Next Speaker Prediction for Multi-Speaker Dialogue with Large Language Models

Lukas Hilgert **Jan Niehues**
Karlsruhe Institute of Technology, Germany
{lukas.hilgert, jan.niehues}@kit.edu

Abstract

Currently Large Language Models (LLMs) are mostly used through a chatbot interface with the user manually deciding when the system should respond. In multi-speaker conversations (e.g., two humans and one robot) it is not clear who speaks when. We therefore investigate the ability of LLMs to predict the dialog structure. First, we frame the task as Next Speaker Prediction (NSP) and create a multi-domain test set. Secondly, we build dedicated systems for the NSP task using LLMs and finally performed automatic and human evaluation. Our final system matches the human performance when tested on unseen data and exceeds it on data of the same domain as the training data.

1 Introduction

In multi-speaker dialogues, it is important for the participants to know when to speak, as talking at the wrong time may be irritating for the other speakers and may even hinder the speakers to reach their goals. It is crucial for dialogue systems to handle this task well as speaking too often may be annoying to the user while speaking rarely may seem unresponsive to the user and opposes the system's purpose.

Large Language Models (LLMs) are the core of modern dialogue systems. Currently they are mostly used through a chatbot interface where they only respond after the user sends a chat message. Here, there is no need for dedicated dialog structure modeling as the user always decides when the model should respond. For spoken dialogue with two speakers, the modeling is not as trivial as it is not clear when one speaker ends their turn. For multi-speaker scenarios it is significantly more challenging when the LLM should respond as the users could be chatting with each other directly during the course of dialogue.

P11: i used to live downtown san jose and every once in a while i just get with garlic and i don't know if it's from gilroy probably not nut i like to think it was so [laugh]
P09: yeah
P09: wow
P12: what are actually some nice places to go around here cause i've moved here recently so [unintelligible]
P09: napa napa is nice
P10: oh
P10: napa is nice [unintelligible]
P12: oh yeah actually i went there last week and they had uhhh i think sonoma had a hot air balloon festival there [...] but it's pretty nice seeing them at sunrise so yeah it was really beautiful yeah

Annotators' votes:
P09: 7, P11: 2, P10: 2

Zero-shot LLM: P12
Fine-tuned LLM: P09

Next utterance:
P09: people like to go wine tasting

Figure 1: Example of a part of a dialogue from DiPCo (Segbroeck et al., 2020). We show the previous utterances, which next speaker our human annotators predicted, what the LLMs in different setting predicted, and what the actual next speaker and utterance are.

We model this ability as the Next Speaker Prediction (NSP) task like Wei et al. (2023). We think it is a suitable proxy task as good performance on predicting the next speaker should indicate the quality of the system's ability to decide the correct time to actively contribute to the conversation.

We want to investigate multi-speaker dialogue from multiple domains to test generalization, estimate the performance of LLMs, and find out how well they have to perform. Therefore, our research on the NSP task covers the following aspects:

- We create a multi-domain benchmark for the NSP task utilizing multiple existing dialogue datasets.
- We run a user study with eleven annotators to gather a human baseline. This evaluation gives insights on the ambiguity of the task.

- We analyze the ability of various size LLMs to perform the NSP task and build dedicated models that reach or exceed our estimate of human performance.

2 Next Speaker Prediction Benchmark

To evaluate how well our approaches perform on the NSP task, we compile a benchmark consisting of multiple datasets. Using datasets from multiple domains enable us to estimate the generalization of the evaluated systems. Additionally, we collect a human baseline on subsamples of the datasets to get an estimate of the human performance on NSP. While the dialogue structure from the datasets offers a ground truth for the next speaker, we want to find out if human annotators would consider other options as equally possible. Also, we obtain an overview how ambiguous the task is for human annotators.

2.1 Datasets

For the NSP task, we need datasets of dialogues where the speaker is denoted for every utterance. As the following utterance then always determines which speaker will be the next after the current one, we can easily model the NSP task. For dialogues with only two speakers, the NSP task is fairly trivial. Therefore, we only investigated multi-speaker dialogue datasets.

We use three dialogue datasets for our benchmark (Table 1) to cover multiple domains. We chose these three datasets to cover multiple domains. Also, there is an existing baseline for the NSP task for MultiLIGHT (ML, Wei et al. (2023)). The other two datasets both include the type of noise that a dialogue system would also encounter in a real-life setting. Additionally, the conversation domains are realistic settings than ML's (Table 1). We use two of these similar datasets as this allows us to compare how well our approaches generalize an unseen domain and different noise as DiPCo includes no training data. The datasets differ in the numbers of participating speakers in one conversation, the domain of the conversation (topic, setting), and the amount of noise in the sense of very short utterances that introduce no or almost no substance to the conversation.

