An LLM Benchmark for Addressee Recognition in Multi-modal Multi-party Dialogue

Koji Inoue, Divesh Lala, Mikey Elmers, Keiko Ochi, Tatsuya Kawahara
Kyoto University, Japan

Summary

- We introduce a novel multi-modal, multi-party dialogue corpus (TEIDAN) designed to advance research on spontaneous triadic conversations.
- Using this corpus, we benchmarked GPT-4o on the task of addressee recognition, revealing that the model performs only marginally above chance, highlighting the unique challenges of multi-party dialogue understanding.

TEIDAN Corpus

- The TEIDAN corpus is a new multimodal dataset consisting of spontaneous three-party conversations in Japanese.
- Each session captures natural and open discussions on everyday topics, recorded with synchronized audio and video from three participants.
- The corpus includes 30 sessions from 10 triads, with high-quality audio from individual microphones and gaze behavior captured via head-mounted cameras.

Annotation of Addressee

- A subset of the TEIDAN corpus was annotated to identify whether each turn had an explicit addressee.
- Step 1: Segment turns by concatenating IPUs and ignoring backchannels
- Step 2: Using both textual content and visual cues (gaze), label each turn as either addressed to a specific participant (A, B, or C), or as non-addressed (labeled O) when the turn was open to the group
- Only **75 turns** (**19.4%**) explicitly addressed a specific participant.

Session ID	#Turn (A/B/C)	#Addressed (A, B, or C)	#Not Addressed (O)
01-city	12 / 16 / 13	9	32
02-city	16 / 22 / 23	14	47
03-city	19 / 29 / 30	6	72
04-city	29 / 21 / 24	10	64
05-city	44 / 43 / 46	36	97
Total (Ave.)	387 (77.4)	75 (15)	312 (62.4)

Benchmark of Addressee Recognition

(1) Addressee Recognition with GPT-4o

- We evaluated GPT-4o on the addressee recognition task.
- The model was prompted to identify the intended addressee (A, B, C, or O) based on the preceding context of five utterances and the current speaker's turn.
- It achieved only **80.9% accuracy**, which is **marginally above the chance level** (80.6%).
- This suggests that even advanced LLMs struggle to interpret conversational roles in multi-party settings.

(2) Incorporating Gaze Information

- To examine the role of non-verbal cues, we also integrated automatically extracted gaze features using OpenFace 2.0.
- Gaze direction in the final second of each turn was added to each utterance in the prompt if the speaker was visually addressing another participant.
- However, **accuracy decreased to 75.2%**, indicating that naive integration of gaze features may not enhance performance.
- This highlights the **difficulty of leveraging multimodal signals** in a format that current LLMs can effectively interpret, and underscores the need for more robust integration methods or fine-tuning.

(3) Next Speaker Prediction

- We also assessed the model's ability to predict the actual next speaker, a task distinct from addressee recognition.
- The model was required to choose A, B, or C as the speaker who would take the next turn. GPT-4o achieved only **46.0% accuracy, below the 50% chance level**.

Addressee Recognition Performance of GPT-4o

LLM Output	#Correct	#Incorrect
Addressed (A/B/C)	9	14
Not addressed (O)	304	60

Correct OK

	Utterance
C	So, if we wanted to change the capital from Tokyo, where do you think would be a good place?
A	I think Osaka would be a good choice. Osaka is the largest city in western Japan, and in terms of population, there's no other city in western Japan that surpasses it. So, I think Osaka is a strong candidate.
B	But one of the reasons for wanting to relocate the capital from Tokyo is likely the population increase, or rather, Tokyo's population is becoming unmanageable, necessitating the transfer of some capital functions. (...) Hokkaido is a bit cold, though, so I think somewhere in Kyushu or, for example, the Tokai region might be better.
A	I see, that makes sense.
B	What do you think, <C's name>-san? Do you have any specific ideas? **(addressee is C)**

Incorrect ✗ (GPT-4o recognized as "O")

	Utterance
A	One of the reasons why I prefer Osaka is that its city planning, including roads and railway networks, is very linear and easy to understand.
C	Like Midosuji?
A	Exactly. If you've ever seen a map of the Tokyo subway, you'll know that it's quite convoluted and complex. In contrast, Osaka's layout is more grid-like.
C	With streets like ``something-suji'' and ``Something-suji Line.''
A	Yes. I think Tokyo is more circular, but a linear layout is easier to understand. Osaka's linear layout with clear divisions, like this area for administrative functions and this area as the central hub where people gather, makes it superior as a city, in my opinion.
C	I feel like in Nagoya, Sakae and Nagoya Station are slightly separated, aren't they? **(addressee is B)**