Human vs LLM in Conversational Repair Annotation: A New Resource And Comparative Study

Anh Ngo[1,5], Nicolas Rollet[1,2], Catherine Pelachaud[4], Chloé Clavel[1,3]
[1]ALMAnaCH, INRIA Paris
[2]Télécom Paris, SES, Institut Polytechnique de Paris, I3-CNRS
[3]Télécom Paris, LTCI, Institut Polytechnique de Paris
[4]CNRS, ISIR, Sorbonne University
[5]ISIR, Sorbonne University
{anh.ngo-ha, nicolas.rollet, chloe.clavel}@inria.fr
catherine.pelachaud@upmc.fr

Abstract

Addressing the scarcity of annotated data for Other-Initiated Repair (OIR), when recipients interrupt conversation progressivity to signal trouble, prompting speakers to provide repair, this work introduces OIR annotations for the NOXI corpus, achieving considerable reliability. We evaluate whether LLMs can reliably annotate OIR sequences using structured Chain-of-Thought prompting and conduct comparative analysis across two corpora: NOXI (natural dialogue) and CABB-S (Dutch, task-oriented), finding weak alignment between LLMs and human annotations, particularly in recognizing trouble signaling. Analyzing human-LLM disagreement using the LLM-generated explanations revealed limitations: models rely on lexical patterns rather than conversational context, construct reasonable-sounding but misleading narratives, highlighting crucial limitations for both automated annotation of complex interactional phenomena.

Keywords: other-initiated repaired, large language model, data annotation, llm reasoning, human-llm disagreement

1. Introduction

Conversation is the fundamental mode of human interaction, in which interlocutors can consider anything as a problem, potentially leading to communication breakdowns. When speakers encounter difficulties in understanding or interpreting their interlocutor's utterances, humans use repair mechanisms to resolve the issues and maintain conversational flow. **Other-Initiated Self-repaired, or in short, Other-Initiated Repair (OIR), is one of such mechanisms, where a recipient interrupts the ongoing conversational activity to signal an issue with the prior speaker's turn, prompting the original speaker to provide the actual repair** (Schegloff et al., 1977; Schegloff, 2000), as illustrated in Figure 1.

The study of OIR sequences has broad implications. Accurate identification of OIR sequences is crucial for dialogue systems to handle communication breakdowns effectively (Ashktorab et al., 2019), as poor handling of conversational errors, such as unexpected responses or ignoring from Conversational Agents can quickly demotivate users (Benner et al., 2021) and easily escalate into a conversational breakdown and interaction failure. Dingemanse and Enfield (2024) shows that current agents are typically fine-tuned to maintain an appearance of confidence, even when misunderstanding users. They attempt simple forms of repair only when dialogue stalls, usually through a simple apology or generic clarification request (e.g., *"Could you repeat that?"*), placing the burden of repair on the human user.

Implementing repair strategies that go beyond simple open requests (e.g., *"Huh? Could you repeat that?"*) and identifying human repair requests remain difficult due to the combinatorial complexity of natural conversation, where any element of a prior turn can become a potential source of trouble (Dingemanse and Enfield, 2024). Another key challenge is the **lack of annotated data for computational modeling of OIR**, although such data are essential for training and evaluation. Existing resources are largely derived from Conversational Analysis (CA) studies, where expert analysts perform in-depth, case-by-case analyses rather than standardized annotation. While these collections provide valuable theoretical insights, they are not directly applicable to computational work. Very few attempts of pre-annotated OIR corpus (e.g., CABB-S, Rasenberg et al. (2022)) exist for computational use, underscoring the critical need for additional resources. In contrast, corpus annotation involves trained, often non-expert annotators applying predefined guidelines to label data systematically across larger scales, which is a labor-intensive process that demands substantial training. This makes manual annotation costly and limited in scope, highlighting the need to explore automatic annotation methods, which remain underexplored for such complex phenomena.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6880–6892
11–16 May 2026.

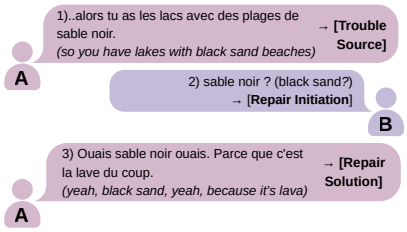

Figure 1: Example of Other-initiated Repaired (OIR) from NOXI corpus (English translated). Speaker B initiates repair request by *"sable noir?"*, signalling a trouble with this part of speaker A's turn 1. Speaker A provides solution by confirmation and addition information for clarification at turn 3.

In parallel, Large Language Models (LLMs) offer promising possibilities for automatic annotation (Brown et al., 2025; Nasution and Onan, 2024; Calderon et al., 2025). They can generate large amounts of labeled data, potentially mitigating the bottleneck of manual annotation. Moreover, LLMs' ability to produce explanations (Ara et al., 2024; Zhang et al., 2024) offers additional potential to interpret and understand annotation decisions, which is crucial to assess annotation quality and reliability.

To address the scarcity of resources for computational modelling of Other-Initiated Repair (OIR), this study introduces new OIR annotations for the NOXI corpus (French) with corresponding guidelines. We further examine whether LLM can serve as a reliable annotator through a structured prompting approach grounded in CA literature. Our contributions are in 3-fold: (1) the OIR sequence annotation for a French corpus (NOXI, Cafaro et al. (2017)) with detailed guidelines and protocols; (2) a systematic evaluation of LLM-based OIR annotation using a decision-tree structure prompting approach, with comparative analysis against human annotation across two corpora: NOXI (French, natural dialogues) and CABB-S (Dutch, task-oriented conversations), enabling assessment across languages and dialogue types; (3) disagreement analysis methodology using LLM-generated explanations to interpret human and LLM perspectives and assess annotation reliability.

