# Evaluating Linguistic Speaker Profiles on Response Selection in Multi-Party Dialogue

Maryam Sajedinia[1], Seyed Mahed Mousavi[2] and Valerio Basile[3]

[1]Modeling & Simulation of Techno-Social Systems, Fondazione Bruno Kessler, Italy
[2]Signals & Interactive Systems Lab, University of Trento, Italy
[3]University of Turin, Italy

**Abstract**

We investigate whether incorporating linguistically derived speaker profiles improves the response selection capabilities of instruction-tuned large language models (LLMs) in multi-party dialogues. Using the Wikipedia Talk Page dataset, we construct lightweight profiles for each speaker based on features extracted from their prior messages, including frequent nouns and verbs, and sentiment tendency. These profiles are incorporated into the input prompts and evaluated using in-context learning with LLaMA 3.2 Instruct (1B and 8B) and GPT-4o, without any model fine-tuning. We compare performance across models and prompt settings, with and without speaker profiles, and analyze the effect of different profile configurations. Results are compared against a Random baseline and a supervised Siamese RNNs (with GRU units) trained on the same data. Our results show that incorporating speaker profiles improves response selection performance across most LLM settings, with the strongest gains observed in larger models such as LLaMA 3.2 (8B). Lexical features (frequent nouns and verbs) demonstrate greater improvements than sentiment information, particularly in low-context or underspecified scenarios. However, profile effectiveness varies by model scale and prompt format, and provides limited benefit in cases where distractors are lexically and semantically similar to the ground-truth response.

**Keywords**

Large Language Model, Multiparty Dialogue, User Profile, Response Selection,

## 1. Introduction

Large Language Models (LLMs) have achieved state-of-the-art performance on a variety of downstream tasks in dialogue systems, including response generation [2, 3, 4], selection [5, 6], and dialogue state tracking [7, 8]. Despite these advances, Multi-Party Dialogue (MPD) remains a more complex setting due to the increased number of participants, diverse conversational roles, and overlapping discourse structures [9, 10]. One key challenge in this context is the absence of explicit user modeling. LLMs typically operate over short dialogue contexts without access to persistent or structured information about individual speakers. This limits their ability to personalize responses or disambiguate interactions based on user-specific traits such as language use, affective tone, or conversational behavior.

Response selection in MPD poses unique challenges due to the presence of multiple speakers, shifting roles, and overlapping conversational threads [5]. Unlike dyadic dialogue, this setting requires distinguishing between several potential interlocutors, resolving ambiguous references, and interpreting speaker-specific cues. These complexities make the task particularly sensitive to both conversational context and speaker identity. However, standard LLM-based approaches primarily rely on surface-level context, without modeling the linguistic behavior of individual speakers, which can limit their ability to correctly select the responses.

**RQ1: Can linguistic speaker profiles improve response selection in multi-party dialogue settings when provided via prompt-only conditioning?** In this work, we investigate whether incorporating cost-effective, linguistically grounded speaker profiles into the input prompt can improve the response selection capabilities of instruction-tuned LLMs in MPD. Our motivation is to test whether such profiles can serve as effective signals for disambiguation and speaker-sensitive response selection. The profiles are derived from users' previous utterances and include their most frequent nouns and verbs, along with a coarse-grained sentiment distribution. Unlike approaches that rely on fine-tuning or persistent user modeling, we adopt a prompt-only strategy using minimal, interpretable features compatible with in-context learning. We focus specifically on response selection rather than generation, as it allows for more controlled experimental conditions and avoids the need for human evaluation. Moreover, automatic metrics for ranking responses are more reliable than those available for open-ended generation, which often struggle to distinguish coherent yet irrelevant outputs [11].

