Context Does Matter: Implications for Crowdsourced Evaluation Labels in Task-Oriented Dialogue Systems

Clemencia Siro **Mohammad Aliannejadi** **Maarten de Rijke**
University of Amsterdam, Amsterdam, The Netherlands
{c.n.siro,m.aliannejadi,m.derijke}@uva.nl

Abstract

Crowdsourced labels play a crucial role in evaluating task-oriented dialogue systems (TDSs). Obtaining high-quality and consistent ground-truth labels from annotators presents challenges. When evaluating a TDS, annotators must fully comprehend the dialogue before providing judgments. Previous studies suggest using only a portion of the dialogue context in the annotation process. However, the impact of this limitation on label quality remains unexplored. This study investigates the influence of dialogue context on annotation quality, considering the truncated context for relevance and usefulness labeling. We further propose to use large language models (LLMs) to summarize the dialogue context to provide a rich and short description of the dialogue context and study the impact of doing so on the annotator's performance. Reducing context leads to more positive ratings. Conversely, providing the entire dialogue context yields higher-quality relevance ratings but introduces ambiguity in usefulness ratings. Using the first user utterance as context leads to consistent ratings, akin to those obtained using the entire dialogue, with significantly reduced annotation effort. Our findings show how task design, particularly the availability of dialogue context, affects the quality and consistency of crowdsourced evaluation labels.[1]

1 Introduction

With recent advances in pre-trained language models and large language models (LLMs), task-oriented dialogue systems (TDSs) have redefined how people seek information, presenting a more natural approach for users to engage with information sources (Budzianowski and Vulić, 2019; Wu et al., 2020). As TDSs become increasingly integral to information-seeking processes, the question of how to accurately and effectively evaluate their performance becomes critical. Due to the poor correlation of automatic metrics with human-generated labels (Deriu et al., 2021), evaluation of TDSs has shifted towards relying on user ratings or crowdsourced labels as ground-truth measures (Li et al., 2019).

Various crowdsourcing techniques have been employed to collect ground-truth labels, such as sequential labeling (Sun et al., 2021), where the annotators go through each utterance and annotate them one by one. This approach introduces certain risks in the annotation process, such as annotators' fatigue and high cognitive load in extra-long dialogues, requiring them to remember and track the state of the dialogue as they annotate the utterances (Siro et al., 2022). While following and understanding the dialogue context is crucial and can influence the annotators' ratings, reading and understanding very long dialogues can lead to degraded performance.

To address this issue, another line of research proposes to randomly sample only a few utterances in each dialogue to be annotated (Mehri and Eskenazi, 2020; Siro et al., 2022, 2023). While addressing the high cognitive load and fatigue, limiting annotators' understanding of the dialogue poses obvious risks, such as unreliable and biased labels (Schmitt and Ultes, 2015; Siro et al., 2022). In particular, the amount of dialogue context can lead to biases. For example, annotators who lack rich context may unintentionally lean towards positive or negative ratings, neglecting the broader quality of the response. Thus, offering annotators too little context risks misleading judgments, potentially leading to inaccurate or inconsistent labels. Conversely, flooding annotators with excessive information can overwhelm them, which can lead to lower returns in terms of label quality.

Prior work has investigated factors that affect the quality and consistency of crowdsourced evaluation labels, including annotator characteristics,

[1]To foster research in this area, we release our data publicly at https://github.com/Clemenciah/Effects-of-Dialogue-Context

task design, cognitive load, and evaluation protocols (see, e.g., Parmar et al., 2023; Roitero et al., 2021, 2020; Santhanam et al., 2020). However, no previous work studies the effect of random sampling and the number of sampled utterances on the annotation quality.

In this study, we aim to address this research gap by investigating how different amounts of contextual information impact the quality and consistency of crowdsourced labels for TDSs, contributing to understanding of the impact of such design choices. We experiment with crowdsourcing labels for two major evaluation aspects, namely, *relevance* and *usefulness* under different conditions, where we compare the annotation quality under different dialogue context truncation strategies.

Addressing the challenge of insufficient context at the turn level, we propose to use heuristic methods and LLMs to generate the user's information need and dialogue summary. LLMs can play the role of annotation assistants (Faggioli et al., 2023) by summarizing the dialogue history, facilitating a more efficient and effective understanding of the dialogue context before annotating an utterance. To this aim, we use GPT-4 for dialogue context summarization and compare the performance of annotators' under different conditions, as well as different context sizes. Through these experiments, we answer two main questions: **(RQ1)** How does varying the amount of dialogue context affect the crowdsourced evaluation of TDSs? **(RQ2)** Can the consistency of crowdsourced labels be improved with automatically generated supplementary context?

Our findings reveal that the availability of previous dialogue context significantly influences annotators' ratings, with a noticeable impact on their quality. Without prior context, annotators tend to assign more positive ratings to system responses, possibly due to insufficient evidence for penalization, introducing a positivity bias. In contrast, presenting the entire dialogue context yields higher relevance ratings. As for usefulness, presenting the entire dialogue context introduces ambiguity and slightly lowers annotator agreement. This highlights the delicate balance in contextual information provided for evaluations. The inclusion of automatically generated dialogue context enhances annotator agreement in the no-context (C_0) condition while reducing annotation time compared to the full-context (C_7) condition, presenting an ideal balance between annotator effort and performance.

Our findings extend to other task-oriented conversational tasks like conversational search and preference elicitation, both relying on crowdsourced experiments to assess system performance.

2 Methodology

We examine how contextual information about a dialogue affects the consistency of crowdsourced judgments regarding *relevance* and *usefulness* of a dialogue response. Here, contextual information refers to the information or conversation that precedes a specific response. We carry out experiments in two phases. **Phase 1** involves varying the *amount* of dialogue context for annotators to answer **RQ1**. In **Phase 2**, we vary the *type* of previous contextual information available to annotators to address **RQ2**.