2. Related Works

OIR Resources. OIR has been investigated across many languages and modalities in CA studies (Kendrick, 2015; Dingemanse and Enfield, 2015; Floyd, 2015; Rossi, 2015), but existing corpora are designed for qualitative in-depth analysis rather than standardized annotation suited for computational models. Resources with explicitly annotated OIR sequences remain scarce, with only one available resource in Rasenberg et al. (2022)'s work (CABB-S). Moreover, the low frequency of OIR occurrences and the variability across corpora further complicate annotation quality. Regarding this, van Dijck et al. (2024) demonstrates that systematic multi-stage review protocols can substantially improve agreement by distinguishing annotation errors from fundamental interpretational differences. Besides, Baledent et al. (2022) introduced validity, agreement, and consensuality metrics to represent distinct aspects of annotation quality, showing that high inter-annotator agreement does not necessarily guarantee annotation accuracy. **To address these limitations, this work provides OIR annotations for a French corpus (NOXI) with annotation guidelines and multi-stage validation.**

LLM-as-Annotator. With advances in LLM, recent research suggests using LLMs as an annotator to reduce human effort in multiple annotation tasks in NLP. Nasution and Onan (2024) compared human and LLM annotations in complex linguistic tasks, found that human annotations consistently outperformed LLMs in intricate contexts that require contextual understanding and ambiguity resolution, which are characteristics central to OIR identification. Brown et al. (2025) reveals that task difficulty is the most influential factor in human-LLM agreement patterns, with demographic factors showing less systematic impact than expected, when using LLM to annotate several subjective tasks. He et al. (2024) demonstrates that structured prompting, particularly "explain-then-annotate" methods, can significantly improve LLM performance by incorporating explicit reasoning processes. Lu et al. (2025) found that LLMs struggle significantly with samples which human annotators show low agreement. Calderon et al. (2025) provides statistical testing frameworks for evaluating LLM annotation quality, demonstrating that LLMs can sometimes reliably replace human annotators, depending on task and dataset. **Building on these findings, our study examines whether structured prompting can enable LLMs to perform reliable OIR annotation in dialogues, across languages and conversational contexts.**

LLM-as-Explanator. LLM capabilities in generating explanations for annotation decisions present novel opportunities to understand disagreement patterns. Chen et al. (2025) showed that LLM-generated explanations can approximate human judgment distributions with minimal guidance, suggesting potential value for illuminating annotation boundary cases rather than simply replacing human judgment. However, Zhang et al. (2024) identified a critical limitation: LLMs consistently fail to recognize legitimate interpretational differences, typically selecting decisive responses even when hu-

man annotators are evenly split. Lu et al. (2025) further demonstrates that LLM confidence often exceeds accuracy, particularly in ambiguous cases where there is substantial human disagreement. Ara et al. (2024) further investigates how knowledge gaps between beginner and expert annotators affect the quality of data annotation and involves LLM to generate explanations and reasoning for the Reasoning interface to assist annotators by offering concise, AI-generated reasoning for decision making, resulting in the most common reasons for disagreements between beginner and expert annotators being confusing words and hidden context. **Inspiring by these works, we examine LLM-generated reasoning to analyze and interpret human-LLM disagreement in OIR annotation.**

3. Methodology

3.1. Data & OIR Annotation Framework

Annotation corpus NOXI. We adopted an existing multimodal corpus NOXI (Cafaro et al., 2017) for OIR annotation. The corpus contains around 7 hours of dyadic conversations in French. For each conversation, one speaker plays the role of an expert while the other speaker plays the role of a novice, featuring spontaneous dialogue with natural OIR sequences. The NOXI corpus was transcribed at word-level by both human annotator and by an Automatic Speech Recognition (ASR) system.

Corpus for comparison CABB-S. For comparative analysis of LLM annotations, we use a pre-annotated corpus from Rasenberg et al. (2022), which is a subset of the Dutch task-oriented multimodal corpus CABB Eijk et al. (2022), here referred to as CABB-S. The corpus has about 5 hours of recordings. In each dyad, participants complete an object-matching task for multiple trials, alternating between the roles of director and matcher. Although the CABB-S (Dutch, task-oriented) and NOXI (French, natural settings) corpora differ in language and settings, this diversity allows us to assess LLM performance and its consistency across languages and dialogue characteristics, leveraging LLM's demonstrated multilingual capabilities.

NOXI's OIR Annotation Framework The two corpora employ different segmentation approaches. CABB-S provides pre-defined turn constructional units (TCU) boundaries, whereas NOXI transcripts contain sentence-level punctuation ("." and "?") but no discourse segmentation. To ensure consistent annotation logic across corpora, we adopted a functional annotation unit called *"segment,"* whose boundaries are determined by the scope of OIR components rather than fixed structural units. NOXI originally provides punctuation marks, which were manually transcribed based on prosodic cues and syntactic completion, offering candidate boundaries for segmentation. Based on that, our annotator identified segments as follows: (1) identifying repair initiation segments through trouble-signaling features (question words, repetition, repair markers); (2) determining the corresponding trouble source retrospectively based on what the repair initiation targeted; (3) marking the repair solution as the acknowledgment and resolution. This approach aligns with the CA principles, where trouble sources are defined retroactively by what recipients treat as problematic (Schegloff et al., 1977; Schegloff, 2000), rather than by predetermined structural boundaries. In CABB-S, the same functional logic was applied: segments corresponded to one or more TCUs based on OIR component logic. This approach ensures that the annotations of both corpora are comparable despite differences in underlying transcription granularity.

We adapted the OIR coding scheme developed by Rasenberg et al. (2022) for CABB-S annotation, which was derived from Kendrick (2015). Based on the definition of OIR, we formulate the steps of identifying repair initiation and its type into a decision tree protocol (see Figure 2). Guided by this protocol, our human annotator labeled each OIR sequence with three components: trouble source, repair initiation, and repair solution segments. The repair initiation was further classified into open request, restricted request, or restricted offer. Additionally, based on our observation from the NOXI corpus, trouble sources were coded as mishearing, non-understanding, or other, and the specific problematic spans were also marked. These refinements ensure a more detailed granularity of OIR phenomena, facilitating further OIR analysis. The detailed coding schema is provided in the Appendix.