**RQ2: To what extent does the effectiveness of speaker profiles depend on model scale, prompt**

*CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy [1]*
*Corresponding author.
msajedinia@fbk.eu (M. Sajedinia); mahed.mousavi@unitn.it (S. M. Mousavi); valerio.basile@unito.it (V. Basile)

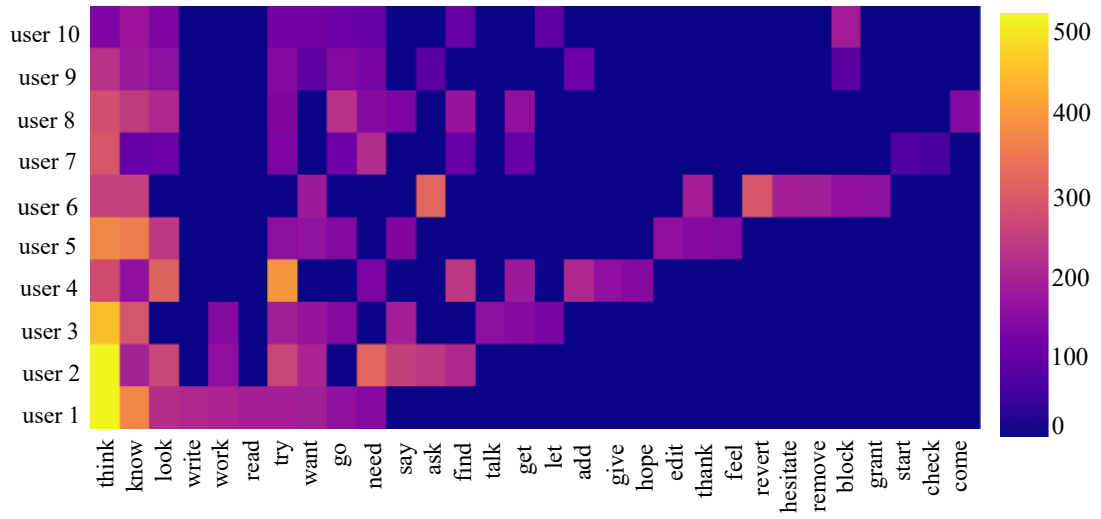


**Figure 1:** Heatmap of the 10 most frequent verbs used by each of the top 10 most active users. The diagonally dominant structure reveals a strong speaker-specific lexicon: verbs frequently used by one user tend to be infrequent or absent in the language of others. This pattern supports the hypothesis that verb usage in MPD can serve as a useful signal for user-aware response selection.

**format, and profile composition?** We evaluate this approach using LLaMA 3.2 Instruct (1B and 8B) and GPT-4o, comparing performance with and without speaker profiles under zero-shot and one-shot prompting. To contextualize results, we include two baselines: a random ranking strategy and a supervised Siamese RNN with GRU units trained on the same dataset. All models are tested on a standardized response selection task using MPDs from the Wikipedia Talk Page dataset.

Our goal is not to build a personalized dialogue system, but to assess whether minimal linguistic speaker information can influence LLM behavior in a selection setting. We do not assume access to long-term user history or stable user identities, and we make no changes to model parameters. Instead, we treat speaker profiling as a lightweight, model-agnostic addition to the input prompt. This setup allows us to isolate the effect of speaker-level information on model performance and to compare its impact across multiple instruction-tuned LLMs.

Our results show that speaker profiles can enhance response selection performance, particularly for larger models and in low-context scenarios. The most consistent gains are observed with lexical profiles (frequent nouns and verbs), while sentiment information yields marginal or mixed improvements. However, model scale and prompt format (e.g. 0-shot and 1-shot) significantly mediate the effectiveness of speaker profiles.

Our contributions can be summarized as follows:

- We introduce a prompt-based method for incorporating lightweight, linguistically derived speaker profiles into LLM-based response selection for multi-party dialogue[1].
- We conduct a systematic evaluation across model scales (1B, 8B, GPT-4o), prompt formats (zero-shot, one-shot), and profile configurations (lexical, lexical+sentiment).
- We present detailed analysis highlighting when and how speaker profiles help, supported by both aggregate performance and error case breakdowns.

## 2. Related Work

Recent work on MPD has explored a range of strategies for modeling speaker identity, roles, and interaction structure. Mahajan and Shaikh [12] introduce a graph-based transformer that incorporates speaker and addressee personas as structured metadata, using crowdsourced profiles to condition response generation. Similarly, Ju et al. [4] build a graph representation of utterances and speaker personas to guide generation through a hierarchical encoder and structured aggregation. These methods emphasize user profile incorporation but assume access to annotated profiles and require complex modeling. Sun et al. [13] use contrastive learning to model speaker-

[1]The code and implementation details will be published in our repository

specific discourse patterns without explicit profiles, learning latent speaker distinctions optimized for generation tasks. Penzo et al. [5] take a diagnostic approach, analyzing how conversation structure affects performance in response selection and addressee recognition. They show that LLMs rely heavily on surface content for response selection and are sensitive to prompt formulation and structural variation. Finally, Hu et al. [9] propose a role-aware modeling framework that combines role-context pretraining with decoding constraints to favor role-consistent outputs. While effective across multiple MPD tasks, the approach depends on predefined role labels and supervised training. Collectively, these studies highlight the importance of speaker- and role-level information in MPD, though most rely on supervised learning, structured annotations, or architectural specialization.

