Universiteit Leiden
The Netherlands

Data Science & AI

LLM performance on multi-interlocutor NLI tasks

Bachelor thesis

Alexander Snepvangers

Supervisors:
Dr. Gijs Wijnholds, Dr. Miros Zohrehvand

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

08/06/2026

Abstract

Large Language Models (LLMs) have demonstrated impressive performance on a wide range of natural language processing (NLP) benchmarks. That being said, their ability to perform inference in multi-speaker dialogues remains poorly understood. This thesis investigates whether increasing dialogue complexity by increasing the number of unique speakers influences performance of LLMs on the Dialogue Natural Language Inference (DNLI) task. Using the DNLI dataset, which consists of faithfully transcribed multi-party conversations from the BNC2014 and CHILDES corpora, we evaluate five state-of-the-art (SotA) LLMs (deepseek-r1:8b, hermes3:8b, mistral:7b-instruct, llama3:8b, and gpt-oss:20b) alongside two older models (llama2:7b and zephyr:7b). Across three controlled experiments; a baseline replication, a unique speaker count test (2 to 7 speakers with context length held constant), and an incremental utterance test, we find that all models plateau at approximately 60% accuracy, regardless of dialogue complexity. Increasing the number of speakers produces no systematic decline in performance, and accuracy does not show a discrete jump when the last speaker (whose perspective the hypothesis is annotated from) first appears. This ceiling closely mirrors the hypothesis-only baseline reported in the original DNLI paper, strongly indicating that models exploit annotation artifacts and surface-level cues rather than building representations of speaker beliefs and perspectives. We conclude that genuine multi-speaker dialogue inference remains an unsolved problem for current LLMs under standard prompting strategies.

Contents

1 Introduction

Large Language Models (LLMs) have become the dominant models in Natural Language Processing (NLP), achieving impressive results across a wide range of benchmarks and tasks. However, before this, the ability to perform Natural Language Inference (NLI) was considered a core benchmark for assessing Natural Language Understanding (NLU). Models would be tested on many different kinds of NLI tasks, which asked a model to determine whether a hypothesis is entailed by, contradicted by, or neutral with respect to a premise. However, as LLMs have grown in scale and capability, NLI benchmarks have largely been set aside. Madaan et al. [MES+25] argues that this abandonment is premature. NLI benchmarks remain informative for distinguishing models across different scales and stages of training. This makes NLI a compelling proxy for LLMs understanding.

Standard NLI benchmarks are typically constructed from single written sentences and fail to capture the specific challenges of natural spoken dialogue. Spoken dialogue is fundamentally different from written text. Spoken dialogue builds meaning in a collaborative, incremental, and distributed way. An actor attempting to understand a dialogue must also understand the beliefs, intentions, and perspectives of the other actors present in the dialogue. Therefore, effective dialogue inference requires a model to track not just what has been said, but by whom, in what context, and from whose point of view. This represents a significant step up in complexity from sentence-level NLI. Recent work by Qamar et al. [QTH25] demonstrates LLMs struggle with the fundamental prerequisites of dialogue understanding and often fail to outperform simple rule-based baselines. A significant gap between human and LLM performance on these tasks further highlights how far current models are from genuine dialogue comprehension.

The Dialogue Natural Language Inference (DNLI) dataset, introduced by Ek et al. [ENC+24], provides a principled testbed for studying NLI in the context of natural spoken dialogue. Drawn from the BNC2014 [LDH+17] and CHILDES [Mac00] corpora, the dataset consists of faithfully transcribed multi-party conversations, complete with repairs, disfluencies, and non-standard speech. Initial experiments with `Llama 2:7b` and `Zephyr:7b` showed modest performance.
This paper builds directly on the DNLI work, extending the preliminary LLM evaluation in two key directions that have not yet been studied. First, this paper investigates how accuracy changes as the number of unique speakers in a dialogue increases from 2 to 7. If LLMs are genuinely tracking speaker beliefs and perspectives, adding more speakers should increase the complexity of inference and drive accuracy down. Second, this paper examines how LLM predictions evolve as a premise is incrementally revealed utterance by utterance. This gives a more granular analysis on model response and could give more insight on how LLMs come to their final prediction.
These experiments address the overarching research question:

- *To what extent does the complexity of a multi-speaker dialogue influence the inference of Large Language Models?*
  - *How does LLM accuracy on the DNLI dataset change as the number of speakers increases from 2 to 7?*
  - *At what point in a conversation (how many utterances) does an LLM typically find enough information to make a correct inference?*

The results across all experiments show that increasing dialogue complexity has no systematic effect on LLM accuracy. Instead, all LLMs tested plateau at around $\sim 60\%$ accuracy. This almost perfectly matches with a hypothesis-only result done by the original DNLI paper [ENC+24], which essentially checks to see if the dataset consists of clues that the models can use to avoid genuinely understanding the dialogue. This strongly suggests that current LLMs are not building structured representations of speaker perspectives, but are instead exploiting surface level features of the hypothesis. The key contributions of this paper are a replication and extension of the DNLI baseline to five modern state-of-the-art (SotA) LLMs; the first controlled evaluation of LLM performance on the DNLI task across unique speaker count strata; an incremental utterance experiment that tests for evidence of adaptive belief updating; and ultimately evidence that genuine multi-speaker dialogue inference remains an unsolved problem for current LLMs under the current prompting strategies.

1.1 Thesis overview

Section 2 discusses related work; Section 3 describes the methodology; Section 4 details the experimental settings; Section 5 presents the results; Section 6 discusses the findings; Section 7 concludes.

2 Related Work

In recent years since the release of ChatGPT, the use of NLI as a performance measure for NLU in LLMs has largely been neglected. However, Madaan et al. [MES+25] recently demonstrated that NLI benchmarks remain a valid way to evaluate the NLU of SotA LLMs.