2.1 Experimental data and tasks

We use the recommendation dialogue (ReDial) dataset (Li et al., 2018), a conversational movie recommendation dataset, comprising of over 11K dialogues. The dataset is collected using a human-human approach, i.e., one person acts as the movie seeker, while the other is the recommender with the goal of recommending a suitable movie to the seeker, thus making the dataset goal-oriented. We randomly select system responses from 40 dialogues for the assignment of relevance and usefulness labels. These dialogues typically consist of 10 to 11 utterances each, with an average utterance length of 14 words. We evaluate the same system responses across all experimental conditions.

The annotation task for the annotators involves two dimensions: (i) *relevance*: Is the system response relevant to the user's request, considering the context of the dialogue? And (ii) *usefulness*: How useful is the system's response given the user's information need? For the *relevance task* we ask annotators to judge how relevant the system's recommendations are to the user's request (Alonso et al., 2008). First, the annotator has to judge whether the system response includes a movie recommendation or not; if yes, the annotator assesses whether the movie meets the user's preference; if not, we ask them to note that the utterance does not recommend a movie. The judgment is on a binary scale for the latter case, where the movie is either relevant (1) or not (0). For each experimental condition (see below), annotators only assess the system

response with access to the previous context. Note that we forego the user's feedback on the evaluated response (next user utterance) so as to focus on topical relevance of the recommended movie, that is, if the movie meets the user request and preference in terms of the genre, actor, director, etc. For the *usefulness task* annotators assess a response with or without a movie recommendation with the aim of determining how useful the system's response is to the user (Mao et al., 2016). The judgment is done on a three-point scale (i.e., very, somewhat, and not useful). Unlike the relevance task, annotators have access to the user's next utterance for the usefulness task; usefulness is personalized to the user, in that even though a movie may be in the same genre, sometimes a user may not like it (e.g., does not like the main actor), thus making the system response relevant but not useful to the user.

2.2 Automatic generation of diverse dialogue contexts

User information need. The user's information need plays a significant role when assessing or improving the quality of the data collected in IR systems (Mao et al., 2016). It refers to *the specific requirement or query made by a user, which guides the system in understanding their preferences and retrieving relevant information to fulfill that need*. For TDSs, understanding the user's intent is crucial for annotators participating in the evaluation, as they are not the actual end users. This understanding improves the alignment of evaluation labels with the actual user's requirements. We define the user's information need as their movie recommendation preference. Given the consistency of user preferences in the ReDial dataset, where users tend to maintain a single preference throughout a conversation, providing the user's initial information need aids annotators in evaluating the current turn for relevance or usefulness.

We adopt two approaches to generate the user's information need. One is to heuristically extract the first user utterance that either requests a movie recommendation or expresses a movie preference, based on phrases such as "looking for," "recommend me," and "prefer." These phrases are extracted from the first three user utterances in a dialogue, with the top 10 most common phrases selected. The second approach relies on LLMs to generate the user's information need. We hypothesize that LLMs can identify pertinent user utterances in a dialogue and generate the corresponding information need. We use GPT-4 (OpenAI, 2023) in a zero-shot setting; with the dialogue context up to the current turn as input, we prompt the model to generate the user's information need.

Generating dialogue summaries. Dialogue summarization is beneficial for providing a quick context to new participants of a conversation and helping people understand the main ideas or search for key contents after the conversation, which can increase efficiency and productivity (Feng et al., 2022). We use dialogue summaries to provide annotators with quick prior context of a dialogue. We use GPT-4 (OpenAI, 2023) in a zero-shot setting, as in the case of user information needs, but vary the prompt. We instruct GPT-4 to generate a summary that is both concise and informative, constituting less than half the length of the input dialogue. Both the generated user information needs and summaries are incorporated in Phase 2 of the crowdsourcing experiments.

Due LLMs' potential for hallucination (Bouyamourn, 2023; Chang et al., 2023), we evaluate the generated summaries and user information need to ensure factuality and coherence. We elaborate the steps we took in Section A.2.

2.3 Crowdsource experiments

Following (Kazai, 2011; Kazai et al., 2013; Roitero et al., 2020), we design human intelligence task (HIT) templates to collect relevance and usefulness labels. We deploy the HITs in variable conditions to understand how contextual information affects annotators' judgments. Our study has two phases: in Phase 1 we vary the *amount* of contextual information; in Phase 2 we vary the *type* of contextual information. In each phase and condition, the annotators were paid the same amount as this study is not focused on understanding how incentive influences the quality of crowdsourced labels. Like (Kazai et al., 2013), we refrain from disclosing the research angle to the annotators in both phases; this helps prevent potential biases during the completion of the HIT.

Phase 1. In Phase 1, the focus is on understanding how the *amount* of dialogue context impacts the quality and consistency of relevance and usefulness labels. We vary the length of the dialogue context to address (RQ1). Thus, we design our experiment with three variations: C_0, C_3, and C_7 (see Section 2.4). The HIT consists of a general task de-

scription, instructions, examples, and the main task part. For each variation, we gather labels for two main dimensions (relevance and usefulness) and include an open-ended question to solicit annotators' feedback on the task. Each dimension is assessed with 3 annotators in a separate HIT, with the same system response evaluated by each. This ensures a consistent evaluation process for both relevance and usefulness.