## 3. Experiments

We evaluate the effect of incorporating linguistic speaker profiles on response selection in MPDs using a set of instruction-tuned LLMs and baseline models. All experiments are conducted on the same unseen test set, using consistent prompt formatting and evaluation metrics. Below, we describe the models, data, and profile features used in our setup.

### 3.1. Dataset

We use the Wikipedia Talk Page Conversations dataset [14], which contains 124,957 multi-party dialogues involving 38,462 unique users and a total of 4,023,376 tokens with a vocabulary size of 108,416. The user activity in the dataset is not balanced, i.e. the top 10 most active users account for over 12% of all turns in the dataset.

To model multi-party interactions, each conversation is represented as a tree, with the root corresponding to the initial post and branches representing reply chains. For each reply path, we extract a linear dialogue history leading to a candidate response. Each instance is framed as a response selection task with one ground-truth response and nine distractors drawn from the same structural depth within other conversations.

We segment this subset into three partitions: a held-out test set of 2,500 previously unseen dialogues shared across all models; a training set of 206,633 samples used to train the Siamese RNN and construct one-shot prompts; and a development set of 25,830 samples for tuning the supervised model. To better understand the conversational domain of the dialogues, we applied topic classification using GPT-4o, following the categorization and methodology of Antypas et al. [15] (50 samples were randomly selected and manually controlled to ensure prediction validity). Table 1 presents the distribution of detected topics

| Topic | Count |
|---|---|
| Business & Entrepreneurs | 20,293 |
| Celebrity & Pop Culture | 19,111 |
| Diaries & Daily Life | 17,150 |
| Arts & Culture | 17,034 |
| Learning & Educational | 16,283 |
| Science & Technology | 11,708 |
| News & Social Concern | 10,970 |
| Relationships | 9,654 |
| Technology | 5,019 |

**Table 1**
Topic distribution in Wikipedia Talk Page dialogues, detected using GPT-4o following Antypas et al. [15].

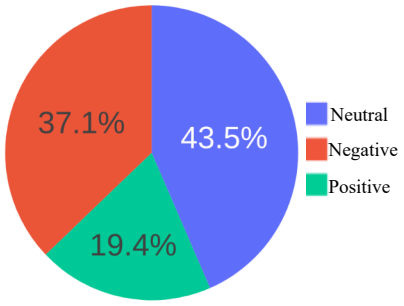


**Figure 2:** Distribution of predicted sentiment labels across all messages. Sentiment labels were derived using GPT-4o and manually verified on a subset of the data.

across the dialogues involving these ten users, covering a broad range of domains including business, popular culture, education, and technology.

We segment the data into three parts: a held-out test set of 2,500 previously unseen dialogues used for evaluation equal for all models; a training set of 206,633 samples used exclusively for training the Siamese-RNN baseline and for constructing one-shot examples; and a development set of 25,830 samples used only for optimizing the RNN architecture and hyperparameters. Each response selection instance consists of a dialogue history and a pool of ten candidate responses drawn from the same depth level in the reply tree. One candidate is the correct continuation, and the remaining nine are randomly sampled distractors from other conversations at the same structural depth.

### 3.2. Models

We evaluate three types of models:

- ***Random Baseline*** generates a uniformly ranked list of candidate responses for each input context. This serves as a lower-bound reference point and helps contextualize performance in the absence