3.2. NOXI Multi-Stage Annotation Protocol

We employed a French native speaker, majoring in linguistics, as the primary annotator (hereafter referred to as the (human) annotator). Annotations were performed in ELAN [1] with predefined tiers for OIR components and three repair initiation types. To ensure annotation consistency and quality, we followed a multi-stage protocol, beginning with a training session with provided guidelines and examples (see Appendix), followed by a pilot annotation of 20 minutes of recordings. The next step involves a discussion and feedback with two reviewers. The first reviewer is the first author, a computational linguist (hereafter referred to as the mid-

[1] https://archive.mpi.nl/tla/elan

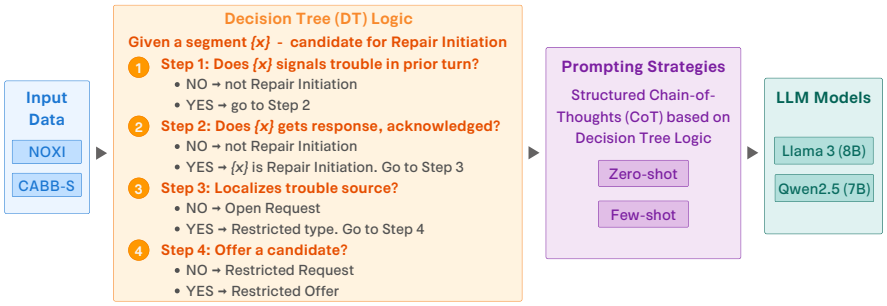

Figure 2: Overview of LLM Annotation Framework

dle reviewer), and the second reviewer is a senior conversational analyst (hereafter referred to as the expert). After the discussion, the annotator revised his decisions for three instances for the pilot study and started the main phase. The main annotation phase proceeded in mini-batches, each comprising two recordings (around 40–45 minutes), followed by iterative reviews after each batch. First, the middle reviewer provided feedback and discussed revisions with the annotator. Subsequently, the expert assessed the revised annotations and had a final discussion with both the annotator and the middle reviewer to reach a final consensus. This two-stage process enabled early detection of systematic issues, iterative refinement of guidelines, and assessment of annotation quality improvement by examining the evolution of the inter-annotator agreement (IAA), which we will discuss in greater detail in Section 4. The annotator re-annotated 10% of the data randomly selected from early batches during the final annotation stage, approximately 2.5 months later, to assess the annotator's self-consistencies.

3.3. LLM Annotation

Task: Identify OIR in Dialogues (Summarized)

Prompt

Sytem: Role: expert conversation analyst..(OIR background). **Task:** Identify Repair Initiation and classify type via decision steps:
(1) Signals trouble? (2) Gets acknowledgment? (3) Localizes trouble? (4) Offers candidate?
Types: Open Request, Restricted Request, Restricted Offer. **Input:** Dialogue segments: [id]. [speaker] [timestamp]: [text]

Example Output

```
{"dialogue_id": "cabb_ex_2", "oir_sequences": [{
"trouble_source": {"id":"1", "spk":"B",
"timestamps":"00:12-00:15", "text":"..."},
"repair_initiation": {"id":"2",..,
"type":"open_request"},
"repair_solution": {"id":"3", "spk":"B", ..},
"decision_steps": {"step1": {"value":true},
...}}]}
```

Figure 2 illustrates our framework for OIR annotation using LLMs. We selected two open-source models: Llama3-8B-Instruct (Grattafiori et al., 2024), for its strong instruction-following capabilities and broad benchmarking across NLP tasks; and Qwen2.5-7B-Instruct (Qwen et al., 2025), the latest release of the Qwen family, for its improved instruction adherence, enhanced role-based prompting, and multilingual proficiency. To ensure comparability with human annotations, our LLM prompting strategy was designed to mirror the four-step decision tree used in human annotation guidelines. We adapted the Chain-of-Thought (CoT) prompting approach (Wei et al., 2022) as its multistep reasoning aligns naturally with our decision logic for OIR identification, allowing the model to explicitly rationalize each classification stage. The CoT logic is embedded within the task prompt, where each step branches through binary (yes/no) decisions rather than the original linear reasoning chain (see the compacted version in Prompt 3.3, the complete prompt is provided in Appendix A). Rather than optimizing performance through advanced prompting techniques, we prioritize transparent reasoning and systematic comparability, facilitating the analysis of interpretive differences between human and LLM annotations. Therefore, we experimented with our structured CoT prompting with zero-shot and few-shot scenarios. Source code with full prompt templates, few-shot examples, and all experiments' parameters [2] is provided.

4. NOXI Human Annotation Results & Discussion

4.1. Overall annotation results

We identified 49 OIR sequences from 21 dyads in NOXI, occupying approximately 3% of the data, with a mean of 2.2 repair initiations per dyad. The distribution includes 22 (47.8%) open requests, 17 (37%) restricted requests, and 7 (15.2%) restricted offers. In comparison, Rasenberg et al. (2022) reports 378 OIR sequences from around 8 hours of recordings in the CABB-S corpus (~14.5% of the data), with 24 open requests, 39 restricted requests,

[2] https://github.com/haanh764/NOXI_annotation

and 315 restricted offers. Despite comparable total durations, OIRs occur less frequently in NOXI, and the distribution of OIR types differs noticeably. These differences appear to stem from both interactional design and epistemic stance characteristic of each corpus. CABB-S is task-based, with referential objects presented to both participants, creating epistemic symmetry and favoring restricted offers as a resource for negotiating mutual understanding. In contrast, NOXI is designed for natural, spontaneous interaction, making repair more likely in natural dialogue settings, resulting in less frequent OIR. Besides, the setting is epistemically asymmetric (expert–novice role play), where participants tend to rely more on open and restricted requests to elicit clarification. The screen-mediated setting further contributes to the prevalence of open requests, as issues such as lag or audio quality can lead to mishearing. In summary, it suggests that epistemic asymmetry in combination with the communicative setting (natural vs. task-based, screen-mediated vs. face-to-face) influences both repair frequency and type preference.

