Efficient Data Generation for Source-grounded Information-seeking Dialogs: A Use Case for Meeting Transcripts

Lotem Golany[*,1,3], **Filippo Galgani**[*,1], **Maya Mamo**[*,1], **Nimrod Parasol**[1], **Omer Vandsburger**[1], **Nadav Bar**[1], **Ido Dagan**[1,2]
[1]Google Research, [2]Bar-Ilan University
[3]**Corresponding author:** mlotem@google.com

Abstract

Automating data generation with Large Language Models (LLMs) has become increasingly popular. In this work, we investigate the feasibility and effectiveness of LLM-based data generation in the challenging setting of source-grounded information-seeking dialogs, with response attribution, over long documents. Our source texts consist of long and noisy meeting transcripts, adding to the task complexity. Since automating attribution remains difficult, we propose a semi-automatic approach: dialog queries and responses are generated with LLMs, followed by human verification and identification of attribution spans. Using this approach, we created MISeD – Meeting Information Seeking Dialogs dataset – a dataset of information-seeking dialogs focused on meeting transcripts. Models finetuned with MISeD demonstrate superior performance compared to off-the-shelf models, even those of larger size. Finetuning on MISeD gives comparable response generation quality to finetuning on fully manual data, while improving attribution quality and reducing time and effort.

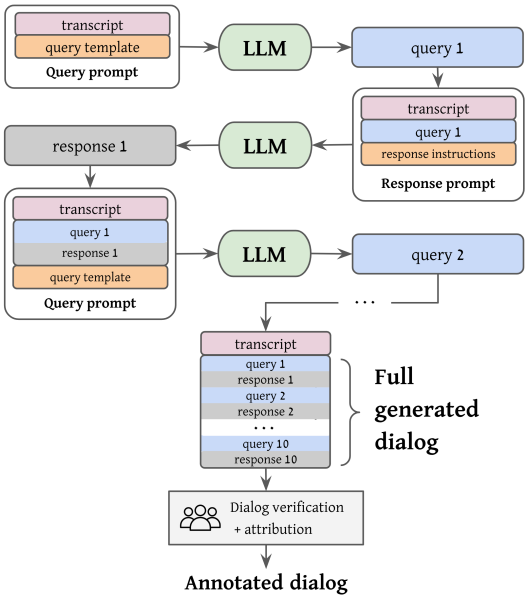

Figure 1: Iterative dialog generation flow. In each turn, a query prompt guides the LLM to generate a user query given the transcript, the accumulated dialog history, and a query template. Then, a response prompt, accompanied by the full context so far, generates the agent response. Iterating this automatic process yields a full dialog, which is then validated by annotators, who further augment it with response attributions.

1 Introduction

Source-grounded information-seeking dialogs allow users to efficiently navigate within a given knowledge source and extract information of interest. In this conversational setting, a user interacts with an agent over multiple rounds of queries and responses regarding the source text (Reddy et al., 2019, Gopalakrishnan et al., 2019, Feng et al., 2020). To train effective agent models, quality dialog datasets are essential.

The prominent (manual) technique for creating dialog datasets is the Wizard-of-Oz (WOZ) method (Kelley, 1984, Budzianowski et al., 2018), in which two human annotators collaboratively produce dialogs: one annotator acts as the user, asking questions about a hidden text, while the other annotator acts as the agent, using the source text to provide answers. The fully manual WOZ methodology is often time-consuming and can result in answers that vary in quality across annotators.

Previous work has explored the application of LLM-based dialog generation in domains like everyday conversations (Chen et al., 2023) or task-oriented dialogs (Sun et al., 2021; Mehri et al., 2022). In this work, we investigate potential automation of the WOZ process to create source-grounded information-seeking dialogs. While LLMs can generate dialog content, attribution

[*]These authors contributed equally.

Findings of the Association for Computational Linguistics: EACL 2024, pages 1908–1925

generation currently requires human involvement for reasonable quality. Therefore we propose a semi-automatic approach: using prompts to guide LLM generation of queries and responses, followed by manual attribution and validation of response correctness (Figure 1). By comparing our semi-automatic approach to the traditional, fully-manual WOZ method, we demonstrate that a model trained on our data achieves comparable response generation quality while improving attribution and reducing costs.

We apply our methodology to create MISeD – the first dataset for information-seeking dialogs over meeting transcripts, supporting the use-case of users catching up on meetings they have missed. Meeting transcripts present additional challenges as they often contain disfluencies, interruptions, and off-topic comments—challenges that are also representative of a broad range of real-world conversational settings. Existing datasets in the meeting domain provide summarization and question-answering data (e.g., Zhong et al. 2021, Prasad et al. 2023), but none supports multi-turn dialogs over meeting content. The dialog setup introduces additional complexity, as each query and response must consider the shared dialog state (e.g., previously shared information) in addition to the transcript, to maintain a coherent dialog flow.