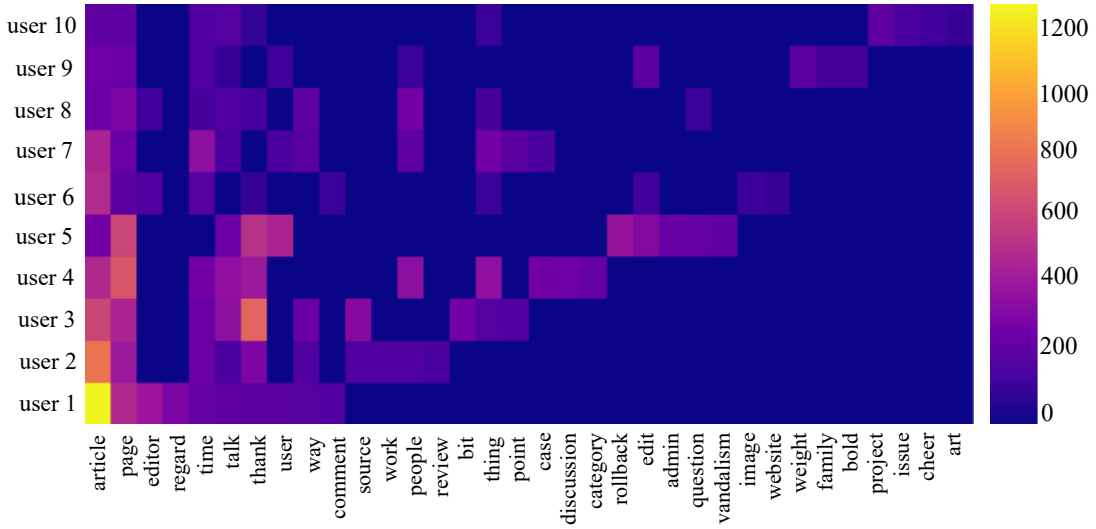


**Figure 3:** Heatmap of the 10 most frequent nouns used by each of the top 10 most active users. Similiar to Figure 1, noun usage is also highly speaker-specific: commonly used nouns for one user are rarely shared with others. This reinforces the utility of lexical profiles, suggesting that noun usage can provide a strong signal for speaker and response selection within MPD.

of data-driven inference.

- ***Siamese RNN*** is a supervised neural baseline using two GRU encoders with shared weights to compute the similarity between a dialogue context and a candidate response. The model outputs a matching score based on pairwise similarity and is trained using labeled context-response pairs with cross-entropy loss. Each GRU encoder uses the following hyperparameters: `MAX_LENGTH = 300`, `input_size = 100`, `hidden_size = 300`, `num_layers = 2`, `dropout = 0`, and `bidirectional = True`. The model is trained for `10 epochs` with a `batch_size = 128` and a learning rate of `0.0001`.
- ***Instruction-Tuned LLMs*** include **LLaMA 3.2 Instruct** (1B and 8B) and **GPT-4o**, representing two families of recent state-of-the-art LLMs. LLaMA 3.2 Instruct is a publicly available model family released by Meta, trained on a diverse multilingual corpus and further instruction-tuned to follow natural language prompts. We include both the 1B and 8B variants to examine the effect of model scale on profile sensitivity. GPT-4o is a proprietary model released by OpenAI, optimized for multimodal interaction and known for strong instruction-following capabilities in both zero-shot and few-shot settings. All models are used via API in inference-only mode without any additional fine-tuning. Inputs are provided as structured natural language prompts, including a system instruction, dialogue history, and a list of candidate responses. When speaker profiles are used, they are appended to the input as plain-text feature descriptions associated with the target speaker. We experiment with both zero-shot prompting (task description only) and one-shot prompting (including a single example of the desired input-output format). Inference is run with $temperature = 0.2$, $top_p = 1.0$, and $max_tokens = 50$, and predictions are parsed to compute Recall@1/2/5.

**Evaluation Metric** We evaluate model performance using Recall@k, a standard metric for response selection tasks. For each dialogue instance, the model ranks a set of ten candidate responses, consisting of the ground-truth response and nine distractors sampled from the same depth level in the conversation tree. Recall@k measures the proportion of instances where the correct response appears in the top $k$ predictions. We report Recall@1, Recall@2, and Recall@5 to assess performance at different levels of ranking sensitivity.