As emphasized in [GSL+18] a long-standing problem with NLI datasets is annotation artifacts, which are statistical regularities in the hypothesis that models can exploit to predict the correct label without reading the premise at all. These arise because systemic linguistic patterns form when annotators are instructed to write constrained examples at scale. For example, a contradiction hypotheses attracts negation words. A premise and hypothesis that share lexical content are easier to classify as entailment [MPL19]. Poliak et al. [PNH+18] demonstrated that a hypothesis-only model significantly outperformed the majority-class baseline on 6 out of 10 NLI datasets. Therefore, this context is important to keep in mind when interpreting any NLI results.

Despite the renewed promise of NLI benchmarks, standard ones often rely on written single-sentence premises and miss the specific challenges posed by natural spoken dialogue. A recent paper [QTH25] shows that when tasked with understanding an utterance in a dialogue, LLMs fail to recognize speaker roles, the broader function of an utterance, and show a bias toward the most recent utterances. Furthermore, a significant gap was found between human and LLM results on these tasks.

This leads to the key paper [ENC+24] this work builds off. This paper provides the natural dialogue-NLI (DNLI) dataset and the preliminary LLM tests that opens the door for more rigorous dialogue testing of modern LLMs. The LLMs, `Llama 2:7b` and `Zephyr:7b`, were tested using few-shot prompts. Llama 2 barely surpassed the majority class baseline at around 33% accuracy, while Zephyr rises from 42% up to 60% across bins organized by utterance count in the premise. Critically, the DNLI dataset is not immune to annotation artifacts. The paper explicitly tests a hypothesis-only

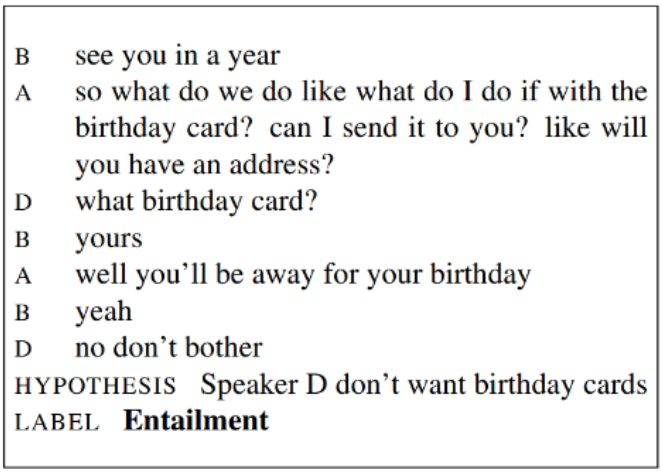


Figure 1: Example of a Dialogue Natural Language Inference (DNLI) task. The premise is a faithfully transcribed natural speech dialogue. The hypothesis and label are annotated from the perspective of the last speaker in the premise.

baseline and finds that a BERT-based classifier achieves 58.9%, well above the 33% majority class baseline. This strongly suggests hypothesis-level artifacts are present in the dataset, which is crucial to keep in mind when analyzing the results.

2.1 DNLI Dataset

The DNLI data was sourced from BNC2014 [LDH+17] and CHILDES [Mac00] corpora. BNC contains faithfully annotated natural language conversations between adult native speakers in British English, which means repairs, disfluencies, and other speech errors are included in the data. The portion of the CHILDES that is added to the DNLI data contains conversations between English-speaking 2 and 5 year olds from sub-urban Atlanta. Including speakers with less linguistic resources forces models to adjust how they interpret meaning, even from less conventional speech. Each dialogue was split into n sub-dialogues of 50 utterances each. From the 50 utterances, 1–5 utterances were chosen at random to be annotated. Annotators from Amazon Mechanical Turk and Master students in the Language Technology program at the University of Gothenburg were instructed to take the perspective of the last speaker and write a hypothesis statement conforming to one of the 3 labels: entailment, contradiction, or neutral.
In total the dataset contains 14,179 DNLI entries with a nearly uniform label distribution as can be seen in table 1. Furthermore, as can be seen in table 2 the CHILDES annotations constitutes a small part of the total annotations.

Table 1: Distribution of labels

Label	Count	Proportion
Entailment	4799	0.338
Contradiction	4677	0.329
Neutral	4723	0.333

Table 2: Distribution across sources

Source	Dialogues	Annotations
BNC	938	13892
CHILDES	17	287

3 Methodology

3.1 Baseline

A baseline test is replicated on new SotA LLMs as well as the 2 older LLMs used in the DNLI paper [ENC+24]. One of those older LLMs, Zephyr, shows accuracy rising from around 40% to 60% as the utterance count increased. Therefore, there is an expectation to see similar results for Zephyr and LLama 2, and an expectation to see improved results from the newer SotA LLMs. In general, this baseline will not directly address any of the research questions set for this paper. However, if the amount of utterances in the premise is seen as a light proxy for complexity, then this test might gather insight on the main research question: *To what extent does the complexity of a multi-speaker dialogue influence the inference of Large Language Models?*

3.2 Unique Speaker Count Testing

LLMs are tested on groups of DNLI data with different unique speaker amounts. The DNLI paper [ENC+24] notes that accurately predicting labels requires a model to compose information given over several turns and take into account a speaker's point of view. Therefore, assuming that LLMs are trying to model speaker beliefs as is necessary to complete the DNLI task at high accuracy, we expect to see accuracy drop as the amount of unique speakers increases. In this way, the amount of unique speakers in the dialogue serves as a general metric for the complexity of the dialogue due to the extra belief states that must be tracked in order to predict labels at a high accuracy. Therefore, this test will address the sub-question, *How does LLM accuracy on the DNLI dataset change as the number of speakers increases from 2 to 7?*, as well as the main research question: *To what extent does the complexity of a multi-speaker dialogue influence the inference of Large Language Models?*

3.3 Incremental Utterance Testing

As a final experiment, the LLMs are asked to give predictions as the premise reveals itself utterance by utterance. This allows for a more granular analysis of each utterance in a premise and might give more insight on how LLMs arrive at their final predicted label. The early utterances are expected to have a low accuracy because the hypothesis is written from the perspective of the last speaker, which means that the model can only really start building a correct prediction after that final speaker starts to reveal itself in the premise. Therefore, if the model is tracking speaker beliefs we expect to see a steady jump in accuracy after the last speaker contributes to the dialogue for the first time. If not, then it is unlikely that the LLMs are attempting to track speaker belief states and could be instead exploiting well-documented annotation artifacts in NLI [PNH+18]. This test will address the sub-question, *At what point in a conversation (how many utterances) does an LLM typically find enough information to make a correct inference?*, as well as the main research question: *To what extent does the complexity of a multi-speaker dialogue influence the inference of Large Language Models?*