Phase 2. In Phase 2, the focus shifts to the *type* of contextual information, to answer (RQ2). We take an approach of machine in the loop for crowdsourcing. We restrict our experiments to experimental variation C_0 (defined below), where no previous dialogue context is available to the annotators. We aim to enhance the quality of crowdsourced labels for C_0 by including additional contextual information alongside the turn being evaluated. Our hypothesis is that without prior context, annotators may face challenges in providing accurate and consistent labels. By introducing additional context, like the user's information need or a dialogue summary, we expect an increase in the accuracy of evaluations. Through this, we aim to approach a level of performance similar to when annotators have access to the entire dialogue context while minimizing the annotation effort required. We enhance the 40 dialogues from Phase 1 with the user's information need or a dialogue summary, as detailed in Section 2.2. Thus, in Phase 2, we have three experimental setups: C_0-llm, C_0-heu, and C_0-sum. Table 3 in Section A.1 summarizes the setups.

The HIT design closely mirrors that of Phase 1. The main task remains unchanged, except for the inclusion of the user's information need or a dialogue summary. Annotators answer the same two questions on relevance and usefulness in separate HITs. While we do not strictly enforce reliance on the additional information provided, annotators are encouraged to use it when they perceive that the current response lacks sufficient information for an informed judgment.

2.4 Experimental conditions

We focus on two key attributes: the *amount* and *type* of dialogue context. For both attributes, we explore three distinct settings, resulting in 6 variations, for both relevance and usefulness; each was applied to the same 40 dialogues:

- *Amount of context*. We explore three truncation strategies: no-context (C_0), partial context (C_3), and full context (C_7), designed to encompass scenarios where no previous dialogue context is accessible to the annotator (C_0), where some previous dialogue context is available but not comprehensively (C_3), and when annotators have access to the complete previous dialogue context (C_7).
- *Type of context*. Using the contexts generated in Section 2.2, we experiment with three variations of context type: heuristically generated information need (C_0-heu), an LLM-generated information need (C_0-llm), and dialogue summary (C_0-sum).

Table 3 in Section A.1 of the appendix summarizes the experimental conditions.

2.5 Participants

We enlisted master workers from the US on Amazon Mechanical Turk (MTurk) (Amazon Mechanical Turk, 2023) to ensure proficient language understanding. Annotators were filtered based on platform qualifications, requiring a minimum accuracy of 97% across 5000 HITs. To mitigate any learning bias from the task, each annotator was limited to completing 10 HITs per batch and participating in a maximum of 3 experimental conditions. A total of 78 unique annotators took part in Phases 1 and 2 and each worker was paid \$0.4 per HIT, an average of \$14 per hour. Their average age range was 35–44 years. The gender distribution was 46% female and 54% male. The majority held a four-year undergraduate degree (48%), followed by two-year and master's degrees (15% and 14%, respectively).

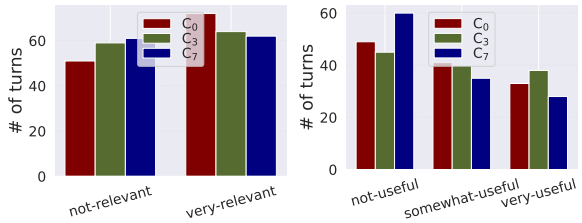
We conduct quality control on the crowdsourced labels to ensure reliability as described in Section A.2 in the appendix.

3 Results and Analysis

We address **(RQ1)** and **(RQ2)** by providing an overview of the results and in-depth analysis of our crowdsourcing experiments. We first describe the key data statistics.

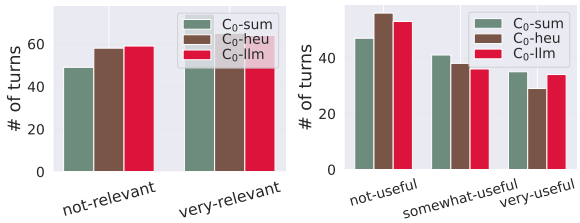
3.1 Data statistics

Phase 1. Figure 1 presents the distributions of relevance and usefulness ratings across the three variations, C_0, C_3, and C_7. Figure 1a indicates a larger number of dialogues rated as relevant when annotators had no prior context (C_0), compared to instances of C_3 and C_7, where a lower number



(a) Relevance (b) Usefulness

Figure 1: Distribution of (a) relevance and (b) usefulness labels for dialogue annotations in Phase 1.



(a) Relevance (b) Usefulness

Figure 2: Distribution of (a) relevance and (b) usefulness ratings when annotators have access to additional context in C_0 Phase 2.

of dialogues received such ratings. This suggests that in the absence of prior context, annotators are more inclined to perceive the system's response as relevant, as they lack evidence to assert otherwise. This trend is particularly prevalent when user utterances lean towards casual conversations, such as inquiring about a previously mentioned movie or requesting a similar recommendation to their initial query, aspects to which the annotators have no access. Consequently, this suggests that annotators rely on assumptions regarding the user's previous inquiries, leading to higher ratings for system response relevance.

We observe a similar trend for usefulness (Figure 1b), compared to C_3 and C_7, C_0 has more dialogues rated as useful. The introduction of the user's next utterance introduced some level of ambiguity to annotators. Evident in instances where the user introduced a new item not mentioned in the system's response and expressed an intention to watch it, the usefulness of the system's response became uncertain. This ambiguity arises particularly when annotators lack access to prior context, making it challenging to tell if the movie was mentioned before in the preceding context.

These observations highlight the impact of the amount of dialogue context on the annotators' perceptions of relevance and usefulness in Phase 1. This emphasizes the significance of taking contextual factors into account when evaluating TDSs.

Phase 2. In Phase 2, we present findings on how different types of dialogue contexts influence the annotation of relevance and usefulness labels. When the dialogue summary is included as supplementary information for the turn under evaluation (C_0-sum), a higher proportion of dialogues are annotated as relevant compared to C_0-llm for relevance (60% vs. 52.5%, respectively); see Figure 2a.

In contrast to the observations made for relevance, we see in Figure 2b that a higher percentage of dialogues are predominantly labeled as not useful when additional information is provided to the annotators. This accounts for 60% in C_0-heu, 47.5% in C_0-llm, and 45% in C_0-sum. This trend is consistent with our observations from Phase 1, highlighting that while system responses may be relevant, they do not always align with the user's actual information need. We find that C_0-sum exhibits the highest number of dialogues rated as useful, indicating its effectiveness in providing pertinent information to aid annotators in making informed judgments regarding usefulness.