**Prompt Design** We structure prompts for the response selection task using a consistent template for ranking responses based on the context. Each prompt comprises three components: (i) a task instruction explaining the ranking objective and expected output format, (ii) a content section containing the dialogue history and 10 candidate responses, and (iii) an optional speaker profile,

**System Prompt (abbreviated)**

<|begin_of_text|>
You will be given:
- A conversation transcript with numbered turns
- 10 candidate responses
- A user profile containing the most frequent nouns and verbs used by the *next speaker*
Your task:
**Rank the candidate response indexes from best to worst** based on how well they continue the conversation and match the speaker profile.
**Example output format:**
1. 3
2. 4
...
Do NOT provide an explanation but the list of numbers.
<|eot_id|>

**User Prompt (example structure)**

**<CONVERSATION>**
Turn 1: Hi, how are you?
Turn 2: I'm doing well, thanks. You?
...
**</CONVERSATION>**

**<Responses>**
1. I'm glad to hear that!
2. What's new with you?
...
**</Responses>**

**<User Profile>**
thank, update, read, discuss, feel, ...
**</User Profile>**
<|eot_id|>

**Table 2**
Prompt structure used for response selection in LLMs. The system prompt defines the task, and the user prompt provides the conversation context, candidate responses, and speaker profile when applicable.

appended when profiling is enabled. The speaker profile provides the most frequent nouns and verbs used by the next speaker, i.e. the user who is expected to respond, extracted from their prior messages. The full prompt is framed in natural language and formatted using system and user tags. The model is explicitly instructed to return a ranked list of response indices without any explanation or commentary. In one-shot settings, we prepend a demonstration example showing the exact input-output structure. The speaker profile, when present, is enclosed in a $< UserProfile >$ section and labeled accordingly. This design follows the practices for LLM prompting in prior work [16]. We provide the prompt template in Table 2.

### 3.3. User Profile

We construct speaker profiles for each user in the dataset, using linguistic features extracted from their prior messages. Each profile is fixed per speaker and remains constant across all dialogue instances in which the user appears. We create a lexical profile consisting of the 10 most frequent nouns and the 10 most frequent verbs used by the speaker, extracted using the *spaCy* dependency parser. These tokens reflect habitual vocabulary choices and serve as coarse indicators of speaker identity and discourse tendencies. This profile is then augmented with a coarse-grained sentiment distribution. Each message authored by the speaker is classified as *positive*, *neutral*, or *negative* using GPT-4o, following a prior work [16] and the resulting counts are normalized to produce a speaker-level sentiment distribution (predictions were manually verified for 50 randomly sampled messages to ensure classifier quality). Profiles are incorporated into the prompt and are explicitly associated with the speaker expected to produce the next turn. This design allows instruction-tuned LLMs to condition their ranking decisions on user-specific linguistic traits without requiring model fine-tuning or structural modifications.

Figure 2 presents the overall sentiment distribution across the turns in the dataset. The majority are neutral (43%), followed by negative (37%) and positive (20%), indicating a generally balanced emotional tone. Figures 1 and 3 show heatmaps of the top 10 most frequent verbs and nouns, respectively, for 10 most frequent users. Each heatmap reveals strong user-specific vocabulary patterns: the most frequent items for a given user tend to be rarely used by others. This lexical asymmetry suggests that even simple word-level statistics can encode informative signals about speaker identity. As a result, lexical profiles may help disambiguate responses in MPD by aligning candidate utterances with user-specific vocabulary preferences.

## 4. Evaluation

We evaluate the effect of incorporating linguistic speaker profiles on the response selection performance of instruction-tuned LLMs in MPDs. Our analysis compares three models, GPT-4o, LLaMA 3.2 Instruct (1B and 8B), and a Siamese RNN baseline, under both zero-shot and one-shot prompting conditions. We assess each model's performance with and without speaker profile information, using two profile configurations as frequent nouns and verbs, and the addition of sentiment tendency.

**Baseline Behavior** The Siamese RNN performs moderately well in the profile-free condition, achieving 31% Recall@1. However, its performance declines when profiles are added. This suggests that the architecture may not effectively integrate linguistic profile information, or