ML is a text-only dataset created specifically for dialogue research. The authors also performed experiments on the NSP task with at time of publication current Transformer-based language models that they fine-tuned on this task. The AMI meeting corpus (Carletta et al., 2005) and the Dinner Party Corpus (DiPCo) (Segbroeck et al., 2020) are primarily audio (and video for AMI) datasets from recorded conversations.

Dataset	AMI	DiPCo	ML
# Speakers	4	4	3
Domain	meeting	dinner party	fantasy role-play
Noisy	yes	yes	no
# Utterance	12627	3400	9164
# Dialogues	16	5	323
Avg. utts.	789.19	680.00	28.37

Table 1: Properties of the investigated datasets (specific numbers from the test splits). We list the number of speakers per dialogue, the topics of the conversations, if they contain some form of noise (short / interrupting utterances), and the number of utterances in total, the number of dialogues, and the average number of consecutive utterances per dialogue.

Dataset	AMI	DiPCo
Speaker 0	32.18	23.93
Speaker 1	26.88	25.75
Speaker 2	23.36	28.04
Speaker 3	18.75	22.28

Table 2: Contributed utterances (in percentage) from each speaker across all dialogues. For AMI, the speaker that speaks earlier in the dialogue, seems to have more dialogue utterances while there seems to be now such accumulation for DiPCo.

The type of conversations in AMI are meetings and in DiPCo dinner party talk. Both contain noise like "Umm", "Hmm", and "Yeah" that introduce no or almost no substance to the conversation in some cases. While these appear to happen at random times, these kinds of utterances are also present in a setting where an LLM gets its input via an Automatic Speech Recognition system. Also, for utterances like "Yeah" it is hard to determine if "Yeah" is just noise or an import acknowledgment of a previous utterance. So, we only filter out obvious irrelevant utterances for the DiPCo dataset like "[Noise]" to reduce the noisiness while keeping potentially important utterances.

Datasets Statistics In a first step, we investigated the dataset statistics in order to identify the various challenges of the datasets.

For example, we analyzed the percentage of con-

Dataset	AMI	DiPCo	ML
4	91.51	89.66	29.32
8	66.79	49.33	14.04
16	40.53	20.42	13.48
32	20.71	6.04	13.48
64	8.60	0.84	13.48

Table 3: Percentage of contexts where at least one speaker is missing depending on the number of recent utterances included in the prompt.

tributed utterances per speaker within each dialogue to see if one specific speaker speaks significantly more often which could lead to a bias to predict that speaker more often as the next one. We number the speakers ascending by their order of appearance. For AMI, the speakers that appear earlier in the conversation seem to speak more often. After qualitative analysis, we concluded that this is the case because in AMI the person opening the meeting is also the organizer of the meeting itself. We saw no such clear trend for DiPCo.

We want to only include the recent dialogue utterances in our benchmark as the dialogues in the datasets are up to several hundred utterances long (Table 1) which could overwhelm both human annotators and NLP systems. We therefore examined the number of times where at least one speaker is missing from our dialogue excerpt to find out in how many cases the context is missing information about some speakers. We start with four included recent utterances and iteratively double the amount up to 64. For ML, the number does not continue to decrease after 16 included utterances (Table 3). This is a result of the fact that in the beginning of the dialogue, not all speakers have spoken yet. As the dialogues in ML are short, this situation is quite common. For the other two, including quadratically more recent utterances linearly reduces the number of excerpts with missing speakers. This shows that very often in a small enough context window only a subset of the speakers interact with each other.

2.2 Human Baselines

In a first step, we analyze the difficulty of the task through a human evaluation. While the dialogues from the datasets were generated by humans, like many other Natural Language Processing (NLP) tasks, the NSP task is also ambiguous. We therefore collect human data on the NSP task for samples of consecutive utterances of the test splits of all three datasets. Our sample size is 63 dialogue utterances for AMI (0.50% of the full test set), 55 for DiPCo (2.00%), and 91 for ML (0.96%). As the dialogues in ML are fairly short, our sample includes three full dialogues. These sample sizes should in our opinion capture the natures of the datasets while also keeping the annotation work at a reasonable level. The user study involved eleven participants for each dataset. We average each's accuracy to get the human baseline (Table 6).