4 Experimental Settings

Across all the experiments the following 5 SotA LLMs are used: `deepseek-r1:8b`, `hermes3:8b`, `mistral:7b-instruct`, `llama3:8b`, and `gpt-oss:20b`. Additionally the 2 models from the DNLI paper [ENC+24], `llama2:7b` and `zephyr:7b`, are used to show comparisons across model versions, but are mainly used to replicate the baseline experiment from the DNLI paper. These 5 SotA models were chosen for a multitude of reasons.
Firstly, barring `gpt-oss:20b`, they all have similar parameter counts, which avoids the results being influenced by parameter size. Secondly, they all come from different families of LLMs, which allows for the analysis across different model types and architectures. Finally, `gpt-oss:20b` was chosen as a reference for how the results change with higher parameter models.
The DNLI data was loaded and processed such that each data entry included a column that indicated the amount of unique speakers in the premise. The 7 models were given a baseline test, where accuracy was measured against utterance length. The 5 SotA models were then tested for accuracy across unique speaker counts. Finally, 3 models were tested incrementally as a premise is slowly revealed to them, utterance by utterance. All of these tests will be explained in further detail in the following subsections.

4.1 Data Stratification/Loading

The entire test, validation, and train DNLI data splits are loaded into one csv file. However, the files containing only dialogues with a single utterance are left out. This main data file is used for the Baseline experiment; for the subsequent experiments the data is re-filtered to only have premise character lengths between 280–350. This range was decided upon because the average characters a premise had across the unique speaker strata were between 287 (for 2 speakers) up to 355 (for 6+ speakers). Table 3 shows the descriptive statistics for the subsequent data pool.

unique speaker count	avg char length	# of entries
2	313.92	9841
3	313.88	5878
4	313.67	1917
5	314.33	317
6–7	312.29	68 + 17

Table 3: Descriptive statistics of the length-filtered data pool. Each row shows the average character length in the premise and number of entries for each strata. Note that the average character length across all strata is around 313, which is crucial in making unique speaker count independent from the context size.

The main data pool is only used for the baseline experiment, whereas the length-filtered data pool is used for the unique speaker testing and incremental utterance testing. This is done such that the results are not affected by premise utterance length.

4.2 Prompting

The few-shot prompts fed to the LLMs during the tests are slightly different from the one shown in the DNLI paper. While the DNLI paper uses examples from the training set, custom examples

were created that don't appear in the data pool at all (Figure 2). The initial prompt is also a little different and is designed to discourage long answers from the LLM.

You are a strict NLI classifier. Output exactly one word: Entailment, Neutral, or Contradiction. Do not output any explanation. Given a dialogue excerpt and a Hypothesis, decide on the semantic relation between them.

SPEAKER A: Hello
SPEAKER B: Hi there
SPEAKER A: How are you
SPEAKER B: I'm good thanks
HYPOTHESIS: The speakers are greeting each other
RELATION: Entailment

SPEAKER A: I love pizza
SPEAKER B: me too
SPEAKER A: What's your favorite topping
SPEAKER B: I hate all toppings
HYPOTHESIS: Both speakers like pizza with toppings
RELATION: Contradiction

SPEAKER A: What time is it
SPEAKER B: Around 3 PM
SPEAKER A: Thanks
SPEAKER B: No problem
HYPOTHESIS: The weather is good
RELATION: Neutral

SPEAKER A: so james was up in th-
SPEAKER B: no no not james, mathew
SPEAKER C: yeah
SPEAKER A: ahh
SPEAKER B: he was trying to get hotdogs
HYPOTHESIS: Speaker A likes hotdogs
RELATION: [FILL]

Figure 2: Standard LLM few-shot prompt shown to each model for each data entry.

4.3 Baseline

The baseline experiments for `zephyr:7b` and `llama2:7b`, shown in the DNLI paper, were replicated by testing for accuracy across bins of DNLI data separated by amount of utterances in the premise. For each random seed out of 2, a sample of 50 DNLI entries was taken from each bin and tested on both models. Next to the 2 older models, the other 5 SotA models are tested on this baseline as well.

4.4 Unique Speaker Count Testing

For each random seed, each model gets few-shot tested on 85 sampled entries from each unique speaker count stratum. This process is repeated across 3 different random seeds and averaged to get the final plot that shows the accuracy of the model across different speaker counts. Additionally, for this experiment in particular the data has been filtered to only allow for premise character lengths between 280 and 350, which should decouple the relationship of increasing speaker count with increasing context length.

4.5 Incremental Utterance Testing

In this experiment each LLM gets prompted in a chat setting, where the first prompt contains the regular few-shot examples (Figure 2), but the test entry includes a partial premise which only includes the first utterance, u_1. In this case u_1 would be `"Speaker A: so james was up in th-"`. After the LLM responds with its prediction, the test premise together with the hypothesis is sent again, but this time the premise contains u_1 and u_2. In which case following the test example in figure 2 would be `"Speaker A: so james was up in th- Speaker B: no no not james, mathew`. This goes on until there is a prediction for every premise size until the total premise length is reached. A new chat is made for every new test entry. For each random seed out of 3, each model gets 30 random samples from the total pool of data.
After the all the results are gathered normalization is performed by computing an utterance offset for each prediction in an entry because the test dialogues have different amounts of utterances. The offset is equivalent to the difference between the current utterance and the utterance at which the last speaker enters the dialogue for the first time. Importantly, each entry contributes multiple data points to the analysis, one for each incremental utterance revealed. Therefore, as seen in Table 4, predictions at different absolute utterance numbers can map to the same offset if they occur at equivalent distances from the last speaker's initial utterance.