3.2 RQ1: Effect of varying amount of dialogue context

Label quality. To gauge the quality of the crowdsourced labels, we rely on inter-annotator agreement (Boguslav and Cohen, 2017; Carletta, 1996). In order to understand how the amount of dialogue context influences the quality of ratings by annotators, we calculate the agreement between annotators for both relevance and usefulness across the three variations; see Table 1. To address potential randomness in relevance ratings, given the binary scale, we randomly drop one rating from each dialogue and compute the agreement. We repeat this process for each annotator and calculate an average Cohen's Kappa score. For usefulness, we compute Kappa for each pair of annotators and then

Table 1: Inter annotator agreement (Cohen's Kappa) and Tau correlation for relevance and usefulness across the three experimental setups in Phase 1.

Aspect	Variation	Kappa	Tau
Relevance	C_0	0.53	0.47
	C_3	0.61	0.49
	C_7	0.70	0.61
Usefulness	C_0	0.64	0.54
	C_3	0.68	0.60
	C_7	0.56	0.41

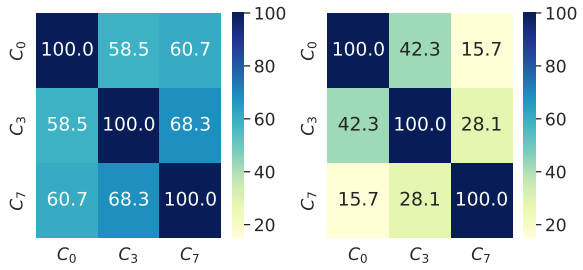
calculate the average. We assess the significance of the agreement using the Chi-squared method. All Kappa scores are statistically significant ($p \leq 0.05$).

We observe an increase in the Kappa and Tau score as the dialogue context increases from C_0 to C_7. Despite the lack of context in C_0, there is a moderate level of agreement regarding the relevance of the current turn. With the introduction of more context in C_3 and C_7, comes an increase in agreement regarding the relevance of the current turn (see Table 1). Providing additional dialogue context seems to lead to higher levels of consensus among annotators. This is likely due to dataset characteristics: users tend to express their preferences early in the dialogue, rather than in subsequent exchanges. Hence, in the case of C_0, which only includes the current turn, when the user's utterance is incomplete, lacking an explicit expression of their preference, annotators rate more dialogues as relevant compared to C_3 and C_7. Overall, we conclude that when annotators have insufficient information to come up with a judgment, they tend to judge the system positively, introducing a positivity bias (Park et al., 2018).

We see in Table 1 (row 3) that despite the lack of context in C_0, there is substantial agreement regarding the usefulness of the current turn. This is due to the availability of the user's next utterance, which serves as direct feedback on the system's response, resulting in higher agreement than for relevance assessment. As more context is provided, there is an even higher level of agreement among annotators regarding the usefulness of the current turn. Access to a short conversation history significantly improves agreement on usefulness.

Surprisingly, despite having access to the entire conversation history in C_7, there is a slightly lower level of agreement than in C_3. The complete dialogue context may introduce additional complexity or ambiguity in determining the usefulness of the current turn. This occurs when conflicting feedback arises from the user's next utterance compared to the previous dialogue context. For example, when the system repeats a recommendation that the user has already watched or stated before, and the user expresses their intent to watch the movie in the next utterance, it leads to divergent labels. Similar trend is observed with the Tau correlations though the values are lower compared to the Kappa scores.

Label consistency across conditions. We examine



(a) Relevance (b) Usefulness

Figure 3: The percentage of agreement in (a) relevance and (b) usefulness labels across the three experimental setups in Phase 1.

the impact of varying amounts of dialogue context on the consistency of crowdsourced labels across the three variations for relevance and usefulness and report the percentage of agreement in Figure 3. We observe moderate agreement (58.54%) between annotations of C_0 and C_3, suggesting that annotators demonstrate a degree of consistency in their assessments when provided with different amounts of context. This trend continues with C_0 and C_7, where the agreement increases slightly to 60.98%. The most notable increase is between C_3 and C_7 (68.29%). As annotators were exposed to progressively broader contextual information, their assessments became more consistent.

Usefulness behaves differently. We observe moderate agreement (41.71%) between C_0 and C_3, indicating a degree of consistency in annotator assessments within this range of context. A notable decrease in agreement is evident when comparing C_3 and C_7, down to 28.3% agreement. The most substantial drop is observed between C_0 and C_7, yielding a mere 14.63% agreement. These findings emphasize the significant impact of context on the consistency of usefulness annotations. For usefulness assessment providing annotators with a more focused context, improves their agreement.

With respect to **RQ1**, we note considerable differences in the labels assigned by annotators as we vary the amount of dialogue context. As the context expands, annotators incorporate more information into their assessments, resulting in context-specific labels. Annotator judgments are shaped not only by response quality but also by the broader conversation. This highlights the complexity of the task and the need for a carefully designed annotation methodology that considers contextual variations. These findings emphasize the significance of dialogue context in annotator decision-making.

Table 2: Inter annotator agreement (Cohen's Kappa) and Tau correlation for relevance and usefulness across the three experimental setups in Phase 2.