We included the last 32 utterances and did not rename the speakers in the prompts. We chose 32 as this number is higher than the number of utterances in full dialogues for the ML dataset and is not overwhelmingly large for human annotators. For the names, we assumed that the annotators should be able to distinguish the names more easily with the original ones from the dataset.

Dataset	Fleiss' kappa
AMI	0.17
DiPCo	0.14
ML 1	0.49
ML 2	0.43
ML 3	0.32

Table 4: Fleiss' kappa for multi-rater agreement on the samples used for the gathering the human baseline.

We provide the Fleiss' kappa multi-rater agreement measure (Fleiss, 1971) for each dataset sample (Table 4). For ML, we show the score for each of three dialogues that are included in our sample. The scores low showing the ambiguity of the task. The difference between ML and the other two datasets are in our opinion a result of it having fewer speakers per dialogue and having less noisy utterances. Manual inspection and anecdotal evidence from the annotators showed that the annotators agreed or were sure in their prediction respectively for some turns (most annotators picked one speaker) but disagreed or were unsure in their prediction respectively in other cases (annotators picked different speakers, no clear "favorite").

3 Next Speaker Prediction with LLMs

We want to use state-of-the-art technology to build a next speaker predictor. This leads to LLMs as they excel on other NLP tasks. Additionally, their task during the pre-training phase is predicting the next token which corresponds to predicting the next speaker when the prompt is a dialogue transcript

with annotated speakers. This implies that the NSP task is "natural" for LLMs given their training.

While the authors of ML perform similar experiments, they were with the smaller encoder-decoder language models R2C2 (Shuster et al., 2022), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020) which are smaller than today's models and were trained on less data and did not receive the extensive post-training of current LLMs. Furthermore, the authors of ML had to fine-tune these models to perform this task while current LLMs can be used with zero-shot prompts.

To model the NSP task as an LLM task, we prompt the LLMs to predict the next speaker by utilizing the information we provide (Appendix A): An instruction for the task and the most recent utterances of the current dialogue as context. Each utterance starts with the corresponding speaker. Each dataset already contains identifiers for the speakers. For ML, each speaker has a descriptive name like "jester". The other two datasets use string identifiers like "P12" or "MTD011UID".

For every dialogue turn we include the last eight utterances as context for the LLMs and rename the speakers to the same generic identifier across all datasets to increase the similarity of the task across the datasets. We replace them with renamings where each speaker has a pseudonym in the format of "speaker <number>". ML also includes descriptions of the character of each speaker and the location of the dialogue. We did not include this information to keep the task comparable.

Although LLMs are able to perform on zero-shot, often specialized models perform better. We therefore train LLMs supervised on the NSP task on multi-domain data by mixing the training splits from the AMI and the ML dataset. We use a balanced mixture (similar number of training data points) to ensure generalization across domains. We train with pairs of the prompt used in the zero-shot setting and the expected speaker from the datasets, so that the model learns how to map the recent dialogue turns to the next speaker.

4 Experiments

We evaluate the LLMs on the test splits of the datasets, compare them to random and human baselines, and perform ablation studies on our modeling decisions.

We chose next speaker accuracy as our main evaluation metrics as this is the most straightforward metric with the given data. As the distributions of utterances per speaker are fairly balanced (Table 2), we did not employ metrics like F_1. While accuracy is a "hard" metric and does not account for ambiguity, we assume that the fairly large dataset size and direct comparison against baselines still gives a good estimate how well the LLMs (and especially our fine-tuned one) do for NSP. Nevertheless, we analyze the agreement of the LLMs with the annotators (section 4.3).

4.1 Setup

We perform all our experiments with models from the Llama 3 family (Dubey et al., 2024). We use the 3B (3.2. 3B) and 8B (3.1 8B) parameter version for zero-shot and fine-tuning experiments while we use the bigger version (3.3 70B) only in a zero-shot setting as fine-tuning this model requires significantly more compute and the smaller models responded already very well to fine-tuning.

The fine-tuning data mixture consists of all the available training data from the AMI meeting corpus and 33% from ML. We use only 33% to balance the number of data points from each dataset. DiPCo has no train split. We conduct ablation studies on all mentioned modeling decisions including the preprocessing (subsection 4.4). We made these decisions that impacted our main results during development on the basis of the validation sets which all utilized datasets provide.