We further present an evaluation framework and few baseline models, whose evaluation demonstrates the benefit of training with our MISeD data.

In summary, our main contributions are: (1) we explore the feasibility of using LLMs for data generation in source-grounded information-seeking dialogs, demonstrating improved overall quality and efficiency compared to a fully manual approach; (2) we release the MISeD dataset[1] – the first dialog dataset over meeting transcripts, consisting of verified source-grounded dialogs with transcript attribution, as well as a corresponding smaller fully manual (WOZ) dataset; (3) we introduce baseline models, an evaluation procedure, and baseline results for meeting-grounded dialog tasks, through which we assess our proposed approach and the MISeD dataset.

2 Related Work

This section provides background on the two research areas which our work bridges: source-grounded information-seeking dialogs (§2.1) and summarization and question answering (Q&A) over meeting transcripts (§2.2).

2.1 Source-grounded Information-seeking Dialogs

Source-grounded information-seeking dialogs are multi-turn conversational interactions, where users seek information from a given source text. For each user query, the agent model provides a response, with supporting references (attributions).

Existing datasets for this task address different types of knowledge sources. Some retrieve answers from large textual corpora (Dinan et al., 2019, Campos et al., 2020, Anantha et al., 2021, Adlakha et al., 2022), others rely on short text passages (Choi et al., 2018, Saeidi et al., 2018, Reddy et al., 2019, Nakamura et al., 2022), long informative conversations (Wu et al., 2022a), or news articles (Li et al., 2023). Our work creates a dialog dataset over the source of long meeting transcripts.

The prominent approach to creating information-seeking dialog datasets is Wizard-of-Oz (WOZ) (Kelley, 1984). In this setup, two annotators role-play as a user and an agent: the user annotator asks questions about the given hidden source, while the agent annotator, who can access the source, provides corresponding answers.

Building on recent work to automate dialog data generation with LLMs (Lin et al., 2022; Wu et al., 2022b; Bao et al., 2023; Zheng et al., 2022; Sun et al., 2021; Mehri et al., 2022; Chen et al., 2023; Li et al., 2022; Chen et al., 2022), we focus on source-grounded information-seeking dialog creation through a combination of LLM-based generation and human verification and attribution.

2.2 Summarization and Q&A over Meeting Transcripts

Meeting transcripts pose unique challenges due to their unstructured and lengthy nature, and potential speech recognition errors. Existing datasets for inquiring meeting content are limited to single-turn settings, focusing on summarization and Q&A. Examples include AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003) for meeting summarization, ELITR (Nedoluzhko et al., 2022) and MeetingBank (Hu et al., 2023) for meeting minuting, MUG (Zhang et al., 2023) with summaries and additional annotations, and ExplainMeetSum (Kim et al., 2023) that also incorporates attributions.

[1] We make our datasets publicly available at https://github.com/google-research-datasets/MISeD.

MeeQA (Apel et al., 2023) and MeetingQA (Prasad et al., 2023) are additional Q&A datasets based on questions asked by the participants during the meeting itself. Differently from this body of work, we focus on a free dialog setup, where queries are not limited, and can build upon each other (e.g., follow-up questions). To our knowledge, there is no existing dataset for information-seeking dialogs over meeting transcripts.

Our work leverages QMSum (Zhong et al., 2021), a widely used dataset for single-turn query-based summarization over meeting transcripts, where summaries are also supported by attribution text spans. QMSum's creation involved annotators generating queries based on a predefined schema, including both general queries addressing the entire meeting and specific queries targeting identified topics and participants. In our work, we modify the QMSum schema to suit LLM prompts (rather than human annotators) and extend it to generate multi-turn dialogs rather than standalone question-answer pairs. After analyzing user data collected in our system prototype, we incorporated additional query types that were not covered by QMSum, such as unanswerable queries, yes/no queries, and follow-up, context-dependent dialog queries. Section 7.5 compares our dataset with QMSum for model training, and Table 16 (Appendix) shows examples of MISeD data compared to QMSum data.

3 Source-grounded Information-seeking Dialogs: Task Definition

As described above, we focus on generating data to train agent models for the task of source-grounded information-seeking dialogs. The agent model task (demonstrated in Figure 2) is defined as follows: at each dialog turn the user issues a query about the source text. The agent model receives the source text, the preceding dialog history, and the current query, and is tasked with generating a response and providing the supporting attributions. Each attribution is a consecutive span in the source text.

