talking about? No way I'm going back over all this again. . .*

Table 35: Discordant disclosure setting (Trn_{Strangers}/Inf_{Peers}). The model receives the interlocutor biography at inference, but the supplied Spider-Man context is not visibly integrated.

Trn_{Peers}/Inf_{Strangers} [Discordant: train WITH, infer WITHOUT]

Target Bio (Walter): as in Table 33.

Interlocutor Bio (Peter): [MASKED] **Topic:** Fiction

Speaker 1: *Listen very carefully – I'm going through this once more for you. This man has become like a son to me. . .*

Speaker 2: *That name sounds familiar somehow.*

Speaker 1: *It's probably because he was an engineer on one of our planes shot down over Laos. . .*

Speaker 2: *What's wrong?*

Table 36: Discordant masking setting (Trn_{Peers}/Inf_{Strangers}). A model trained to expect interlocutor context produces generic dialogue when that context is absent at inference.

Trn_{Peers}/Inf_{Peers} [Concordant: both WITH]

Target Bio (Walter): as in Table 33.

Interlocutor Bio (Peter): as in Table 33. **Topic:** Fiction

Speaker 1: *Oh boy! Did you see? He's gonna get it again. . . What the hell was wrong with him?*

Speaker 2: *What does this thing do?*

Speaker 1: *It takes a picture when they hit me or something. Come check it out. . .*

Speaker 2: *Can it keep up? My camera could make money doing this all by itself!*

Table 37: Concordant mutual-disclosure setting (Trn_{Peers}/Inf_{Peers}). Aligned training and inference produce thematic adaptation to Spider-Man's photographer context (underlined).