4.2. Annotation Evaluation

Metric	Stage 0→1	Stage 1→2
Agreement rate	72.73%	89.09%
Cohen's κ	0.372	0.667
Revision rate	27.27%	10.91%
Sequences changed	15	6

Table 1: NOXI annotation reliability metrics across 2 validation stages. Stage 0→1: human annotator vs. middle reviewer; Stage 1→2: stage 1 consensus vs. expert reviewer.

Table 1 summarizes the reliability metrics across the two-stage validation, showing progressive improvement. Cohen's κ was calculated for IAA between our annotator and the middle reviewer at stage 1, and between the stage 1 consensus and the expert at stage 2. The score increased markedly from 0.372 (fair agreement) to 0.638 (substantial agreement), indicating a significant improvement in annotation consistency. The agreement rate rose from 72.73% to 89.09%, while revision rates decreased from 29.63% to 12.96%, indicating convergence toward stable annotations. Overall, 61.82% of sequences remained unchanged through both review stages, showing a fair stability rate. For revised samples, stage 1 systematically resolved annotation uncertainties and guideline ambiguities, with most revisions addressing "undecided" annotations. The rest involved misclassified OIR types or misidentified OIR sequences. For stage 2, revised samples are mostly genuinely ambiguous boundary cases (see Appendix), where the annotator and middle reviewer had considered these as regular dialogue but were reclassified by the expert as repair initiations. The annotator shows high self-consistency. After 2.5 months, 10/11 OIR sequences (90.9%) were re-annotated identically, with only one reclassified as non-OIR, corresponding to a Jaccard similarity of 0.91 and 100% agreement on repair initiation types.

5. LLM Annotation Results (NOXI+CABB-S) & Discussion

Table 2 summarizes the overall results of LLM annotation on NOXI and CABB-S corpora, using four model configurations (LLaMA and Qwen, each in zero-shot and few-shot settings), evaluated at both sequence and segment (repair initiation) levels. The human annotations in NOXI correspond to the final consensus results after two-stage validation. In addition, Table 3 presents the LLM's performance in identifying repair initiation types.

Sequence-level annotation. At the sequence level, the human annotator identified 47 OIR sequences in NOXI and 366 in CABB-S, whereas the two models showed different patterns. LLaMA consistently overpredicts across both corpora and strategies, predicting approximately double the number of OIR sequences annotated by the human annotator in NOXI and around 4-5 times more in CABB-S. Conversely, Qwen exhibits extreme conservatism in both settings, identifying far fewer sequences (coverage only ~2–3% in zero-shot and 30–50% in few-shot) for both corpora. The number of fully matched sequences, where all three components (trouble source, repair initiation, and repair solution segments) align with human annotation, remains low, typically below 10 per each setting. In NOXI, models achieve 8.5-12.8% exact sequence match rates (4-6 sequences out of 47), with LLaMA-FS and Qwen-FS tied at 12.8%. Meanwhile, CABB shows higher absolute numbers of exact matches for LLaMA (77-79 sequences in both settings), though this represents only around 21% of the human annotations. Given the small number of matched sequences, similar metrics were not computed at this level, as they are more sensitive and may not reflect a meaningful difference.

Segment-level annotation. At the segment-level, we examined repair initiation segments identified by both the LLMs and the human annotator. LLM models' performance improves slightly, with higher numbers of LLM identified repair initiations aligned with human annotator, but remains weak overall. Precision and recall are low across corpora (≤ 0.50),

Corpus	Model	Sequence-Level			Segment-Level				
		Human	LLM	# Exact Matched	# RI Matched	P	R	F1	κ
NOXI	LLaMA-ZS	47	85	4	14	0.165	0.298	0.212	-0.618
	LLaMA-FS	47	106	6	18	0.170	0.383	0.235	-0.477
	Qwen-ZS	47	1	1	1	1	0.021	0.042	0.000
	Qwen-FS	47	26	6	12	0.462	0.255	0.329	-0.488
CABB-S	LLaMA-ZS	366	1823	77	139	0.076	0.380	0.127	-0.242
	LLaMA-FS	366	1530	79	153	0.100	0.418	0.161	-0.268
	Qwen-ZS	366	10	0	6	0.60	0.016	0.032	-0.022
	Qwen-FS	366	114	18	37	0.327	0.101	0.154	-0.388

Table 2: Overall LLM annotations results, with 2 strategies: ZS=zero-shot, FS=few-shot. **Sequence-Level:** Counts of OIR sequences identified by LLM, and *# Exact matched*=numbers of exactly matched sequences between LLM and human annotator (matched all 3 components trouble source, repair initiation, and repair solution). **Segment-Level:** LLM performance on identification of whether each dialogue segment is a repair initiation (*# RI matched*=numbers of matched repair initiation segment between LLM vs. human annotator, *P*=precision, *R*=recall, *F1*=F1-score, κ=Cohen's kappa for agreement between LLM vs. human annotator).

yielding F1 scores below 0.3 for most configurations. Qwen-FS achieves the highest F1 score and precision on both NOXI and CABB-S, but lower recall, while LLaMA-FS shows a reverse trend, higher recall but lower precision, reflecting its over-inclusiveness predictions, where models correctly identify some repair initiations but flag many non-repair segments as positive. Qwen-FS thus identifies only a few but salient repair initiations. Nevertheless, κ values again hover around zero or negative, confirming limited alignment with human labels in both corpora. This weak alignment shows differences between how LLMs and human annotators identify repair. While human annotation relies on conversational progressivity by analyzing a turn's relation to surrounding turns, LLMs appear to rely largely on surface-level lexical cues to detect trouble signalling, as further confirmed in Section 6. Moreover, the extreme difference in behaviour between LLaMA's over-prediction and Qwen's under-prediction indicates that neither model captures the OIR patterns in conversation.