4 Dataset Creation Methodology

This section details our semi-automatic dataset creation methodology, as applied to meeting transcripts (Figure 1). We first automatically generate dialogs using a pre-trained LLM (§4.1), simulating the typical WOZ process. Then we employ human annotators to generate response attributions, a task that currently necessitates human expertise. As part

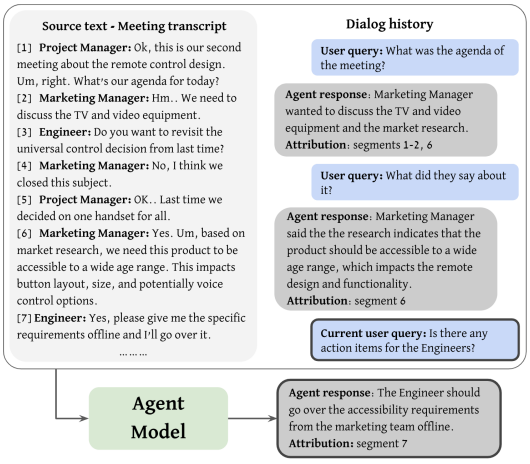

Figure 2: An illustration of the agent model task. The agent receives the source text, dialog history, and the current user query. It then generates a corresponding response along with supporting attributions in the source text. Each attribution is a sequence of consecutive transcript segments.

of the attribution process, annotators also verify the accuracy and validity of the generated queries and responses (§4.2).

4.1 Automatic Dialog Generation

We generate dialog turns iteratively, with each turn consisting of a user query and an agent response. These are generated via targeted LLM prompts, designed for queries (§4.1.1) and responses (§4.1.2).

4.1.1 User query prompts

In each turn, a user query is generated by a "query prompt" (Table 9 in Appendix A). This prompt incorporates the transcript, the dialog history, and a templated instruction randomly selected from a pool of query instructions (Table 10 in Appendix A), designed to include various query types.

Starting from the QMSum schema (§2.2), we create a corresponding set of query templates adapted for guiding LLMs. We include both *General* queries for overall meeting themes, and *Specific* queries which focus on particular topics or individuals. We also expand the schema and include *Yes-no* questions, an *Unanswerable* variant, to generate queries that cannot be answered from the transcript, and *Context-dependent* queries that rely on the existing dialog for context.

While our set of templates does not cover all possible query types in the meeting domain, we suggest that it represents a sufficiently broad range to yield a useful dataset, as assessed in our ex-

periments (§7). Further, we propose that crafting domain-specific query template schemas can be similarly feasible and effective in other domains.

4.1.2 Agent response prompt

After generating the user query, we provide the LLM with a prompt that includes the meeting transcript, dialog history, the current query, and an instruction on how to generate the agent response (Table 11). The instruction guides the LLM to generate an answer that aligns with certain length and format guidelines, and that is grounded in the transcript. It further emphasizes ethical and unbiased communication by instructing the model to use neutral language and avoid direct quotes.

4.2 Dialog Annotation and Validation

Following automatic dialog generation, trained annotators assessed the generated queries and responses while identifying supporting attribution spans within the source text. We first employed 3 professional annotators (details in Appendix B) and conducted a pilot study, measuring Cohen's Kappa agreement between each pair of annotators, over 525 dialog turns. Agreement scores are presented in Table 1.

The high level of agreement observed confirmed the annotation task's clarity and criteria consistency, allowing us to proceed with a single annotator annotating each dialog in the main phase.

4.2.1 Query assessment

Annotators verify query validity by answering the question: "*Is the question understandable and makes sense?*". If a query is marked as invalid, it is removed from the dialog. We also remove all subsequent dialog turns to avoid cascading errors. Valid queries are then annotated with metadata, including their type ('general', 'specific', or 'yes/no') and whether they depend on prior turns ('context-dependent' or not). Context-Dependent query templates specifically target context-dependent queries, so we expect these turns to be context dependent, but annotators still check for context-dependency to confirm that the query does depend on previous turns.

4.2.2 Response annotation: attribution, validation, and editing

Due to the current limitations of LLM-based attribution detection (which we also observed in §7.3.2 and is reported in the literature e.g. (Gao et al., 2023)), we opted for fully manual attribution annotation. For each generated response, annotators identify supporting text spans in the transcript while also verifying that the response is factually correct and editing the response as necessary to ensure its accuracy (more details in Appendix B). If a query cannot be answered from the transcript, it is marked as 'unanswerable' and annotators ensure the response accurately conveys the lack of information.