4.2 Random Baselines

To compare our results to another baseline, we present three random baselines. Each is designed to model two very distinct types of dialogue flow and a combination of both. These baselines are: One where the speaker is picked randomly but always switches after each dialogue utterance (denoted as **always**). Then, we assume that the speaker never switches, so we predict the last speaker to also be the next speaker (denoted as **never**). At last, we model a combination of both where we pick the speaker completely randomly without excluding the last speaker (denoted as **usually**). We run each method five times and average the results.

4.3 Main Experiments

We differentiate between the results on the full test sets and the samples for the human baseline.

Results on the full Tests Sets The accuracy scores (Table 5) for the random baselines illustrate

Dataset	AMI	DiPCo	ML
Random Baselines			
Always	22.10	26.62	45.91
Usually	25.17	25.21	33.36
Never	33.41	19.32	8.91
Zero-shot			
Llama 3.2 3B	25.28	25.66	28.10
Llama 3.1 8B	34.88	30.94	40.41
Llama 3.3 70B	35.81	32.98	52.06
Fine-tuned			
Llama 3.2 3B	45.91	36.91	59.40
Llama 3.1 8B	47.85	38.48	59.85

Table 5: NSP accuracy on the **full test splits**. We compare the accuracy of the random baselines and the Llama 3 models in a zero-shot and fine-tuned setting. Fine-tuning improves performance beyond the 70B model's performance. Even the dataset we did not train on (DiPCo) benefits from fine-tuning on the NSP task.

Dataset	AMI	DiPCo	ML
Human	30.88	33.22	48.65
Random Baselines			
Always	20.63	27.64	45.49
Usually	17.78	26.91	35.16
Never	32.06	14.55	11.87
Zero-shot			
Llama 3.2 3B	15.87	21.82	25.27
Llama 3.1 8B	34.92	23.64	32.97
Llama 3.3 70B	30.16	34.55	51.65
Fine-tuned			
Llama 3.2 3B	47.62	40.00	61.54
Llama 3.1 8B	58.73	34.55	59.34

Table 6: NSP accuracy on the samples of the test splits for the **human baseline**. We compare the accuracy of the human annotators, random baselines, and the Llama 3 models in a zero-shot and fine-tuned setting. Fine-tuning beats human accuracy on the datasets with training data but also on DiPCo.

what we already saw during qualitative analysis of the datasets: In the AMI meeting corpus, the speakers often deliver multiple utterances after another while in the ML dataset the speaker almost always switches. Llama 3.1 8B performs a bit or clearly better than the random baselines on AMI and DiPCo, which highlights the importance of a multi-domain benchmark. On ML however, simply randomly picking one of the other two speaker as the next performs better. The smallest model we tested (3.2 3B) only manages to predict next speaker as well as completely randomly picking one. The bigger 70B model outperforms the random baselines clearly on DiPCo and ML. We see a clear trend that scaling the model size increases the ability to predict the next speaker.

When fine-tuning 3.1 8B on the task, it significantly outperforms itself in the zero-shot setting, the random baselines, and the bigger version. The performance even improves beyond the 70B model on the DiPCo dataset, which has no training split meaning that this dataset is out-of-domain for the fine-tuned models, and we see generalization for different domains. The case for 3.2 3B is similar but with slightly lower scores than 3.1 8B.

Results on the Samples of Tests Sets for Human Baselines On DiPCo and ML, our collected human baseline outperforms the random baselines albeit not all of them by a big margin (Table 6). For AMI, it is even slightly below the best technique ("never") that assumes the last speaker will always be the next speaker. In the zero-shot setting, the smallest Llama model shows the same pattern as on the full test sets. The medium LLM however achieves a higher accuracy on AMI as the human baseline, while struggling to reach the random baseline on the other two datasets which may be specific to these samples. The 70B version roughly matches the human performance on all datasets. The scaling trends we observed on the full test sets is also present on the samples except for AMI, where the 70B model underperforms the 8B model.

Fine-tuning the two smaller models shows similar effects as we saw on the full test split: The NSP accuracy is increased greatly compared to the zero-shot setting and even sightly outperforms the 70B model on the datasets where training data exists. For DiPCo, the performance of Llama 3.1 8B is the same as the one of 3.3 70B. The fine-tuned 3B model manages to outperform both the 8B and 70B model on DiPCo. As it showed reduced performance compared to the 8B model on the full test sets and as this sample set is small, we assume that these differences between the models are partly noise while still showing the effectiveness of our fine-tuning in general for the NSP task.

Agreement of Annotators and LLMs As mentioned before, this task is a highly ambiguous task. However, there are also situations where only a small set of possible next speakers are correct. We

wanted to investigate this and therefore use the human annotations as additional references.