Entry	Offset -2	Offset -1	Offset 0	Offset +1	Offset +2
A (Last speaker @ utt. 3)	Utt. 1	Utt. 2	Utt. 3	Utt. 4	Utt. 5
B (Last speaker @ utt. 7)	Utt. 5	Utt. 6	Utt. 7	Utt. 8	Utt. 9

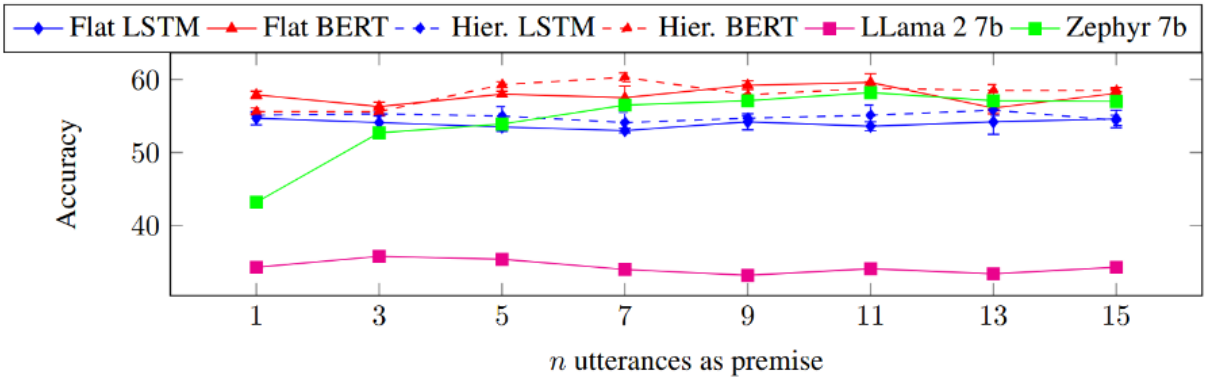
Table 4: Mapping of utterance numbers to normalized offsets for two entries. When aggregated across all entries, accuracy at each offset represents the mean correctness across all predictions at that relative distance from the last speaker's first appearance.

The last speaker is important because the hypothesis is annotated from the perspective of the last speaker. This means that predicting the hypothesis before the last speaker appears in the premise should be very difficult. However, once the last speaker comes up in the premise then the accuracy should experience a jump, because only then does the context for the hypothesis start to become clearer.

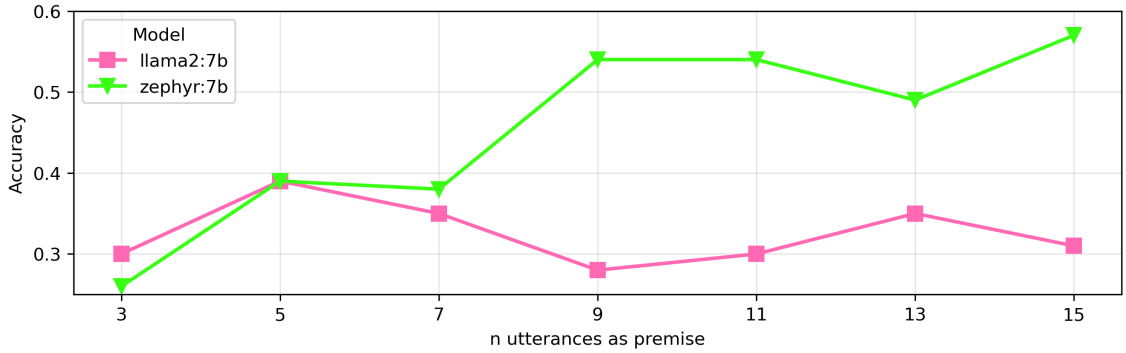
5 Results

5.1 Baseline results

Figure 3 shows the replicated baseline results alongside the original DNLI paper results for Llama 2 and Zephyr models. Despite testing on fewer samples, the replicated results show the same general behavior for Llama 2 and Zephyr. Llama 2 remains near the majority class baseline of 33%, while Zephyr rises from around 30% to 60% as utterance count increases. This confirms that the experimental setup and data is reliable.



(a) Original baseline results from DNLI paper [ENC+24] where the mean accuracy of Llama 2 and Zephyr is shown over three runs on the whole training set.



(b) Replicated baseline results with mean accuracy across two runs of 50 samples for each utterance premise stratum.

Figure 3: Comparison of original baseline results from the DNLI paper [ENC+24] with the replicated baseline results. Despite the replicated results testing on less data (50 samples per strata), the 2 plots show the same general behavior for the Llama 2 and Zephyr model.

Figure 4 shows the same baseline results, but now including the 5 SotA models. All 5 SotA models plateau around the 50–65% range. Notably, the SotA models don't exhibit the same clear upward trend that Zephyr displays, suggesting that the modest context length benefit observed in the DNLI paper does not generalize strongly to newer models. Furthermore, this more stable accuracy response is the same across different families, parameter sizes, and architecture types.

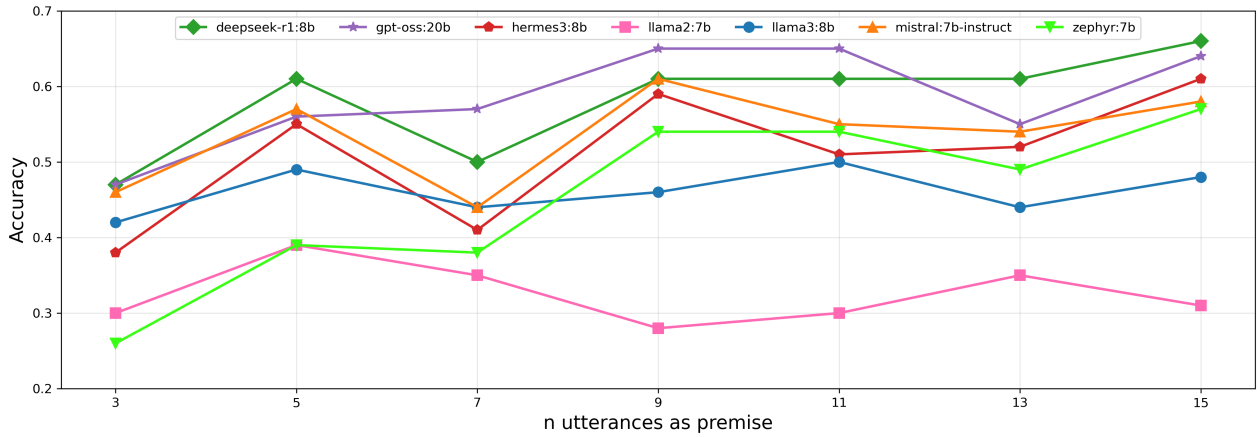


Figure 4: Baseline results including 5 SotA models showing mean accuracy across two runs of 50 samples for each utterance premise stratum.