5 Datasets

This section describes the MISeD dataset (§5.1), constructed using our semi-authomatic method (§4). We also present a smaller independent dataset created using the fully manual Wizard-of-Oz methodology (§5.2), created to assess whether MISeD data is comparable to human-generated data. Finally, we present a manual assessment of MISeD response quality (§5.3).

5.1 MISeD Dataset

Meeting sources Our data creation method was applied on transcripts from the QMSum meeting corpus. We used 225 meetings across three domains: 134 Product Meetings (AMI; Carletta et al., 2005), 58 Academic Meetings (ICSI; Janin et al., 2003), and 33 public Parliamentary Committee Meetings. Figure 3 presents the transcript lengths of meetings in MISeD.

LLM We used the public Gemini Pro model (Gemini Team Google, 2023) to automatically generate dialogs based on the meeting transcripts, as described in Section 4.1.

5.1.1 Dataset structure

Each dataset instance includes a single dialog about a specific meeting transcript, containing up to ten query-response turns with associated metadata (§4.2.1). When relevant, a response is accompanied by a set of attributing transcript spans.[2] In some cases, the response inherently lacks attributions, mostly for 'unanswerable' queries.

For training and evaluating an agent model, each dialog is divided into task instances. Each such instance represents a single current query, incorporating its preceding dialog history along with the corresponding target response and attributions.

[2]Each span is a sequence of consecutive transcript segments.

Assessment Question	Pair A	Pair B	Pair C
"Is the question understandable and makes sense?"	1.0	1.0	1.0
"Does the meeting transcript contain the right answer?"	1.0	0.93	0.93
"Is the provided answer correct based on the meeting transcript?"	0.86	0.90	0.80

Table 1: Cohen's Kappa agreement between each pair of annotators, for each assessment question. Annotators were asked to edit the model response, if needed, in case they replied "yes" to the first two questions.

Overall MISeD Statistics	
# meetings	225
# dialogs	432
# query-response pairs	4161
query type: general	20.91%
query type: specific	52.37%
query type: yes/no	26.72%
context-dependent	13.17%
unanswerable	30.62%
avg. query-response pairs per dialog	9.63
avg. query length (# words)	15.22
avg. response length (# words)	40.80

Table 2: Overall statistics of the MISeD dataset. Queries are classified as either 'general', 'specific', or 'yes/no'. Additionally, queries may be tagged as 'context-dependent' and 'unanswerable'. Table 14 provides statistics by dataset split. Figure 4 shows an additional breakdown by question words.

5.1.2 Dataset statistics

Statistics of the final MISeD dataset are presented in Table 2.

Splits We followed the dataset splits from QMSum, with train, validation and test splits in an approximate ratio of 70:15:15. For each meeting, we aimed to generate two dialogs, each containing ten query-response pairs. Some pairs and dialogs were later filtered during the annotation process.

Context-dependency Through manual analysis of a sample of 100 queries annotated as 'context-dependent', we observed 3 primary characteristics defining context-dependent queries: (1) Detail Seeking: 75 queries sought specific details or clarification regarding a prior response (e.g., "*Besides the ease of use, were there any other advantages mentioned for having an LCD screen?*"); (2) Topic Shifting: 17 queries aimed to change topics or explore other aspects of the meeting not previously discussed ("*What other topics did the participants discuss?*" – response must consider the dialog history to avoid repeating previous topics); (3) Anaphoric Reference: 13 queries utilized pronouns, necessitating interpretation through preceding dialogue turns ("*What were her recommendations to address this?*"); These characteristics can co-occur within a single query ("*What were her other ideas for the input mechanism for the remote?*").

Attribution Nearly all (99%) MISeD responses for answerable queries are supported by transcript attribution, with a median of two attributing transcript spans per response. Attributing spans are relatively long and scattered, with a median length of 96 words, and a median distance of 350 words between subsequent spans (Figure 3).

Process Following the process in Section 4, we generated 443 dialogs comprising 4430 query-response turns. During validation, annotators eliminated 6% of the queries, and corrected 11% of the remaining responses. The average annotation time was 105 minutes per dialog.

5.2 Wizard-of-Oz Dataset

To test the value of our semi-automatic MISeD data compared to fully-manual data, we also collected a smaller set of dialogs using the Wizard-of-Oz (WOZ) methodology. As typical in similar WOZ processes (Choi et al., 2018, Adlakha et al., 2022), the 'user' annotator received a short meeting description, simulating their prior knowledge of the meeting context. They were then instructed to ask free questions to understand aspects of choice about the meeting content. The 'agent' annotator received the full meeting transcript and was tasked with providing free-form answers to the user queries, with supporting attributions.