We analyze the agreement of the LLMs with the human annotators. To do this, we remove one annotator at a time from the pool of annotators. We then compare their agreement with the rest of the annotators and with the LLMs by measuring the accuracy of their predictions. We then average the results for all annotators.

Additionally, we show how many of the predictions can be counted as correct with these conditions which decreases with the number of required agreeing annotators increasing (row "Correct answers", column "all").

In this setup, we counted a prediction as correct if at least n annotators propose this prediction. This allows for situations where then no answer is correct and therefore it does not matter what the model predicts and for situation where multiple solutions are correct. Additionally, we show how many of the predictions can be counted as correct with these conditions which decreases with the number of required agreeing annotators increasing (row "Correct answers", column "all").

Also, we show how many choices a predictor has with the given threshold as for example only three possible next speakers can be counted as correct if the number of annotators is ten and the threshold for the number of agreeing annotator is three. Therefore, the number of possible correct answers also decreases with a higher threshold. The reported numbers for the annotators and the models display the percentage of correct predictions (given a threshold) out of the possible correct answers. We then also list the distribution of choices within this set – how many predictions are possibly correct. Per bin of possibly correct prediction, we also report the accuracy of each predictor.

For the AMI dataset, we see mixed results: From a threshold of three and more, the larger model has lower agreement than the 8B model. The fine-tuned model shows a similar regression for a threshold of three and five. For seven agreeing annotators, the fine-tuned model has a slightly higher agreement, yet the 70B model is lower than Llama 3.1 8B in zero-shot. We think that these results come from the fact that fine-tuning on the AMI training data pushed the 8B LLM towards the distribution by the dataset increasing the NSP accuracy, which disagrees with our human annotators. That the 70B model also has a lower agreement could be a sign of its training data containing part of AMI and it memorizing it better than the 8B model.

For DiPCo, we see that the 8B model in the fine-tuned setting has a clearly higher (threshold of one and three) or slightly higher (threshold of five) agreement than in the zero-shot setting (Table 8). Here, we also see that the 70B version has higher agreement than the 8B model in zero-shot. This matches our observations from the accuracy scores before that increased model sizes correlates with an improved NSP ability. Fine-tuning Llama 8B therefore improves for most tested thresholds the agreement with the human annotators on DiPCo and moves it closer to that of the 70B model. As we did not fine-tune the 8B model on data from DiPCo, we think that these results together with the increase in NSP accuracy show that training on the NSP task with dialogue datasets does generalize to better NSP performance – matching or exceeding human performance in NSP accuracy.

4.4 Ablation Studies

We also examine our modeling decisions when fine-tuning Llama 3.1 8B.

Dataset	AMI	DiPCo	ML
Speaker Renaming			
Original	42.04	39.35	54.58
Renamed	47.85	38.48	59.85
Context Length			
4	46.32	34.66	59.47
8	47.85	38.48	59.85
16	47.92	39.46	60.13
32	47.58	37.86	60.21
64	46.72	36.29	59.73
Training Data Mixture			
Zero-shot	34.88	30.94	40.41
AMI	47.84	37.35	42.75
ML	24.08	28.25	60.07
AMI + 33% ML	47.85	38.48	59.85

Table 9: Comparison of the accuracy results from the ablation studies. Renaming the speakers to a dataset-across scheme increases performance in general. Including more previous utterances in the prompt only helps until 16 utterances. Training only on one of the two available datasets is worse than using both.

Speaker Renaming We compare the unmodified versions of the datasets with our renamed versions. Renaming improves performance on all datasets except for DiPCo (Table 9). This is probably the case as the speaker names in DiPCo (e.g., “P12”) are already fairly generic but distinct. This also

# choices	all	1	2	3	4
At least one out of ten agreeing annotator					
Correct answers	100.00	2.31	21.07	43.00	33.62
Annotators	92.06	68.75	86.99	89.60	100.00
Zero-shot 8B	88.46	75.00	63.70	92.28	100.00
Fine-tuned 8B	88.74	81.25	87.67	80.87	100.00
Zero-shot 70B	89.32	75.00	69.18	91.61	100.00
At least three of ten agreeing annotator					
Correct answers	100.00	42.14	54.98	2.89	0.00
Annotators	68.40	56.51	76.12	95.00	0.00
Zero-shot 8B	66.67	43.49	82.68	100.00	0.00
Fine-tuned 8B	54.83	44.86	61.42	75.00	0.00
Zero-shot 70B	61.18	34.93	79.27	100.00	0.00
At least five of ten agreeing annotator					
Correct answers	82.40	98.42	1.58	0.00	0.00
Annotators	56.39	56.23	66.67	0.00	0.00
Zero-shot 8B	45.01	44.13	100.00	0.00	0.00
Fine-tuned 8B	39.93	39.15	88.89	0.00	0.00
Zero-shot 70B	39.23	39.32	33.33	0.00	0.00
At least seven of ten agreeing annotator					
Correct answers	29.29	100.00	0.00	0.00	0.00
Annotators	59.61	59.61	0.00	0.00	0.00
Zero-shot 8B	44.33	44.33	0.00	0.00	0.00
Fine-tuned 8B	45.81	45.81	0.00	0.00	0.00
Zero-shot 70B	42.36	42.36	0.00	0.00	0.00