In addition, at the sequence-level, a critical insight from the results is the ratio of matched repair initiations (same repair initiation segment identified by both human and LLM) to exact sequence matches (all three components align), suggesting clues for the source of disagreement between the LLM and human annotator. Since a repair initiation is theoretically defined by its relationship to a trouble source and repair solution, matching on the repair initiation segment alone indicates that both LLM and the human annotator recognized a repair phenomenon but disagreed on its boundaries. In NOXI, this ratio varies sharply: LLaMA models achieve around 29–33% (4/14 and 6/18), while it is around 50% in Qwen-FS (6/12), implying severe boundary disagreement in NOXI (around 67-71% for LLaMA and around 50% for Qwen-FS). Noticeably, Qwen-ZS achieving 1/1 (100%) does not mean it is better at boundaries, as it identified only one OIR sequence in total, which happened to be exactly aligned with humans' for both sequence and repair initiation segment-level. CABB shows better boundary agreement (approx. 50%) except for Qwen-ZS, suggesting either the OIR sequences in CABB-S have clearer boundaries or the LLM models happen to align better with human judgment in CABB-S.

Repair initiation types classification. Among matched repair initiations, Table 3 reports the numbers of each repair initiation type identified by both LLM and human annotator. Across corpora, open request emerges as the most consistently identifiable type, whereas restricted offer is the most difficult type overall. For open request type, in NOXI, both LLaMA-FS and Qwen-FS achieve 7 matched instances out of 23 human annotations (30.4%). However, their classification patterns are different, in which LLaMA-FS labeled a total of 16 open request instances, while Qwen-FS identified only 14, resulting in a precision of around 44% and 50% for each, respectively. This pattern becomes more visible in CABB-S, where both LLaMA experiments achieve more than 50% (12/24 and 15/24) matched open request instances with human labels compared to only 12.5% (3/24) of Qwen-FS. However, examining total predictions tells a different story: LLaMA-FS made 1347 open request predictions to achieve these 15 correct identifications (1.1% precision), while Qwen-FS made only 63 predictions (4.8% precision, nearly 4 times better). Restricted request shows intermediate difficulty for both models and across corpora. In NOXI, LLaMA-ZS achieved 4 matches out of 17 instances identified by the human annotator, followed by LLaMA-FS and Qwen-FS with 2 matches only.

In CABB-S, the similar pattern between precision and recall persists, in which LLaMA-ZS identifies 8

Data	Model	Types (# matched/# identified) OR	RR	RO	Agree. κ
NOXI	LLaMA-ZS	3/7	4/7	0/0	0.125
	LLaMA-FS	7/16	2/2	0/0	0.165
	Qwen-ZS	0/0	0/0	0/1	0.000
	Qwen-FS	7/14	2/8	0/4	0.438
	Human	23	17	7	-
CABB-S	LLaMA-ZS	12/981	8/678	23/164	0.113
	LLaMA-FS	15/1347	3/182	0/1	0.014
	Qwen-ZS	2/6	0/1	2/3	0.333
	Qwen-FS	3/63	4/49	0/2	-0.060
	Human	24	39	315	-

Table 3: LLM annotation performance per repair initiation types, among matched repair initiation indentified by LLM and human annotator (per each type, *# matched*: numbers of repair initiations of that type by both LLM and human, *# identified*: total numbers of repair initiation of that type identified by LLM including unmatched sequences)

correct instances from 678 predictions (1.2% precision) versus Qwen-FS's 4 correct from 49 predictions (8.2% precision). Restricted offer presents the most challenging across nearly all configurations, especially in the NOXI corpus, where this type is rare even in human annotation; no model correctly classifies any of the 7 restricted offers. In CABB-S, even the restricted offers dominate (315/378 instances, 83.3%), performance remains low. LLaMA-ZS achieves the best absolute count (23 correct) but only 7.3% (23/315) recall and 14.0% precision (23/164), followed by Qwen-ZS with 2 matches out of its total 3 predictions only, and when it comes to few-shot prompting, both models perform even worse. This noticeable performance degradation from NOXI to CABB also reflects the effects of class distribution. Regarding the agreement between LLMs and human annotation, Qwen-FS on NOXI attains the highest overall Cohen's κ (0.438), showing limited but fair agreement with human annotation, while Cohen's κ values on CABB-S are varied around zero or negative, except that Qwen-ZS surprisingly achieves moderate agreement (κ=0.333) but with minimal predictions.

Taken together, these results reveal that few-shot strategies outperform zero-shot ones, particularly on NOXI, indicating that a few examples help the models better perceive the OIR concept. Regarding the model family, LLaMA identified more potential OIRs overall but is less selective, while Qwen found fewer OIRs, but when they do, they are more likely to be correct. Although overall Cohen's κ is still far from inter-human reliability, both repair initiation detection and type classification show slightly better performance in NOXI than in CABB-S, likely due to the greater contextual complexity of task-oriented dialogues, where repair initiation is highly context-dependent and embedded in directive or referential actions. It also suggests that the LLM annotation is sensitive to discourse organization, especially for restricted offer instances, which rely on implicit candidate understandings rather than explicit interrogative markers. Despite having a decision-tree structure to guide the model through explicit reasoning steps, Step 1 (*'Does x signal trouble?'*) fundamentally requires understanding how a turn relates to its surrounding turns rather than just surface form. Although the proposed structured prompts can mimic the reasoning procedure, they cannot provide the interactional knowledge needed for this evaluation. Especially in restricted offers, the surface forms can appear as a demonstration, and the model must infer that a speaker proposes a candidate understanding of what another speaker said, which requires tracking both speakers' epistemic states across turns. This process of reasoning with theory-of-mind over conversational context cannot be induced by prompt structure alone.

6. Disagreement Analysis

6.1. Overview of Disagreement Cases

To understand the systematic differences between human and LLM annotations, we extracted and analyzed disagreement cases between final human annotations and Qwen-FS annotations on NOXI, resulting in 41 instances for analysis. Figure 3 presents the overall distribution of disagreement types: 75.6% of disagreements occurred when the human annotator identified repair sequences that the LLM did not. For most of them ($>$ 77%), LLM found nothing, while a few showed partial (6.5%) or completely different (16%) overlaps. The other disagreements fell into groups of LLM found but human not (approx. 17%), and the same repair initiation found but classified as different repair types (7.3%). Across decision steps, LLM and human annotation diverge the most at the first step in recognizing trouble-signaling in the target segment, followed by identifying repair solutions, and classifying repair initiation types is not a controversial problem. These patterns raise a critical question about what the main reasons or cues lead to the disagreement in recognizing trouble-signaling in the first place.