5.2 Unique Speaker Count Results

Figure 5 shows accuracy across unique speaker count strata with premise length controlled between 280–350 characters. The results for all 7 models show no consistent downward trend as speaker count increases. Despite the context length being held constant, adding more speakers clearly doesn't seem to have an effect on the performance of the models. This evidence paired with the fact that the models can't seem to exceed an accuracy of $\sim 60\%$ suggests that models are not sensitive to the number of unique speakers and are most likely using other features of the premise and hypothesis to make predictions.

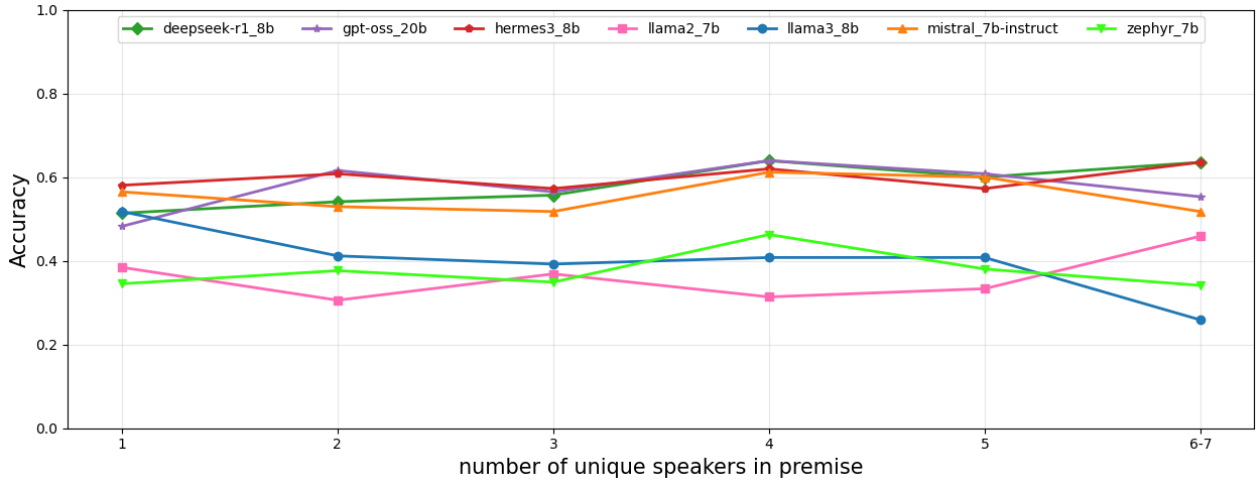


Figure 5: Mean accuracy across 3 runs, where every run tested an LLM on 85 samples from each unique speaker count stratum. Zephyr performs significantly worse compared to the baseline results in Figure 3b mainly due to timeout errors which were set at 20 seconds. Overall, across parameter size and model types there seems to be a limit at $\sim 60\%$ accuracy, while the amount of unique speakers in the premise has no substantial effect on the accuracy. This result indicates that LLMs don't track unique speakers.

5.3 Incremental Utterance Results

Figure 6 shows mean accuracy across all entries at each utterance relative to the first appearance of the last speaker. Offset 0 indicates the point at which the last speaker enters the dialogue for the first time. However, when looking at Figure 6 there is no clear jump in accuracy at the onset of the last speaker. Instead the results show a gentle upward drift in the later utterances, which is more consistent with the general utterance count benefit seen in the baseline. This suggests models are not updating their predictions in response to the last speaker's contributions.

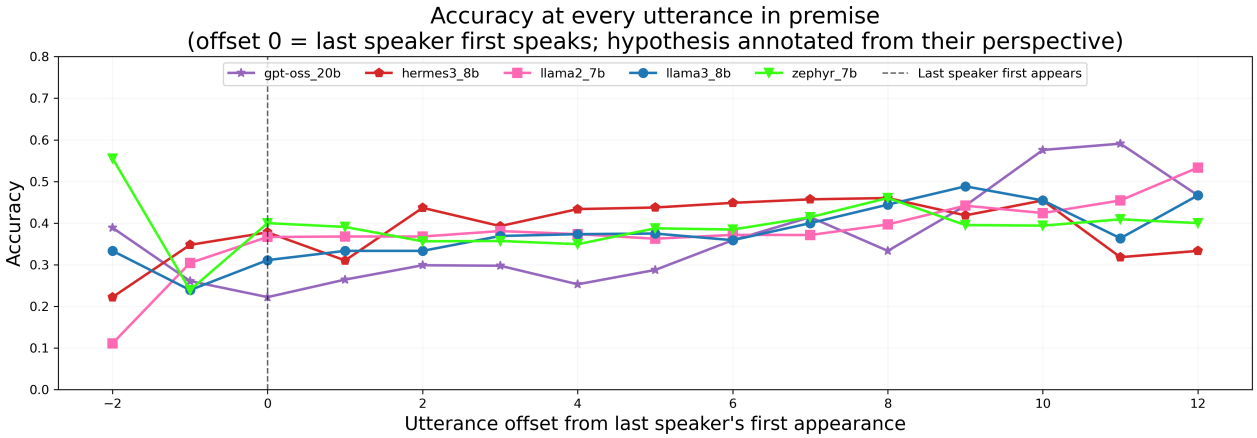
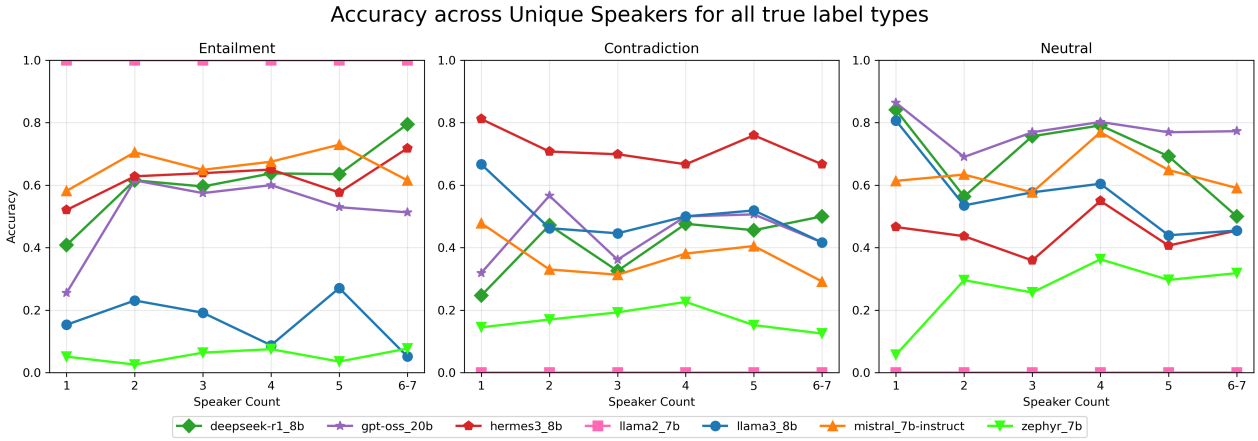