| Model | ICL Setting | Profile | Recall@$k$ in 10 | | |
|---|---|---|---|---|---|
| | | | ***R@1*** | ***R@2*** | ***R@5*** |
| **Random** | - | - | 0.10 | 0.20 | 0.50 |
| **Siamese-RNN** | - | *w.o. User Profile* | 0.31 | 0.35 | 0.35 |
| | - | Freq. Nouns & Verbs | 0.15 | 0.28 | 0.56 |
| | - | + Sentiment | 0.14 | 0.26 | 0.55 |
| **Llama 3.2$_{Ins.}$ 1B** | 0-shot | *w.o. User Profile* | 0.10 | 0.21 | 0.51 |
| | | Freq. Nouns & Verbs | 0.11 | 0.21 | 0.52 |
| | | + Sentiment | 0.11 | 0.20 | 0.51 |
| | 1-shot | *w.o. User Profile* | 0.10 | 0.21 | 0.52 |
| | | Freq. Nouns & Verbs | 0.10 | 0.21 | 0.50 |
| | | + Sentiment | 0.11 | 0.22 | 0.52 |
| **Llama 3$_{Ins.}$ 8B** | 0-shot | *w.o. User Profile* | 0.20 | 0.33 | 0.61 |
| | | Freq. Nouns & Verbs | 0.40 | 0.49 | 0.72 |
| | | + Sentiment | 0.36 | 0.46 | 0.68 |
| | 1-shot | *w.o. User Profile* | 0.22 | 0.24 | 0.55 |
| | | Freq. Nouns & Verbs | 0.25 | 0.34 | 0.60 |
| | | + Sentiment | 0.22 | 0.32 | 0.59 |
| **GPT-4o** | 0-shot | *w.o. User Profile* | 0.56 | 0.66 | 0.82 |
| | | Freq. Nouns & Verbs | 0.59 | 0.69 | 0.83 |
| | | + Sentiment | 0.62 | 0.70 | 0.83 |
| | 1-shot | *w.o. User Profile* | 0.60 | 0.69 | 0.83 |
| | | Freq. Nouns & Verbs | 0.62 | 0.70 | 0.84 |
| | | + Sentiment | 0.62 | 0.69 | 0.84 |

**Table 3**
Response selection performance (Recall@1,2,5) across models, prompting settings (zero-shot, one-shot), and speaker profile configurations. Each instance includes 10 candidates (1 ground-truth, 9 distractors). Across LLMs, incorporating speaker profiles improves performance in nearly all settings, with the strongest gains observed for LLaMA 3.2 Instruct (8B) in the zero-shot condition. Lexical profiles (frequent nouns and verbs) consistently outperform sentiment-augmented profiles, particularly in lower-capacity models. These results suggest that shallow linguistic profiles can enhance LLM-based response selection, but their effectiveness varies with model size and prompting regime.

that the additional features introduce noise in the learned similarity space. The random baseline performs as expected, confirming that all models operate well above chance.

**LLM Performance** Table 3 presents the performance scores across models, prompt settings, and speaker profile configurations. GPT-4o achieves the highest performance in all conditions, with Recall@1 reaching 62% under one-shot prompting with profile information. LLaMA 3.2 Instruct (8B) performs substantially better than its 1B variant, particularly in the zero-shot setting, where the addition of speaker profiles yields the largest relative improvements.

**Speaker Profiles** Incorporating speaker profiles leads to consistent gains across most LLM configurations. For LLaMA 3.2 Instruct (8B), the inclusion of frequent nouns and verbs improves Recall@1 from 20% to 40% in the zero-shot setting. However, sentiment augmentation does not produce additional gains and, in some cases, slightly degrades performance. Nevertheless, the smaller LLaMA model (1B) shows minimal sensitivity to profile input, suggesting that profile utility may depend on model size. Meanwhile, GPT-4o demonstrates strong baseline performance without profiles, but still benefits from profile inclusion. The highest Recall@1 for GPT-4o is 62% with both lexical and sentiment features in the one-shot setting. These improvements, though smaller in magnitude compared to LLaMA 8B, indicate that even high-performing models can leverage cost-effective linguistic speaker information.

**Prompt Structure** Prompting style has non-uniform impact on models' performance. For LLaMA 3.2 Instruct (8B), zero-shot prompting outperforms one-shot in several configurations, particularly when profiles are included. In contrast, GPT-4o benefits more consistently from one-shot prompting, though the margin is small. These results highlight interactions between model scale, prompt format, and profile effectiveness.

### 4.1. Error Analysis

To better understand the limitations and strengths of speaker profiles, we manually analyzed several subsets of the test set. In our analysis, we define a *misclassified* instance as one in which the ground-truth (GT) response does not appear among the top five ranked candidates (i.e., not within Recall@5), and a *correct* instance as one where the GT response is ranked first (i.e., Recall@1).