To answer this, we examined the LLM's capabilities and reliability in explaining both humans' and the LLM's reasoning. Inspired by recent works in LLM-as-a-judge, we prompted Qwen2.5-7B to generate structured explanations for disagreements between two given OIR annotations for the same dialogue, identifying the likely reasoning behind each annotator's decisions and assessing which interpretation better aligns with OIR theory. (see the summarized version of our Prompt 6.1; the full

Case	Extract	LLM explanation	Key Issue
Human identified, LLM not	[1]N: "Exactly, yes I'm actually planning to leave" [2]E: "sorry" [3]N: "..I was saying I'm actually planning to leave very soon, at the end of April."	LLM's reason: "'Pardon' does not contain any cues suggesting trouble or repair initiation." Human's reason: "...'Pardon' indicating trouble with the previous statement, 'j' envisage ..peu fin avril' provides the repair solution."	- Contingency/adequacy problem, LLM missed the link to common conversational orientation/actions - The "pardon" is due to mishearing, because of overlap speaker turns, need non-verbal cues to decide, which LLM missed
LLM identified, Human not	[1]E: "..The banks of the Arno, which are also very, very nice to do... [2]E: "It's very pretty.." [3]N: "okay, can you explain to me what the Arno is" [4]E: "Well, it's the river that flows through the middle of Rome"	LLM's reason: "The phrase 'd' accord' (okay) suggests agreement, but the following question indicates confusion or need for clarification" Human's reason: The segment appears to be a natural follow-up to the previous statement without any signs of confusion or need for clarification.	LLM explanation used words like "need", pretends to be able to infer human intention (theory-of-mind), yielding false positive

Table 4: Sample Cases of LLM Explanation Analysis

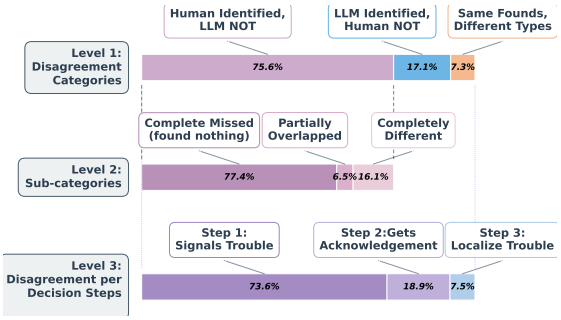

Figure 3: Disagreement cases overview statistics

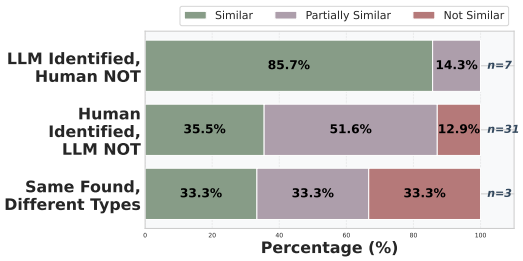

Figure 4: Overview of similarity between LLM's generated human reasoning vs the real human annotator's reasoning

version is given in the Appendix B). However, we focus on analyzing LLM's reasoning process to examine how the LLM conceptualizes OIR sequences compared to CA methods, rather than investigating its judgments as authoritative. Therefore, for each disagreement case, we provide an LLM with anonymous annotations (eg, annotator A's annotation and annotator B's annotation) to ensure the LLM does not always favor human annotations. The proportion of human and LLM annotations being either annotator A or B is randomly assigned.

Disagreement Analysis Prompt (Summarized)

Input: Dialogue + Annotator A's annotation + Annotator B's annotation
Task: Analyze disagreement and provide JSON output with:
- Disagreement summary; Possible reasons (explanation, supporting evidence)
- Each annotator analysis (annotation summary, rationale, supporting cues)
- Recommendation (favorability towards annotators)

6.2. LLM's Explanation Analysis & Discussion

The model generated six categories of disagreement reasons, with the most common being repair initiation interpretation (37%), type classification (34%), sequence boundary dispute (29%), and trouble source identification (21.9%). We validate the reliability of these reasons, finding correct reasons for most categories (93-100%), except for type classification and sequence boundary dispute, where approximately 50% of disagreements stemmed from different causes rather than those generated. When the LLM identified OIR that humans did not, its rationales closely resembled human reasoning (85.7% similar, 14.3% partially similar). In contrast, when humans identified OIRs missed by the LLM, the LLM struggled significantly to reconstruct human reasoning (35.5% similar, 51.6% partial, and 12.9% not similar). This asymmetry pattern can be explained by the nature of the reasoning task: when LLM identifies an OIR that humans do not, reconstructing human reasoning is relatively straightforward by applying the decision tree logic in reverse to deduce why that annotator rejected the sequence. In contrast, cases where humans identified an OIR but the LLM did not are more complex, as human annotators may rely on more subtle interactional cues, such as contextual factors, making it harder for the LLM to infer.

Table 4 presents examples of LLM-generated reasoning, illustrating the earlier observed patterns. In "LLM identified, Human not" cases, the LLM's reconstruction of human reasoning (e.g., *"a natural follow-up without confusion"*) correctly aligned with our annotator's reason, but can be easily inferred by reversing the decision-tree logic. For the opposite case, LLM could produce only partial reasoning with supporting lexical cues, but remained incomplete: the actual trigger for *"pardon"* was the speaker's mishearing due to overlapping speech, which requires inferring acoustic from temporal patterns rather than lexical content solely. When explaining its own annotations, we observed several issues. In sample 1, the LLM dismissed *"pardon"* as non-trouble signaling, reasoning solely based

on lexical or semantic cues rather than interactional context, reflecting a contingency problem. In sample 2, it combined *"confusion"* (a cognitive state) with *"clarification"* (a conversational action), treating them as interchangeable, and introduced intent interpretation (e.g., *"need"*), which goes beyond current observation. In other words, when generating explanations, LLM models tend to produce mentalistic narratives by attributing mental states such as *"confusion"* or intents such as *"need"* to speakers. These explanations may sound reasonable, which can make them appear convincing even when they are incorrect annotations. However, under the CA methodology, analysis should be grounded in observable interactional behavior rather than inferred speaker intentions. This reveals that LLM explanations may appear persuasive but misidentify regular requests for information as repair.