Table 7: Agreement between annotators and LLMs (**AMI**): We show the NSP accuracy for each annotator (results averaged) and the LLMs when the other annotators serve as the ground truth. We show different thresholds for agreeing annotators that an answer counts as correct. We also display the accuracy grouped by the number of choices a predictor has (if too many annotators have to agree, the number of possible correct answers shrink).

means that not renaming the speakers for the user study should not skew our comparison.

Context Length We also compare how the number of included most recent dialogue utterances influences the accuracy of the predictions: We vary the number of included utterances in the prompt as context for the models in steps of the power of two from four to 64. There seems to be a limit on how much context in the form of previous dialogue utterances helps the model in its decision even with the number of not included speakers decreasing (Table 3). We picked eight recent utterances for our experiments as it showed the best performance on the validation sets, and it enables faster inference than for 16 utterances. As the accuracies differ only sightly across the context lengths we tried, it seems that the model mostly relies on the last few utterances for its decision while also being able to focus on them even if the included dialogue is longer.

Training Data Mixture As we have two datasets from our benchmark with training data, we want to find out how the specific selection of training data impacts the generalization ability of the fine-tuned models. Only training on the AMI data already shows large improvements for the two similar datasets (AMi and DiPCo) but only small improvements for ML. Only training on this dataset however reduces the performance on the other two datasets. A weighted combination of both datasets (roughly equal amount of datapoints from both) resulted in performance similar like training on the "corresponding" dataset. We even saw slight transfer learning for DiPCo.

5 Related Work

Previous research on dialogue turns is different from our approach as we assume both the setting of a multi-speaker dialogue either in text form or as a transcript and a text-only LLM as the predictor.

# choices	all	1	2	3	4
At least one of ten agreeing annotator					
Correct answers	100.00	0.00	5.95	50.58	43.47
Annotators	92.07	0.00	61.11	88.89	100.00
Zero-shot 8B	92.56	0.00	88.89	86.60	100.00
Fine-tuned 8B	98.18	0.00	94.44	97.06	100.00
Zero-shot 70B	96.53	0.00	61.11	97.71	100.00
At least three of ten agreeing annotator					
Correct answers	100.00	44.30	51.74	3.97	0.00
Annotators	61.98	54.85	67.73	66.67	0.00
Zero-shot 8B	55.21	37.69	67.41	91.67	0.00
Fine-tuned 8B	61.98	50.00	70.93	79.17	0.00
Zero-shot 70B	61.98	48.13	72.52	79.17	0.00
At least five of ten agreeing annotator					
Correct answers	70.74	100.00	0.00	0.00	0.00
Annotators	55.61	55.61	0.00	0.00	0.00
Zero-shot 8B	46.26	46.26	0.00	0.00	0.00
Fine-tuned 8B	46.73	46.73	0.00	0.00	0.00
Zero-shot 70B	53.04	53.04	0.00	0.00	0.00
At least seven of ten agreeing annotator					
Correct answers	25.12	100.00	0.00	0.00	0.00
Annotators	43.42	43.42	0.00	0.00	0.00
Zero-shot 8B	45.39	45.39	0.00	0.00	0.00
Fine-tuned 8B	42.76	42.76	0.00	0.00	0.00
Zero-shot 70B	50.66	50.66	0.00	0.00	0.00

Table 8: Agreement between annotators and LLMs (**DiPCo**): We show the NSP accuracy for each annotator (results averaged) and the LLMs when the other annotators serve as the ground truth. We show different thresholds for agreeing annotators that an answer counts as correct. We also display the accuracy grouped by the number of choices a predictor has (if too many annotators have to agree, the number of possible correct answers shrink).