Out of 2,500 total instances, 1,500 cases were consistently misclassified by all models across all conditions. In these cases, the distractors were often semantically and lexically similar to the GT responses, making the ranking task inherently difficult. Moreover, frequent nouns and verbs extracted for profile construction were typically generic (e.g., "thanks," "help," "response"), and occurred in both GTs and distractors, limiting their discriminative value. In such cases, the profile provided little to no additional context to support accurate disambiguation.

In contrast, 611 instances were correctly classified by all models across all settings. Here, the GT responses were clearly more contextually grounded and lexically aligned with the dialogue history, and the distractors were often generic acknowledgements (e.g., "thanks," "okay") or off-topic continuations. The linguistic profiles were more distinctive in these examples and appeared to support the model's ability to prioritize the correct response.

Finally, in 77 cases, all models failed without speaker profiles but they all correctly selected the GT response once profile information was added. These instances were typically characterized by minimal dialogue history (one-turn inputs), where contextual grounding was insufficient for accurate prediction. The added speaker profile appeared to serve as an auxiliary context that supported correct ranking in these otherwise under-specified dialogues. Conversely, there were 2 cases in which the inclusion of sentiment in the profile led to improved predictions in all models. These examples featured strong affective alignment between the dialogue history and the GT response, while the distractors were neutral and short, allowing the model to benefit from the added sentiment context.

Interestingly, in 12 cases the models ranked the correct response at R@1 without speaker profiles, but failed to do so when profiles were added. In these cases, sentiment distribution was nearly uniform across responses in these cases, providing no additional signal. Furthermore, the distractors were uniformly generic, with some distractors including non-English text or irrelevant long-form content. Thus, the profile content introduces more noise rather than useful contrast, confusing the model.

Overall, speaker profiles provide most benefit when dialogue context is minimal or generic, but lose effectiveness when distractors are lexically similar or the profiles themselves are noisy.

## 5. Conclusion

We investigate whether linguistically derived speaker profiles can improve the response selection capabilities of instruction-tuned LLMs in multi-party dialogue. We constructed user profiles based on frequent nouns, verbs, and sentiment tendencies from prior utterances, and incorporated them into prompts without any model fine-tuning. Our experiments with LLaMA 3.2 and GPT-4o show that lexical profiles consistently improve performance, particularly for larger models and in zero-shot settings. Our results show that lexical speaker profiles improve performance in nearly all LLM settings, especially in larger models and zero-shot conditions. This supports RQ1, demonstrating that even lightweight user information can help response selection in MPD. In addressing RQ2, we find that model scale and prompt design play a crucial role in how effectively speaker profiles are used. Larger models benefit more from profile information, suggesting that they can better leverage user context. However, the sentimental features show mixed results, in some cases adding noise rather than clarity. We also observe that profiles are particularly useful in low-context situations, but their impact diminishes when distractors are semantically close or when the profiles themselves lack specificity.

In future work, we plan to explore richer profile representations, investigate cross-domain generalizability, and test the applicability of this approach in real-time or streaming dialogue systems. We also see potential in extending our method to multilingual MPD and combining profile signals with structural or discourse-level features.

## Limitations

This study relies exclusively on in-context learning and does not involve any fine-tuning of the evaluated models. While this makes our approach lightweight and accessible, it also constrains the models' ability to adapt more deeply to user-specific behaviors. Due to computational constraints, we did not experiment with larger LLMs beyond LLaMA 3.2 (8B) and GPT-4o, and were unable to explore open-weight models at scale requiring GPU access. Our data is limited to English Wikipedia Talk Pages, which restricts the generalizability of our findings to multilingual or informal conversational domains. Additionally, speaker profiles are based on automatic extraction of lexical and sentiment features, which may introduce noise or inaccuracies that affect profile quality. Finally, we focus exclusively on response selection and did not experiment with response generation. While this

choice enables robust and reproducible automatic evaluation, it leaves open the question of how linguistic speaker profiles might affect the quality of generated responses in more open-ended dialogue settings.

## References


[1] C. Bosco, E. Ježek, M. Polignano, M. Sanguinetti, Preface to the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[2] S. Alghisi, M. Rizzoli, G. Roccabruna, S. M. Mousavi, G. Riccardi, Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue, in: S. Mahamood, N. L. Minh, D. Ippolito (Eds.), Proceedings of the 17th International Natural Language Generation Conference, Association for Computational Linguistics, Tokyo, Japan, 2024, pp. 180–197. URL: https://aclanthology.org/2024.inlg-main.15/. doi:10.18653/v1/2024.inlg-main.15.