7. Conclusion

This work provides annotated Other-Initiated Repair sequences for the NOXI corpus with systematic reliability evaluation across two-stage validation, achieving substantial inter-annotator agreement. We also investigated LLM capabilities in annotating OIR sequences in task-oriented and natural dialogue, revealing fundamental limitations in automated repair annotation. While few-shot prompting with decision-tree structured outperformed zero-shot approaches, overall alignment with human annotations remained weak across both corpora. LLaMA models consistently over-predicted while Qwen models exhibited extreme conservatism, with both struggling particularly at sequence boundary identification and restricted offer classification, showing limitations in identifying implicit trouble-signaling rather than explicit markers.

Disagreement analysis revealed that LLMs primarily failed at recognizing trouble-signaling, with the majority of disagreements involving repairs that humans identified but LLMs missed. The models demonstrated limited reasoning capabilities, struggling to understand the nuanced cues underlying human repair identifications. LLM self-explanations exposed critical limitations that rely on lexical patterns rather than conversational context and progressivity, and produce plausible but misleading mentalistic narratives that misidentified routine information-seeking as repair. The plausibility of LLM explanations poses risks for users without expertise. This work provides empirical evidence of LLM limitations in capturing complex interactional phenomena.

Future work could explore several directions to improve LLM performance on OIR annotation. Retrieval-augmented prompting with examples from CA literature could help ground the model in interactional rather than lexical reasoning. In addition, multimodal LLMs, incorporating multimodal inputs, such as prosodic and visual signals, may help LLM resolve ambiguous cases.

8. Acknowledgements

We thank the anonymous reviewers for their constructive feedback. Data were provided (in part) by the Radboud University, Nijmegen, The Netherlands. This work has been supported by the Paris Île-de-France Région in the framework of DIM AI4IDF. It was also partially funded by the ANR-23-CE23-0033-01 SINNet project. Additional support was provided by the ANR under the France 2030 program PRAIRIE (ANR-23-IACL-0008).

9. Bibliographical References

Zinat Ara, Hossein Salemi, Sungsoo Ray Hong, Yasas Senarath, Steve Peterson, Amanda Lee Hughes, and Hemant Purohit. 2024. Closing the knowledge gap in designing data annotation interfaces for ai-powered disaster management analytic systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, page 405–418, New York, NY, USA. Association for Computing Machinery.

Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin. 2022. Validity, agreement, consensuality and annotated data quality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2940–2948, Marseille, France. European Language Resources Association.

Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. 2021. What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents.

Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Guangyuan Piao, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jurgen Müller, Lukas Gianinazzi, Ales Kubicek, Hubert Niewiadomski, Aidan O'Mahony, Onur Mutlu, and Torsten Hoefler. 2025. Demystifying chains, trees, and graphs of thoughts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.

Megan A. Brown, Shubham Atreja, Libby Hemphill, and Patrick Y. Wu. 2025. Evaluating how llm annotations represent diverse views on contentious topics. *ArXiv*, abs/2503.23243.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. 2025. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10777–10802, Vienna, Austria. Association for Computational Linguistics.

Mark Dingemanse and N. J. Enfield. 2015. Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguistics*, 1(1).

Mark Dingemanse and N.J. Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

Simeon Floyd. 2015. Other-initiated repair in cha'palaa. *Open Linguistics*, 1(1).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay

Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Kobin H. Kendrick. 2015. Other-initiated repair in english. *Open Linguistics*, 1:164–190.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Is LLM an overconfident judge? unveiling the capabilities of LLMs in detecting offensive language with annotation disagreement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5609–5626, Vienna, Austria. Association for Computational Linguistics.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062, Miami, Florida, USA. Association for Computational Linguistics.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Kseniia Petukhova and Ekaterina Kochmar. 2025. A fully automated pipeline for conversational discourse annotation: Tree scheme generation and labeling with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15829–15852, Vienna, Austria. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Marlou Rasenberg, Wim Pouw, Asli Özyürek, and Mark Dingemanse. 2022. The multimodal nature of communicative efficiency in social interaction. *Scientific Reports*, 12.

Giovanni Rossi. 2015. Other-initiated repair in italian. *Open Linguistics*, 1(1).

Emanuel A. Schegloff. 2000. When 'others' initiate repair. *Applied Linguistics*, 21:205–243.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361.

Gijs van Dijck, Carlos Aguilera, and Shashank Chakravarthy. 2024. Deciphering disagreement in the annotation of eu legislation. *Artificial Intelligence and Law*, pages 1–36.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Michael J.Q. Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2024. Diverging preferences: When do annotators disagree and do models know? *ArXiv*, abs/2410.14632.

10. Language Resource References

Cafaro, Angelo and Wagner, Johannes and Baur, Tobias and Dermouche, Soumia and Torres Torres, Mercedes and Pelachaud, Catherine and André, Elisabeth and Valstar, Michel. 2017. *The NoXi database: multimodal recordings of mediated novice-expert interactions*. Association for Computing Machinery, ICMI '17.

Lotte Eijk and Marlou Rasenberg and Flavia Arnese and Mark Blokpoel and Mark Dingemanse and Christian F. Doeller and Mirjam Ernestus and Judith Holler and Branka Milivojevic and Asli Özyürek and Wim Pouw and Iris van Rooij and Herbert Schriefers and Ivan Toni and James Trujillo and Sara Bögels. 2022. *The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses*. Academic Press Inc.

Marlou Rasenberg and Wim Pouw and Asli Özyürek and Mark Dingemanse. 2022. *The multimodal nature of communicative efficiency in social interaction*. Nature Research.