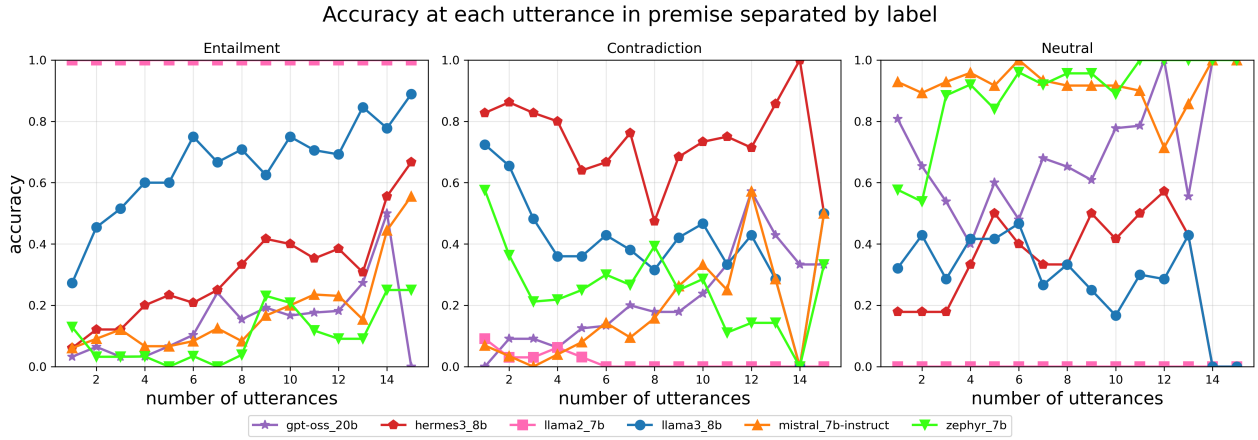
Figure 6: Mean accuracy across all entries at each utterance offset relative to when the last speaker first appears in that entry's premise. For each prediction in an entry, the offset is computed as the difference between the utterance where the last speaker first appears and the current utterance. This creates a normalized comparison across varied dialogue lengths. There is a gentle upward trend across all offsets.

5.4 Further Results Analysis

These results raise an important question: *why is there a consistent limit at* $\sim 60\%$ *across different features?* In order to help answer this question the results are plotted differently by splitting them per label. Figures 7a and 7b show both experiments split into subplots for Entailment, Contradiction, and Neutral.



(a) Mean accuracy across 3 runs of 85 samples for each unique speaker stratum. Mistral, deepseek, gpt-oss, and hermes have a very similar response in the Entailment subplot. Hermes and deepseek show the most stable response across all plots. Gpt-oss shows a slight bias towards neutral.



(b) Mean accuracy across 3 runs of 30 samples each, where in each sample the premise is slowly revealed to an LLM utterance by utterance. Mistral and zephyr show heavy bias towards Neutral, while Llama 2 essentially acts as a majority baseline as it shows 100% accuracy on Entailment. Overall hermes 3 looks to have the most stable response.

Figure 7: Shows both Figure 5 and Figure 6 split into 3 subplots for each label type: Entailment, Contradiction, and Neutral.

It is now clear that Llama 2 is essentially acting as a majority class baseline, consistently choosing Entailment regardless of the premise. Overall, the per-label split reveals that several models have collapsed to a label prior to performing inference. Zephyr and Mistral in Figure 7b show a heavy bias toward Neutral. Where this bias is stemming from is unclear and likely requires a reconstruction of the experimental approach among other considerations. However, it is likely some models are defaulting to a certain label due to artifacts and poor annotation in the dataset. Furthermore, a

model that reasons about speaker beliefs would show more stable performance across all three labels. Hermes 3 comes closest to this, showing one of the more balanced responses despite a slight bias toward contradiction.

6 Discussion

The main research question guiding this work was: *To what extent does the complexity of a multi-speaker dialogue influence the inference of Large Language Models?* Based on these results the simple answer is that the complexity of multi-speaker dialogue seems to show no influence on the inference of LLMs. However, this doesn't mean that LLMs find it trivial to track multiple speaker perspectives. Rather, these results indicate that these LLMs are so poor at the DNLI tasks that they aren't capable of being tested on higher level skills like tracking speaker perspectives. It appears likely that under the current prompting strategy LLMs are not building speaker belief representations in order to track an increasing number of speakers in the premise.