Transition Relevance Places Methods for turn-taking use LLMs to predict transition-relevant places within a stream of words. Transition-relevant places are points in a dialogue where a turn-shift can happen. Ekstedt and Skantze (2020) fine-tuned GPT-2 to predict these spots in written and spoken dialogues. Later work (Umair et al., 2024) investigated if more recent LLMs (e.g., Llama 3.1 8B) can do the same.

Audio / Visual Cues Multimodal approaches for NSP use visual cues like gaze and hand gestures (Ishii et al., 2016; Malik et al., 2020). This research incorporates gaze transition patterns and eye contact timing structure (Ishii et al., 2016) or head movement (Ishii et al., 2015) to predict the next speaker using support vector machines. Malik et al. (2020) utilized focus of attention among others to train classic machine learning classifiers for NSP. Other systems rely on voice activity projection for turn-taking prediction (Inoue et al., 2024a,b) which predicts future voice activity based on the current audio signal.

6 Conclusion

Our research goal was to investigate the ability of LLMs to predict the next speaker in a multi-speaker dialogue setting. We also compared their performance with humans and fine-tuned LLMs to improve them on NSP. The experiments on our compiled benchmark show that LLMs like Llama 3.3 70B can match the human performance on the NSP task in accuracy and it also shows very high agreement with human predictors. Smaller LLMs can achieve this performance or even exceed it by fine-tuning on dialogue datasets when the dialogue flow (e.g., with some short noisy utterances) is similar. We think that these results imply an ability of LLMs to "know" when to talk at transition-relevant places in a multi-speaker dialogue – either through large model size or fine-tuning on dialogues. Fu-

ture work will investigate how multimodal LLMs handle the NSP task as this work did not investigate the impact of additional auditory and visual information about the dialogue.

Limitations

Our investigation is limited to text-only dialogues and does not cover the use of audio or visual cues. We do not predict the next speaker on a per-token or per-word basis but rather after a full utterance. This assume that the system only receives full utterances as input which is the case if the dialogue participants interact via text or through an audio transcript.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback.
This work was partially funded by the Ministry of Science, Research and Arts Baden-Württemberg (MWK BW) as part of the state's "digital@bw" digitization strategy.

References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried M. Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 herd of models. *CoRR*, abs/2407.21783.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2981–2990. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024a. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11873–11883. ELRA and ICCL.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024b. Real-time and continuous turn-taking prediction using voice activity projection. *CoRR*, abs/2401.04868.

Ryo Ishii, Shiro Kumano, and Kazuhiro Otsuka. 2015. Predicting next speaker based on head movement in multi-party meetings. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 2319–2323. IEEE.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, 6(1):4:1–4:31.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training

for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Usman Malik, Julien Saunier, Kotaro Funakoshi, and Alexandre Pauchet. 2020. Who speaks next? turn change and next speaker prediction in multimodal multiparty interaction. In *32nd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2020, Baltimore, MD, USA, November 9-11, 2020*, pages 349–354. IEEE.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. 2020. DiPCo - dinner party corpus. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 434–436. ISCA.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 373–393. Association for Computational Linguistics.

Muhammad Umair, Vasanth Sarathy, and Jan Peter de Ruiter. 2024. Large language models know what to say but not when to speak. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 15503–15514. Association for Computational Linguistics.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multiparty chat: Conversational agents in group settings with humans and models. *CoRR*, abs/2304.13835.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

A Prompt

Here, we present the prompt that both the tested LLMs and the human participants received to complete the NSP task:

```
Your task is to predict the next speaker given the full conversation history. Do not provide any explanation. Do not complete the conversation.

This is the conversation history:

<conversation history>

Predict the next speaker by outputting the name and only the name of the next speaker. Carefully consider the motives of the participating speakers in the conversation. Do not provide any explanation. Do not complete the conversation.
```

B Inference and Training Details

- Hugging Face Transformers library (Wolf et al., 2020) for loading and running the models.[1]
- Inference
  - All models were loaded in 8-bit precision via bitsandbytes. [2]
  - Temperature: 0.0 (no sampling)
- Training
  - Supervised Fine-tuning Trainer script from Hugging Face Transformer Reinforcement Learning library. [3]
  - LoRA (Hu et al., 2022) with rank $r = 8$.
- Hardware equipment: Up to two NVIDIA RTX 6000 Ada Generation GPUs at the same time.

C Data Collection for Human Baseline

We describe our process of collecting data for the human baseline in detail.