Aspect	Variation	Kappa	Tau
Relevance	C_0-heu	0.75	0.54
	C_0-sum	0.60	0.45
	C_0-llm	0.51	0.44
Usefulness	C_0-heu	0.71	0.59
	C_0-sum	0.63	0.49
	C_0-llm	0.53	0.44

3.3 RQ2: Effect of automatically generated dialogue context

Label quality. In Phase 2, our experiments aim to establish the impact of presenting annotators with different types of context during crowdsourcing. Different from conventional dialogue context, we provide the annotators with the dialogue summary (C_0-sum), the user's information need in the dialogue (C_0-heu and C_0-llm). We also aim to uncover if we can improve the quality of the crowdsourced labels in C_0 to match those in C_7. We calculate the Cohen's Kappa similar to Section 3.2; see Table 2.

The heuristic approach (C_0-heu) yields the highest agreement (Kappa and Tau), indicating a noteworthy degree of agreement in relevance assessments. The LLM-generated context (C_0-llm and C_0-sum) results in a moderate to substantial level of agreement, signifying a reasonable level of agreement regarding the relevance of the system response. We observe similar results for usefulness. The heuristic approach (C_0-heu) again leads with the highest level of agreement (0.71 and 0.59), C_0-sum follows with a kappa score of 0.63, while C_0-llm has a kappa score of 0.53. This high level of agreement (Kappa) for the two aspects indicates the quality of the labels; the additional context provided, generated either heuristically or with LLMs, is effective in conveying relevant information to annotators, leading to more consistent assessments.

For both relevance and usefulness, C_0-heu consistently improves agreement among annotators, while the LLM-generated context (C_0-llm and C_0-sum) has a substantially lower agreement than C_7. This difference reflects the limitations of LLMs in capturing context and generating a factual summary. While they generate coherent text, LLMs sometimes fail to correctly represent the sequential order of the dialogue and users' language patterns.

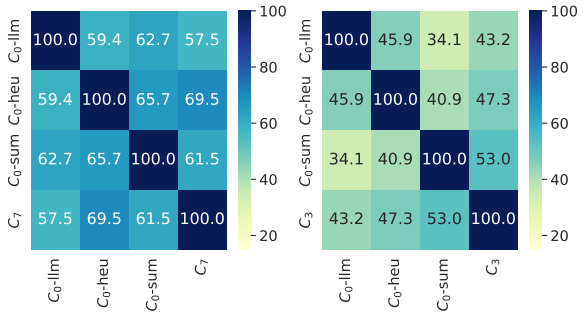
Label consistency across conditions. In Figure 4a we report the agreement between the setups in Phase 2 and compare them to C_7 (relevance) and C_3 (usefulness) due to their high inter-annotator agreement (IAA) and label consistency. For the relevance annotations, varying levels of agreement emerge. There is substantial agreement between C_0-heu and C_0-llm (59.36%), showing a significant overlap in the labels assigned using both methods, although there are instances where annotators differ in their assessments of relevance. C_0-sum exhibits moderate label agreement with C_0-llm (62.74%) and C_0-heu (65.67%), pointing to relatively similar label assignments across the setups.

We observe similar results for usefulness in Figure 4b. While the heuristically generated approach achieves high IAA, the C_0-sum method demonstrates greater consistency with all other setups in terms of usefulness. This suggests that while annotators using the C_0-heu approach often agreed on a single label, the chosen label may not have always been the most accurate. We note slightly low agreement levels for a similar label between the three setups, consistent with results in Phase 1. Unlike relevance, which used a binary scale, usefulness was rated on a 1–3 scale. This finer-grained scale may explain the lower agreement compared to relevance, as different types of contextual information can influence usefulness scores.

Regarding **RQ2**, we show that we can improve the consistency of the labels assigned by crowdworkers in C_0 condition by augmenting the current turn with automatically generated supplementary dialogue context. The heuristic approach demonstrates higher consistency in both IAA and label consistency for relevance and usefulness compared to C_0 and C_7. Providing annotators with the user's initial utterance expressing their preference, particularly in scenarios lacking context, can significantly enhance the quality and consistency of crowdsourced labels. This approach can yield performance comparable to a setup involving the entire dialogue C_7, without imposing the cognitive load of reading an entire conversation on annotators. This streamlines the annotation process and maintains high-quality results, offering a practical strategy for obtaining reliable labels for dialogue evaluation.

4 Discussion and Implications

Our findings reveal intriguing insights into the impact of context size and type on crowdsourced rel-



(a) Relevance (b) Usefulness

Figure 4: The percentage of agreement in (a) relevance and (b) usefulness labels across the three experimental setups in Phase 2.

evance and usefulness labels for TDS. Expanding the dialogue context from C_0 to C_7 significantly improves agreement among annotators, indicating that annotators rely on comprehensive context to make more accurate assessments. This trend does not hold for usefulness, where we notice a decrease in agreement when all previous dialogue context is available. The optimal amount of context required for reliable labels relies on the aspect evaluated.

Consistent with prior work (Eickhoff, 2018; Kazai et al., 2011a), we observe an inconsistency in relevance labels across variations, with the same system response being rated differently depending on the context provided. Given the lack of label consistency across variations, future studies should carefully tailor their annotation task design and test various settings to ensure high-quality and consistent labels. Additionally, much care should be taken when comparing the performance of a system across several datasets when labels are crowdsourced with a different strategy to ensure a fair comparison as models similar to humans can be sensitive to the annotation strategy (Kadasi and Singh, 2023; Kern et al., 2023).

We also analyzed data from the open-ended question asking annotators about their experience with the annotation task. Annotators note that dialogue summaries fail to convey a user's emotion, limiting their annotation process. Additionally, lower accuracy of the context generated by an LLM may lead to low agreement among annotators. This signifies the importance of carefully considering the quality and accuracy of generated content in the evaluation process. We provide examples in Section A.5 in the appendix. While there may be constraints in presenting user information need and dialogue summary as dialogue context, one key consideration to take into account is the cognitive load of annotators. Providing a shorter, focused context reduces the cognitive burden on annotators, allowing them to devote more attention to actually evaluating a response. This not only streamlines the annotation process but also helps maintain high-quality results. Reducing the amount of content to be assessed may lead to faster annotation times without compromising the quality of ratings (Santhanam et al., 2020). Another approach to using LLMs in annotation, is for researchers to consider co-annotation (Li et al., 2023) between humans and LLMs.