6.1 The ∼60% Ceiling and Annotation Artifacts

Across all experiments, SotA models plateau at $\sim$ 60% accuracy regardless of utterance count, speaker count, or model family. This ceiling closely mirrors the 58.9% accuracy achieved by the hypothesis-only classifier reported in the DNLI paper [ENC+24]. The near-identical figures might suggest that these LLMs are exploiting hypothesis-level artifacts, such as negation words, vague phrasing, or lexical overlap, rather than engaging in genuine dialogue understanding [GSL+18, PNH+18, MPL19].

6.2 Baseline Shows Modest Performance

The baseline experiment demonstrated that the SotA LLMs across family types and up to 20b parameters achieved approximately 50–65% accuracy on the DNLI task. While this shows a genuine improvement above the 33% majority class baseline, the overall improvement is modest compared to Zephyr's original results seen in Figure 3a. Moreover, despite all 5 SotA models having differences across architecture, parameter count, and training approach, a common plateau seems to be appearing at 60–65% accuracy. Rather than a limitation specific to any single model class, this suggests a fundamental challenge or barrier that the DNLI task poses.
The original DNLI paper [ENC+24] observed a steady upward trend from Zephyr as utterance count increased. The authors interpreted this as potential evidence of context-length benefit. The replicated baseline results confirm this trend for Zephyr, but show that newer SotA models do not exhibit the same clear upward trajectory. Rather, despite their performance still gradually rising as utterance count increases, relative to Zephyr this trend is much less pronounced. This suggests that SotA LLMs require less context when attempting to perform inference. In conclusion, this indicates that newer models may have a different way of using contextual information, as opposed to Zephyr.

6.3 Evidence against Speaker Perspective Tracking

The unique speaker count experiment directly addressed the sub-question: *How does LLM accuracy on the DNLI dataset change as the number of speakers increases from 2 to 7?* It is important to

note that for this test the unique speaker strata were filtered to have the same average context length, which eliminates the effect context size has on the results. Therefore, with the unique speaker strata made independent from context length, we expect to see accuracy decline as the amount of unique speakers increases. However, the results show a remarkably flat response across all speaker counts. Furthermore, if models were genuinely tracking speaker perspective it is likely that their baseline accuracy on the DNLI task would score higher than $\sim 60\%$ accuracy. Therefore, the absence of any kind of reaction to increasing speaker counts and an inability to exceed $\sim 60\%$ accuracy suggests that under the current prompting strategy LLMs are not developing models of speaker perspectives. However, changing the prompt to instruct the models to explicitly or implicitly track speaker perspectives as can be seen in Figure 8 might lead to better results.

> You are a strict NLI classifier. Output exactly one word: Entailment, Neutral, or Contradiction. Do not output any explanation. Given a dialogue excerpt and a Hypothesis, decide on the semantic relation between them. Track speaker perspectives in order to be successful at the task.
>
> You are a strict NLI classifier. Output exactly one word: Entailment, Neutral, or Contradiction. Do not output any explanation. Given a dialogue excerpt and a Hypothesis, decide on the semantic relation between them. Note the Hypothesis has been annotated from the perspective of the last speaker in the premise.

Figure 8: Examples of two different potential prompts that could initiate speaker perspective tracking from the LLMs.

Moreover, instructing the models to be strict NLI classifiers discourages using typical world knowledge and may restrict the models to perform correct inferences on the DNLI task. Therefore, using "`You are an NLI classifier`" may encourage the model to use more typical world knowledge, which should help with inference.
This conclusion is similar to prior work showing that LLMs have documented biases and shortcuts in NLI tasks [PNH+18]. Rather than engaging in pragmatic reasoning about speaker beliefs and intentions, models may rely on surface level features and annotation artifacts potentially learned from data they were trained on. Moreover, the 20b parameter model gpt-oss also doesn't appear to enable this pragmatic reasoning of speaker beliefs, despite the increase in parameter count. That being said, it is potentially possible as shown by Madaan et al. [MES+25] that the parameter count must simply be much larger in order for these higher level tasks, like tracking speaker perspective, to be possible.

6.4 No Evidence of Adaptive Belief Updating

The incremental utterance test was addressing the sub-question: *At what point in a conversation does an LLM typically find enough information to make a correct inference?* If models were tracking speaker perspectives, there is an expectation for a noticeable increase in accuracy at or shortly after the last speaker's first utterance in a premise. This is because the hypothesis for the premise is written from the last speaker's perspective, so most dialogues will become more clear once the last speaker becomes a part of the premise.
However, despite this expectation the incremental utterance results don't show a substantial increase

in accuracy directly after or shortly after the last speaker enters the context. Instead, accuracy slowly drifts upwards as more utterances are revealed, which is more consistent with the slight context-benefit seen in the baseline results. The absence of a noticeable jump in accuracy might indicate that the models are not selectively attending to information about specific speakers or updating their internal models of speaker beliefs as the dialogue unfolds.
Furthermore, the results reveal interesting model-specific biases when split by label. Llama 2 essentially acts as a majority class baseline because it nearly always predicts entailment. Zephyr and Mistral show strong biases toward the neutral label. Hermes 3 exhibits relatively stable performance across labels, suggesting less extreme biases. These label-specific patterns indicate that models are not making nuanced decisions when predicting the labels, but rather show a systemic preference towards one label. This may mean that models might often resort to bad guessing by resorting to a biased label preference. A model that tracks speaker beliefs and intentions will have to resort to fewer such guesses.