A. LLM OIR Annotation Prompt

Task: Identify Other-Initiated Repair (OIR) in Dialogues (Full Prompt)

System Prompt

You are an expert in conversation analysis specializing in identifying Other-Initiated Repair (OIR) sequences using a structured decision-tree approach. **Definition.** Other-Initiated Repair (OIR) occurs when a recipient interrupts the ongoing conversational activity to signal possible trouble in speaking, hearing, or understanding, leaving it to the speaker of the trouble source to perform the repair. **Key components:** *Trouble Source:* The segment in prior speaker's turn retrospectively signaled by another speaker as containing a repairable; *Repair Initiation:* The segment that signals trouble/reparable with the prior turn and initiates the repair sequence ; *Repair Solution:* The Trouble Source speaker's subsequent turn that acknowledges the Repair Initiation and resolves the repairable.

Task Instructions

Given a dialogue containing a list of segments, in format: `[segment_id].[speaker_name] [start_time-end_time]: [text]` Your task is to identify the Repair Initiation segment in an OIR sequence and classify its type. First, read through the entire dialogue to understand the context. Then, for each segment that might be a Repair Initiation, follow the decision tree below. For each decision step, explain your reasoning before concluding YES or NO:
Step 1. Does a candidate segment signal a trouble/reparable presence in the prior speaker's segment?
YES → Step 2 (identify prior segment as Trouble Source); **NO** → Not Repair Initiation.
Step 2. Is the signal acknowledged in the subsequent segment?
YES → Segment is Repair Initiation; subsequent segment is Repair Solution → Step 3; **NO** → Not Repair Initiation.
Step 3. Does the candidate segment localize the trouble source?
NO → `open_request`; **YES** → Step 4
Step 4. Does the candidate segment offer a candidate understanding?
NO → `restricted_request`; **YES** → `restricted_offer`

Input

```
{dialogue}
```

Required Response JSON Format

```
{"dialogue_id": "...", "language": "...",
"oir_sequences": [{"sequence_id": "...",
"trouble_source": {segment_id, speaker_name,
start_time, end_time, text},
"repair_initiation": {segment_id, speaker_name,
start_time, end_time, text, type},
"repair_solution": {segment_id, speaker_name,
start_time, end_time, text}, "confidence":
0.0-1.0, "decision_steps": {
"step1_signals_trouble": {value, explanation,
confidence},
"step2_gets_acknowledgment": {value, explanation,
confidence},
"step3_localizes_trouble": {value, explanation,
confidence},
"step4_offers_candidate": {value, explanation,
confidence}}}]}
```

B. LLM-Human Annotation Disagreement Analysis Prompt

Task: Explain Human-LLM Disagreement in OIR Annotation (Full Prompt)

System Prompt

You are an expert in conversation analysis and Other-Initiated Repair (OIR) annotation. Your task is to analyze disagreements between two annotators who analyzed the same dialogue for OIR sequences but produced different annotations. Provide objective, balanced explanations that help understand why the disagreement occurred.

Task Instructions

Other-initiated Self-repaired, or Other-initiated Repair (OIR), refers to the practices in interaction where a recipient interrupts the ongoing conversational activity to address possible trouble/reparable in speaking, hearing, or understanding, but leaves it to the speaker of the trouble source to accomplish the actual repair. Two annotators A and B have analyzed the same dialogue for OIR sequences but produced different annotations. Given a dialogue and two annotations from two annotators, your task: read the full dialogue to understand sequential context, compare Annotator A vs Annotator B annotation, identify **where** they differ, reason for each annotator's decision, and provide recommendation about which annotation better aligns with the OIR definition, and justify it with evidence.

Input

```
{dialogue} | {annotation_A} | {annotation_B}
```

Required Output (Structured Format)

Response a JSON object with the following structure:

```
{"dialogue_id": string, "disagreement_summary":
string, "disagreement_types": one of
[sequence_identification,
boundary_dispute, repair_initiation_type,
component_identification, others (specified)],
"supporting_evidence": [Evidence],
"annotator_analysis": {
"A": {"annotation_summary": string,
  "rationale": [Reason],
  "supporting_evidence": [Evidence]},
"B": {"annotation_summary": string,
  "rationale": [Reason],
  "supporting_evidence": [Evidence]}},
"recommendation": {"favors": select one
[strongly_favors_a, slightly_favors_a, neutral,
slightly_favors_b, strongly_favors_b],
"justification": [Reason], "supporting_evidence":
[Evidence], "confidence": float in [0,1]}}
```

C. Sample of Few-shot Example

Few-shot Example: OIR Annotation (NOXI)

Dialogue

```
1.A [00:02:34.120-00:02:38.450]: j'ai vraiment
adoré l'exposition sur le mouvement
impressionniste au musée
2.B [00:02:38.780-00:02:39.890]: le mouvement quoi?
3.A [00:02:40.120-00:02:43.560]: l'impressionnisme
tu sais Monet Renoir et les autres
4.B [00:02:43.890-00:02:45.230]: ah oui d'accord
je vois
```

Annotation Output

```
{"dialogue_id":"noxi_ex_1","language":"french",
"oir_sequences":[{"sequence_id":"seq_1",
"trouble_source":{"segment_id":"1",
"speaker_name":"A", "start_time":"00:02:34.120",
"end_time":"00:02:38.450",
"text":"j'ai vraiment adoré l'exposition sur le
mouvement impressionniste au musée"},
"repair_initiation":{"segment_id":"2",
"speaker_name":"B", "start_time":"00:02:38.780",
"end_time":"00:02:39.890", "text":"le mouvement
quoi?", "type":"restricted_request"},
"repair_solution":{"segment_id":"3",
"speaker_name":"A", "start_time":"00:02:40.120",
"end_time":"00:02:43.560",
"text":"l'impressionnisme tu sais Monet Renoir et
les autres"}, "confidence":0.9,
"decision_steps":{
"step1":{"value":true,"explanation":"B signals
trouble with 'le mouvement quoi?', indicating
issue with the term connecting to 'mouvement'",
"confidence":0.95},
"step2":{"value":true,"explanation":"A clarifies
'l'impressionnisme' and provides examples",
"confidence":0.9},
"step3":{"value":true,"explanation":"partial
repetition 'le mouvement' + 'quoi' localizes
trouble","confidence":0.9},
"step4":{"value":false,"explanation":"no candidate
understanding offered","confidence":0.9}}}]}
```