Optimal context varies by the aspect under evaluation, challenging the idea of a universal strategy. The consistent reliability of automatic methods suggests their potential as dependable tools for evaluation. This implies their use in generating supplementary context, eliminating the need for manual determination of context amounts. This streamlines evaluation, enhancing efficiency in context-driven evaluations for TDS. For data lacking topic or preference shifts, heuristics perform effectively. However, LLMs are recommended for shifting conditions, showcasing adaptability not easily discernible with heuristics.

While our primary focus was limited to relevance and usefulness, the proposed experimental design can be extended to other aspects of TDSs evaluation. Moreover, our findings may be task- or dataset-specific, prompting the need for further investigation into their generalizability. As to future work, we aspire to enhance the robustness of our findings by conducting studies on larger-scale datasets. In addition following previous work by Kazai et al. (2012, 2013), we would also want to understand the effect of annotator background: experience of interacting with conversational system or prior experience in doing the annotation task on label consistency for TDSs.

5 Related Work

We review related work not covered in the paper so far. Several user-centric dialogue evaluation metrics (Ghazarian et al., 2019; Huang et al., 2020; Mehri and Eskenazi, 2020) have been proposed. For TDSs, high-level dimensions such as user satisfaction (Al-Maskari et al., 2007; Kiseleva et al., 2016) and fine-grained metrics such as relevance and interestingness (Siro et al., 2022) have gained interest. Due to the ineffectiveness of standard evaluation metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), which show poor

correlation with human judgments (Deriu et al., 2021), a significant amount of research on these metrics relies on crowdsourcing dialogue evaluation labels to improve correlation with actual user ratings. Crowdsourcing ground-truth labels has gained momentum in information retrieval (IR) for tasks like search relevance evaluation (Alonso et al., 2008) and measuring user satisfaction in TDS. A major challenge is ensuring quality and consistency of crowdsourced labels. Task design and annotators' behavioral features and demographics can affect the quality of the collected labels (Hube et al., 2019; Kazai et al., 2012; Pei et al., 2021). Kazai et al. (2013) examine how effort and incentive influence the quality of labels provided by assessors when making relevance judgments. Other factors such as judgment scale (Novikova et al., 2018; Roitero et al., 2021), annotator background (Kazai et al., 2011b; Roitero et al., 2020), and annotators' demographics (Difallah et al., 2018) have also been studied. Most studies focus on search systems, not dialogue systems. Closer to our work, Santhanam et al. (2020) study the effect of cognitive bias in the evaluation of dialogue systems. Providing an anchor to annotators introduces anchoring bias, where annotators' ratings are close to the anchor's numerical value. Like Santhanam et al. (2020), we focus on the effect of task design on the evaluation of TDSs. In particular, we investigate how the amount and type of dialogue context provided to annotators affect the quality and consistency of evaluation labels and the annotator experience during the evaluation task.

6 Conclusion

In this work, we investigated the impact of varying the dialogue context size and type on crowdsourced evaluation labels. In particular we crowdsourced evaluation labels for two aspects: *relevance* and *usefulness*. Our findings reveal that optimal context is dependent on the aspect under evaluation. For relevance annotators tend to agree more on a label when they have access to the whole dialogue context. However this does not hold for the usefulness aspect where we witness high annotator agreement when partial context is available. We show that a simple approach like providing an automatically generated user need through heuristics without revealing the entire dialogue can consistently increase annotator agreement across the two aspects. This implies that we can rely on automatic methods such as the use of LLMs to improve the productivity of the crowdworkers by reducing the amount of dialogue they have to read before evaluating the current response.

This study contributes towards how LLMs can be integrated in the annotation process to ensure quality labels from the crowdworkers. In this work we used GPT-4 API which is not open source. For future work we will explore the use of open-source LLMs, like Llama-chat (Touvron et al., 2023), to facilitate a more transparent and reproducible experimental framework.

Limitations

In this work, we dived into the effect of task design on crowdsourced evaluation labels, specifically the amount and type of context available. Nonetheless our study faces some limitations: the absence of actual user ratings hinders us from claiming an optimal strategy for presenting previous dialogue history. Despite this limitation, we highlight the noteworthy observation of high label consistency in C_7 for relevance and C_3 for usefulness aspect, which served as our basis for comparison. It is crucial to note that our study is exploratory in nature and thus may be data or task specific. To ensure the applicability and generalizability of our findings, it is imperative to undertake further investigations to ascertain the extent to which these findings can be extrapolated across different tasks and datasets.

Ethical Considerations

Anotator diversity

All participants in this research were master workers recruited exclusively from the United States through Amazon Mechanical Turk (MTurk). While this selection ensured a level of language proficiency and familiarity with the context, it is crucial to note that the findings of this study may not generalize universally due to the specific demographic representation. The restriction to U.S.-based annotators may introduce a limitation in terms of cultural diversity and global perspectives, influencing the external validity of the study.

Annotator bias

Despite the provision of detailed instructions and examples to annotators, potential biases may still arise during the evaluation process due to the diverse backgrounds of the annotators. Cultural biases may be more pronounced if annotators from

different cultural backgrounds interpret movie preferences, relevance, or usefulness in divergent ways. Subjective biases may also be influenced by the diverse interpretations of guidelines, as individuals from different backgrounds may have distinct views on dimensions like "relevance" or "usefulness."

To mitigate these potential biases, continuous monitoring and feedback mechanisms were incorporated into the study design. Additionally, the study refrained from disclosing the specific research angle to annotators to prevent potential biases related to the research objectives.