6.5 Takeaway

Overall there are two main takeaways from these results. Firstly, there is a general lack of performance of LLMs on the DNLI task as seen by the baseline results. It seems unlikely that increasing the parameter size will improve this performance, mainly due to the much larger gpt-oss with 20 billion parameters performing exactly the same as the models with 7 billion parameters. Secondly, there is an indication that LLMs are not tracking speaker perspectives. The uniform performance across different speaker counts and the absence of larger jumps in accuracy as hypothesis-relevant speakers become involved in the dialogue both indicate that LLMs don't appear to develop structured representations of speaker perspectives. Furthermore, the label separation of results exposed how some models default to a specific label when they don't know the answer, which indicates that the models are generally struggling with this task and could be lacking the proper representations needed to understand the dialogue.
These takeaways seem to extend Qamar et al. [QTH25], who revealed that LLMs fail to recognize speaker roles and the purpose of utterances. These takeaways show that not only do models struggle with speaker roles, but they also appear fundamentally limited in their ability to deal with information expressed across multiple turns of dialogue in the way required for natural dialogue understanding. Furthermore, despite impressive performance on many NLP benchmarks, LLMs may not have genuinely "solved" natural language understanding, as expressed by Madaan et al. [MES+25] as well. Current transformer-based models may be limited in their capacity to maintain and reason about structured, multi-agent belief states.

6.6 Limitations & Future Work

As explained in the Related Work section, NLI datasets commonly contain annotation artifacts. The DNLI dataset is no exception to this phenomenon. In fact, the $\sim 60\%$ accuracy ceiling observed across all SotA models closely mirrors the accuracy achievable with a BERT hypothesis-only classifier shown in the DNLI paper [ENC+24]. Therefore, this suggests that instead of understanding dialogue the LLMs are instead using clues like negation words, generic phrasing, or lexical overlap with dialogue content in order to predict labels. Testing these LLMs for artifact exploitation is an important task for future research. There are a few ways to do so. Firstly, run the same baseline

test on the LLMs, but use an hypothesis-only prompting method. If the LLMs score similarly to the baseline test with the premise this would suggest use of hypothesis artifacts. Secondly, shuffling the order of utterances in the premise or randomly reassigning speaker labels can test to see if the LLMs are actually tracking turn sequencing or tracking speaker perspectives. If the results are similar it would provide extra evidence that the LLMs are not tracking speaker perspectives and turn sequencing. Finally, using the pointwise mutual information calculation described in [GSL+18] on the DNLI data itself would reveal which words are overrepresented for each label type. Allowing us to know which words are highly correlated to certain labels. These are some of the few ways one could improve upon the results shown in this paper.

When manually scanning the data for irregularities, some of the premise hypothesis pairs were found to be badly annotated. The DNLI paper [ENC+24] noted that the Amazon Mechanical Turk workers struggled with the annotation task, thus some of those bad annotations might have slipped through to the final dataset. For example, Figure 9 shows one of multiple bad annotations in the dataset. Future research should re-assess the dataset for bad annotations and errors.

SPEAKER A:	like I mean he wrote like
SPEAKER B:	yeah
SPEAKER A:	does n't really seem to be any maths
SPEAKER B:	although they say er to be
SPEAKER A:	yeah
SPEAKER A:	yeah but Einstein was scientific and he was like imagination is more
SPEAKER B:	imagination is erm
SPEAKER A:	yeah more so and he's like if you want your
SPEAKER B:	see all those fairy tales I used to make up for you
SPEAKER A:	well
HYPOTHESIS:	Person A like the way she looks
RELATION:	Entailment

Figure 9: Example of poor DNLI data entry. This is clearly Neutral not Entailment.

The way in which the LLMs were prompted most likely has a much larger impact on results than initially expected. For example, due to the annotations for the DNLI task [ENC+24] being written from the perspective of the last speaker, informing the models that the hypothesis is written from the perspective of the last speaker could lead to active tracking of speaker intentions, beliefs, and perspectives. Therefore, future research could observe how models perform under different prompting strategies.

Furthermore, it is also possible that the LLMs didn't have sufficient parameter size to deal with the difficulty of the task. Madaan et al. [MES+25] shows that across almost all modern NLI datasets increasing the amount of parameters from `8B` to `70B` to `405B` significantly increases the performance of LLMs. Therefore, future work can test on LLMs with larger parameter size.

Overall, discovering how complexity of multi-speaker dialogue effects LLMs will only be possible after the aforementioned limitations are dealt with.

7 Conclusion

This paper attempted to investigate to what capacity the complexity of multi-speaker dialogue influenced the inference of LLMs on the DNLI task. The results revealed that increasing the complexity of dialogue, such as increasing the amount of unique speakers, had no effect on the inference accuracy of the models. SotA models plateau at roughly $\sim 60\%$ accuracy regardless of utterance count, number of unique speakers, or model family. This ceiling closely mirrors the 58.9% accuracy reported for a hypothesis-only classifier in the original DNLI paper [ENC+24], strongly suggesting that models are exploiting annotation artifacts rather than engaging in genuine dialogue understanding.

The unique speaker count experiment is particularly telling. Even with context length held constant, accuracy remained flat as the number of speakers increased from 2 to 7. If LLMs were building structured representations of speaker perspectives and belief states, increasing the number of speakers should impose a greater cognitive load and produce a measurable decline in at least some of the models. The absence of any such effect indicates that models are not tracking speaker perspectives at all under the current prompting strategy.

In short, genuine multi-speaker dialogue inference remains an unsolved problem for LLMs. Progress on this front will likely require a cleaner and artifact-controlled DNLI dataset, as well as testing different prompting strategies that encourage speaker perspective tracking from the LLMs.

References

[ENC+24] Adam Ek, Bill Noble, Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, Eleni Gregoromichelaki, Christine Howes, Staffan Larsson, Vladislav Maraev, Gregory J. Mills, and Gijs Wijnholds. I hea- umm think that's what they say: A Dataset of Inferences from Natural Language Dialogues. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, 2024.

[GSL+18] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data, 2018.

[LDH+17] Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. The spoken bnc2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3):319–344, 2017.

[Mac00] Brian MacWhinney. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, 3rd edition, 2000.

[MES+25] Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. Lost in inference: Rediscovering the role of natural language inference for large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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 9229–9242, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[MPL19] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[PNH+18] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference, 2018.

[QTH25] Ayesha Qamar, Jonathan Tong, and Ruihong Huang. Do LLMs understand dialogues? a case study on dialogue acts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26219–26237, Vienna, Austria, July 2025. Association for Computational Linguistics.