[3] S. M. Mousavi, S. Caldarella, G. Riccardi, Response generation in longitudinal dialogues: Which knowledge representation helps?, in: Y.-N. Chen, A. Rastogi (Eds.), Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1–11. URL: https://aclanthology.org/2023.nlp4convai-1.1/. doi:10.18653/v1/2023.nlp4convai-1.1.

[4] D. Ju, S. Feng, P. Lv, D. Wang, Y. Zhang, Learning to improve persona consistency in multi-party dialogue generation via text knowledge enhancement, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 298–309. URL: https://aclanthology.org/2022.coling-1.23/.

[5] N. Penzo, M. Sajedinia, B. Lepri, S. Tonelli, M. Guerini, Do LLMs suffer from multi-party hangover? a diagnostic approach to addressee recognition and response selection in conversations, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11210–11233. URL: https://aclanthology.org/2024.emnlp-main.628/. doi:10.18653/v1/2024.emnlp-main.628.

[6] Z. Yin, Q. Sun, Q. Guo, Z. Zeng, X. Li, T. Sun, C. Chang, Q. Cheng, D. Wang, X. Mou, X. Qiu, X. Huang, Aggregation of reasoning: A hierarchical framework for enhancing answer selection in large language models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 609–625. URL: https://aclanthology.org/2024.lrec-main.53/.

[7] Y. Feng, Z. Lu, B. Liu, L. Zhan, X.-M. Wu, Towards LLM-driven dialogue state tracking, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 739–755. URL: https://aclanthology.org/2023.emnlp-main.48/. doi:10.18653/v1/2023.emnlp-main.48.

[8] Z. Li, Z. Chen, M. Ross, P. Huber, S. Moon, Z. Lin, X. Dong, A. Sagar, X. Yan, P. Crook, Large language models as zero-shot dialogue state tracker through function calling, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 8688–8704. URL: https://aclanthology.org/2024.acl-long.471/. doi:10.18653/v1/2024.acl-long.471.

[9] Z. Hu, Q. He, R. Li, M. Zhao, L. Wang, Advancing multi-party dialogue framework with speaker-ware contrastive learning, 2025. URL: https://arxiv.org/abs/2501.11292. arXiv:2501.11292.

[10] S. Liu, P. Li, Y. Fan, Q. Zhu, Enhancing multi-party dialogue discourse parsing with explanation generation, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 1531–1544. URL: https://aclanthology.org/2025.coling-main.103/.

[11] S. M. Mousavi, G. Roccabruna, M. Lorandi, S. Caldarella, G. Riccardi, Evaluation of response generation models: Shouldn't it be shareable and replicable?, in: A. Bosselut, K. Chandu, K. Dhole, V. Gangal, S. Gehrmann, Y. Jernite, J. Novikova, L. Perez-Beltrachini (Eds.), Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 136–147. URL: https://aclanthology.org/2022.gem-1.12/. doi:10.18653/v1/2022.gem-1.12.

[12] K. Mahajan, S. Shaikh, Persona-aware multi-party

conversation response generation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 12712–12723. URL: https://aclanthology.org/2024.lrec-main.1113/.

[13] T. Sun, K. Qian, W. Wang, Contrastive speaker-aware learning for multi-party dialogue generation with llms, 2025. URL: https://arxiv.org/abs/2503.08842. arXiv:2503.08842.

[14] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, J. Kleinberg, Echoes of power: language effects and power differences in social interaction, in: Proceedings of the 21st International Conference on World Wide Web, WWW '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 699–708. URL: https://doi.org/10.1145/2187836.2187931. doi:10.1145/2187836.2187931.

[15] D. Antypas, A. Ushio, J. Camacho-Collados, V. Silva, L. Neves, F. Barbieri, Twitter topic classification, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3386–3400. URL: https://aclanthology.org/2022.coling-1.299/.

[16] Y. Zhao, T. Nasukawa, M. Muraoka, B. Bhattacharjee, A simple yet strong domain-agnostic debias method for zero-shot sentiment classification, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3923–3931. URL: https://aclanthology.org/2023.findings-acl.242/. doi:10.18653/v1/2023.findings-acl.242.

## Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Improve writing style and Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.