7 Acknowledgements

This research was supported by the Dreams Lab, a collaboration between Huawei Finland, the University of Amsterdam, and the Vrije Universiteit Amsterdam, by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, by project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO), and by the FINDHR (Fairness and Intersectional Non-Discrimination in Human Recommendation) project that received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*, page 773–774, New York, NY, USA. Association for Computing Machinery.

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15.

Amazon Mechanical Turk. 2023. `https://www.mturk.com`.

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. In *MEDINFO 2017: Precision Healthcare through Informatics - Proceedings of the 16th World Congress on Medical and Health Informatics, Hangzhou, China, 21-25 August 2017*, volume 245 of *Studies in Health Technology and Informatics*, pages 298–302. IOS Press.

Adam Bouyamourn. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3181–3193. Association for Computational Linguistics.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2 - How can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.*, 54(1):755–810.

Djellel Eddine Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh Association for Computing Machinery International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 135–143. Association for Computing Machinery.

Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh Association for Computing Machinery International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 162–170. Association for Computing Machinery.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 Association for Computing Machinery SIGIR International Conference on Theory of Information Retrieval, ICTIR*

2023, Taipei, Taiwan, 23 July 2023, pages 39–50. Association for Computing Machinery.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5453–5460. ijcai.org.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Pritam Kadasi and Mayank Singh. 2023. Unveiling the multi-annotation process: Examining the influence of annotation quantity and instance difficulty on model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1371–1388. Association for Computational Linguistics.

Gabriella Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, volume 6611 of *Lecture Notes in Computer Science*, pages 165–176. Springer.

Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. 2011a. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceeding of the 34th International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 205–214. Association for Computing Machinery.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2011b. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th Association for Computing Machinery International Conference on Information and Knowledge Management*, CIKM '11, page 1941–1944, New York, NY, USA. Association for Computing Machinery.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st Association for Computing Machinery International Conference on Information and Knowledge Management*, CIKM '12, page 2583–2586, New York, NY, USA. Association for Computing Machinery.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14874–14886. Association for Computational Linguistics.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 Association for Computing Machinery on Conference on Human Information Interaction and Retrieval*, CHIIR '16, page 121–130, New York, NY, USA. Association for Computing Machinery.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1487–1505. Association for Computational Linguistics.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9748–9758.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 2016. When does relevance mean

usefulness and user satisfaction in web search? In *Proceedings of the 39th International Association for Computing Machinery SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 463–472. Association for Computing Machinery.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Kunwoo Park, Meeyoung Cha, and Eunhee Rhim. 2018. Positivity bias in customer satisfaction ratings. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 631–638. Association for Computing Machinery.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't blame the annotator: Bias already starts in the annotation instructions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1771–1781. Association for Computational Linguistics.

Weiping Pei, Zhiju Yang, Monchu Chen, and Chuan Yue. 2021. Quality control in crowdsourcing based on fine-grained behavioral features. *Proc. Association for Computing Machinery Hum. Comput. Interact.*, 5(CSCW2):442:1–442:28.

Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for information retrieval evaluation. *Inf. Process. Manag.*, 58(6):102688.

Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can the crowd identify misinformation objectively?: The effects of judgment scale and assessor's background. In *Proceedings of the 43rd International Association for Computing Machinery SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 439–448. Association for Computing Machinery.

Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–13. Association for Computing Machinery.

Alexander Schmitt and Stefan Ultes. 2015. Interaction quality: Assessing the quality of ongoing spoken dialog interaction by experts - and how it relates to user satisfaction. *Speech Commun.*, 74:12–36.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding user satisfaction with task-oriented dialogue systems. In *Proceedings of the 45th International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2018–2023, New York, NY, USA. Association for Computing Machinery.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2023. Understanding and predicting user satisfaction with conversational recommender systems. *ACM Transactions on Information Systems*, 42(2):Article 55.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*, page 2499–2506, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 917–929. Association for Computational Linguistics.

A Appendix

In this section we provide supplementary materials used to support our main paper. These materials include: experimental conditions elaborated in Section A.1, quality control measures undertaken to ensure high quality crowdsourced labels and generated supplementary context in Section A.2 and the prompts used to generate the supplementary context in Section A.3. In Section A.4 we include the annotation instructions and screen dumps of our annotation task. Section A.5 shows sample supplementary context generated by GPT-4.

A.1 Experimental conditions

We list the experimental conditions used for our crowdsource experiments in Table 3.

A.2 Data quality control

Generated user information need and summary. To address the potential hallucination of LLMs (Chang et al., 2023), we implemented a quality control process for the generated user information needs and summaries, ensuring their coherence and factual accuracy. We automatically cross-reference the movies mentioned in both the input dialogues and the summaries. A summary must contain at least two-thirds of the movies mentioned in the input dialogue to be considered valid. If this criterion is not met, the summary is discarded, and a new one is generated following the specified prompt requirements. In total, we discarded and regenerated 15 dialogue summaries. To further ensure coherence, we randomly sampled 30% of the generated summaries and information needs. The authors reviewed them to confirm their coherence and alignment with the information presented in the input dialogue. This process enhanced the quality and reliability of the generated content.

Crowdsourced labels. To ensure a high quality of the collected data, we incorporated attention-checking questions into the HIT. Annotators were required to specify the number of utterances in the dialogues they were evaluating and to identify the last movie mentioned in the system response being evaluated. 10% of the HITs were rejected and returned back to collect new labels. In total, we gathered 1440 data samples from the crowdsourcing task, spanning six variations for relevance and usefulness. We employed majority voting to establish the final relevance and usefulness dialogue label.

A.3 Prompts

In Table 4 we show the final prompts used to generate the user information and dialogue summary with GPT-4.

A.4 Annotation instructions and screen dumps

Table 5 details the annotation instructions for the relevance and usefulness evaluations. In Figure 5 and 6 we show the annotation interface used for Phase 1 and Phase 2, respectively.