[1] https://github.com/huggingface/transformers

[2] https://github.com/bitsandbytes-foundation/bitsandbytes

[3] https://github.com/huggingface/trl/

C.1 Sample Selection

We targeted a sample of 1% of each test sets to keep the amount of work for the voluntary annotators small while still capturing the nature of the datasets. However, the different natures added additional constraints. For ML, we only selected three full dialogues leading to approximately 1% of the data. For AMI, a sample of 1% would have been outside of our annotator budget. Therefore, we selected a sample of 0.5%. For DiPCo, 1% was not enough to capture the dataset's nature, so we doubled the sample size here.

To decide which samples of the test sets to use during data collection, we performed several random samples of consecutive dialogue utterances and selected the one showing the most similar accuracy in a zero-shot setting to the full dataset.

C.2 Annotation Acquisition

We asked colleagues working in the field of NLP and Computer Vision to fill out the forms for our user study to acquire a human baseline. The participation was not mandatory, and we offered no compensation. We informed the participants that the data created by them during this user study will be incorporated into a scientific publication.

We presented the participants of our data collection for the human baseline the following introduction texts:

- **Human Baseline for Next Speaker Prediction on the AMI Meeting Corpus Dataset**
  I want to acquire a human baseline for the task of next speaker prediction on (a sample of) the AMI Meeting Corpus Dataset (https://groups.inf.ed.ac.uk/ami/corpus/). You will be presented the same prompt as the models I am testing. You will then be asked to select the next speaker out of the given ones. Please read the instructions in the first prompt carefully. The following questions (63 in total) will have the same prompt and will only change the newest (and oldest) conversation step.
- **Human Baseline for Next Speaker Prediction on the Dinner Party Corpus (DiPCo) Dataset**
  I want to acquire a human baseline for the task of next speaker prediction on (a sample of) the Dinner Party Corpus (DiPCo) Dataset (https://arxiv.org/abs/1909.13447). You will

be presented the same prompt as the models I am testing. You will then be asked to select the next speaker out of the given ones. Please read the instructions in the first prompt carefully. The following questions (55 in total) will have the same prompt and will only change the newest (and oldest) conversation step.

- **Human Baseline for Next Speaker Prediction on the MultiLIGHT Dataset 1/3**
  I want to acquire a human baseline for the task of next speaker prediction on (a sample of) the MultiLIGHT dataset (https://arxiv.org/abs/2304.13835). This is the first of three full conversations I want your help for. You will be presented the same prompt as the models I am testing. You will then be asked to select the next speaker out of the given ones. Please only enter your choice after your prediction is final (as picking a choice will show the next conversation state). Please read the instructions in the first prompt carefully (whose conversation history will be empty). The following questions (26 in total) will only show the newest conversation step. You can use the conversation steps of the previous questions for your decision for the current question (as the models also get the full conversation history for the current conversation state).

- **Human Baseline for Next Speaker Prediction on the MultiLIGHT Dataset 2/3**
  I want to acquire a human baseline for the task of next speaker prediction on (a sample of) the MultiLIGHT dataset (https://arxiv.org/abs/2304.13835). This is the second of three full conversations I want your help for. You will be presented the same prompt as the models I am testing. You will then be asked to select the next speaker out of the given ones. Please only enter your choice after your prediction is final (as picking a choice will show the next conversation state). Please read the instructions in the first prompt carefully (whose conversation history will be empty). The following questions (31 in total) will only show the newest conversation step. You can use the conversation steps of the previous questions for your decision for the current question (as the models also get the full conversation history for the current conversation state).

- **Human Baseline for Next Speaker Prediction on the MultiLIGHT Dataset 3/3**
  I want to acquire a human baseline for the task of next speaker prediction on (a sample of) the MultiLIGHT dataset (https://arxiv.org/abs/2304.13835). This is the third of three full conversations I want your help for. You will be presented the same prompt as the models I am testing. You will then be asked to select the next speaker out of the given ones. Please only enter your choice after your prediction is final (as picking a choice will show the next conversation state). Please read the instructions in the first prompt carefully (whose conversation history will be empty). The following questions (34 in total) will only show the newest conversation step. You can use the conversation steps of the previous questions for your decision for the current question (as the models also get the full conversation history for the current conversation state).

The introduction text for ML differs from the other datasets as we used a different setup for the online form. This switch from the setup for ML to the one used for AMI and DiPCo was mostly done out of convenience during the creation of the online form and should not impact the results of the data collection.

After this introduction text, the participants were shown the exact same prompt template as they were presented to the LLMs (subsection 2.2). To select the next speaker, they could choose from all appearing speakers in that dialogue with a radio button control element.