A.5 Sample supplementary context

In Table 6 we show sample user information need and summary generated by GPT-4.

Table 3: Descriptions of the experimental setups used for the crowdsourcing experiments with corresponding relevance and usefulness labels. Unlike relevance, usefulness includes the user's next utterance as feedback. A "turn" denotes a user-system exchange.

Variations	Description
C_0	Current turn with no previous dialogue context
C_3	Current turn with three system-user utterances as previous context
C_7	Current turn with 7 user-system utterances as previous context
C_0-llm	Current turn with an LLM-generated user information need as dialogue context
C_0-heu	Current turn with a heuristically generated user information need as dialogue context
C_0-sum	Current turn with a dialogue summary as dialogue context

Table 4: Prompts used to generate the supplementary context; user information need and dialogue summary with GPT-4.

Dialogue summary prompt
Below you are provided with dialogues between a user and the system about movie recommendations. Generate a complete short and informative summary extractively which is half the length of the dialogue.
User information need prompt
Given the following user and system dialogue in a movie recommendation conversation, generate a concise user's goal in a natural manner. State only the goal without extra text. Start the sentence with "the user wants."

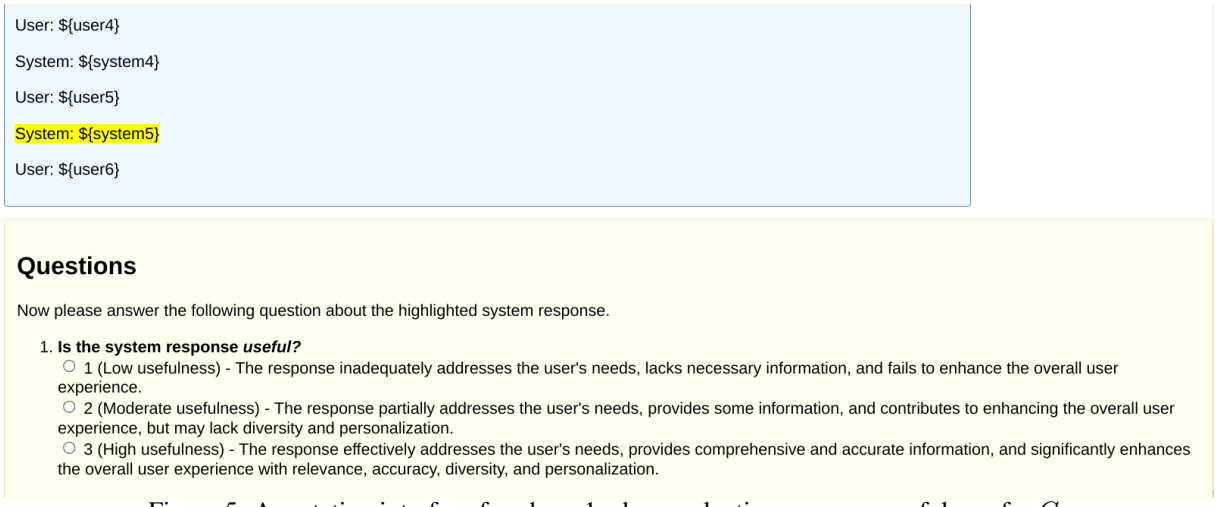


Figure 5: Annotation interface for phase 1 when evaluating response usefulness for C_3

Figure 6: Annotation interface for phase 2 when evaluating response usefulness with supplementary context

Table 5: Annotation instructions provided to the annotators for relevance evaluation. The instructions are the same for usefulness apart from the aspect being evaluated.

Introduction
Thank you for helping us out! Below we explain everything in full detail. Please make sure to read the instructions carefully.
Purpose
The aim of this survey is to evaluate the quality of a system's response. We want to evaluate the dialogue system's performance and gather insights for improvements. We will ask you to evaluate the system response on one metric, that we will discuss in more detail below.
Scenario Outline
Imagine you are evaluating a dialogue system that generates a response to user queries. Your task is to assess the response based on relevance. We will provide examples and detailed explanations of this criteria below.
Task
In each HIT, you will be presented with a dialogue chunk. Your task is to evaluate the last system response based on the given criteria. Please review the explanations and examples for the criteria to ensure your understanding before proceeding with the evaluation. Keeping the scenario that was outlined above in mind, we would like to ask you to judge the system response on relevance.

Table 6: Sample dialogue summaries as supplementary context generated by GPT-4.

Dialogue 1
User inquires about a good family movie recommendation similar to "Real Steel (2011)" or "The Lego Movie (2014)". System recommends "Super (2010)", an action-comedy about a regular guy who becomes a self-made superhero, describing it as hilarious and entertaining. The user shows interest in this recommendation.
Dialogue 2
The user asked for coming-of-age movie recommendations and mentioned they enjoyed "My Girl (1991)" and "Lucas (1986)". The system suggested watching "The Spectacular Now (2013)", a film where Shailene Woodley stars as a character who forms a bond with a troubled classmate.
Dialogue 3
User seeks a dramatic love story to watch. System recommends "The Notebook (2004)", but the user has watched it, as well as "Titanic (1997)". Both films are favored by the user; they desire to watch something new.
Dialogue 4
The user requests animated movie recommendations following their enjoyment of "The Incredibles (2004)". The system suggests other movies, including "Monsters, Inc. (2001)" and its sequel "Monsters University (2013)", which the user approves. The conversation pivots to the topic of successful sequels, citing "Toy Story 3 (2010)" as an example despite the user's disagreement, favoring the original movie, "Toy Story (1995)".
Dialogue 5
The user wants to find a thrilling crime movie like "Thor: Ragnarok (2017)" for their weekend. The system suggested they watch "The Snowman (2017)" but the user declined. However, the system then gave another recommendation, "First Kill (2001)".