The collected WOZ data comprises 70 dialogs based on meetings from the test split, with a total of 700 query-response pairs. This data is used in subsequent evaluations (Section 7) for a comparison of model performance. Table 15 in the Appendix presents a comparison between the WOZ data and the MISeD data, with respect to response vocabulary and length.

WOZ dialog annotation took 161 minutes on average, making it 1.5 times more time-consuming than MISeD. While in this work we show the feasibility of automating the generation of queries and

Response evaluation categories	Count	Sum
MISeD 'substantially better'	29	55
MISeD 'slightly better'	26	
'equally good'	2	
Human-only 'slightly better'	21	43
Human-only 'substantially better'	22	

Table 3: Results of MISeD answer quality assessment. Annotators compared 100 pairs of MISeD and fully-manual responses to the same query, determining the better response in each pair.

responses, followed by annotator reviewing, we expect that higher speed-up ratios would be achieved once attribution automation of a reasonable quality is obtained in future research.

5.3 MISeD Response Quality Assessment

MISeD dialogs were generated by an LLM and then validated and corrected by humans (§4). To assess the quality of the agent responses created through this semi-automatic process, we aimed to compare them with responses created fully-manually. To that end, we collected human responses to a randomly selected sample of 100 MISeD queries. Annotators were provided with the meeting transcript and the dialog history up to the selected query and were tasked with answering the query.

Next, we provided a new group of annotators with the manually-generated responses, alongside the original MISeD response for each query (in randomized order between the two), and the corresponding query, dialog history, and meeting transcript. Annotators assessed which response was overall superior, considering correctness, grounding, and clarity, using a scale of 'Equally Good', 'Slightly Better', or 'Substantially Better'. Table 3 presents the results.

Overall, MISeD responses were ranked as better in 55 pairs, compared to 43 pairs for fully-manual responses. A statistical sign test yields $P(x \geq 55) = 0.13$ ($H_0 : p = 0.5$; $n = 98$). These results suggest that the quality of MISeD responses is at least as good as fully-manual responses, and possibly somewhat better, though a larger sample would be needed to establish statistical significance.

6 Evaluation Methodologies

We evaluate the agent models along two dimensions: the quality of the generated responses (§6.1), and the accuracy of the provided attributions (§6.2), through both automatic and human evaluations.

6.1 Response Quality Evaluation

Agent responses are evaluated against the gold responses in the test dataset. Modern LLMs produce outputs that very often exhibit high readability and consistency, therefore we chose to focus our evaluation on the content.

6.1.1 Human evaluation

We conduct human evaluation on a random subset of 100 queries. The annotator is presented with the current user query, the gold response and the model responses, as well as the meeting transcript and the dialog history. To quantify the content overlap between the model response and the gold response, annotators provide scores for recall (how much of the gold response is covered by the model response), and precision (how much of the model response is covered by the gold response). Here covered refers to content overlap between the gold and model responses, in a semantic sense. Both scores are on a Likert scale of 1 to 4, corresponding to 'Not Covered', 'Slightly Covered', 'Mostly Covered' and 'Fully Covered'.

6.1.2 Automatic evaluation

To automatically score model responses we use the standard Rouge-1, Rouge-2 and Rouge-L scores[3] (Lin, 2004), for lexical overlap. To capture semantic overlap, we report BLEURT scores[4] (Sellam et al., 2020), a learned evaluation metric based on BERT (Devlin et al., 2019), trained to model human judgments for reference-based text generation evaluation.

6.2 Attribution Quality Evaluation

To evaluate model attribution, we adopt the AIS protocol (Rashkin et al., 2023) as extended by (Liu et al., 2023) and (Gao et al., 2023). In this approach, attribution evaluation is modeled as a Textual Entailment (Natural Language Inference) task where the generated response text should be entailed by its attributions.

6.2.1 Human evaluation

We sample 100 queries from each test set for human evaluation, after filtering out responses for which it

[3]https://github.com/google-research/google-research/tree/master/rouge

[4]https://github.com/google-research/bleurt

was judged (manually) that attribution is not needed (20% of all responses).

For each model response, annotators assess recall and precision of attributions. For recall, we break the response into sentences asking for each sentence whether it is fully supported by the set of attributions (score of 1) or not (score of 0), reporting micro-average over all sentences. For precision, we ask for each attribution span whether it entails some piece of information in the model response ("partially supports" the response, score of 1), or not (score of 0), reporting micro-average over all attributions.

6.2.2 Automatic evaluation

For automatic attribution evaluation we use the NLI model TrueTeacher (Gekhman et al., 2023) to approximate human entailment judgement, automatically computing recall (the proportion of response sentences fully entailed by the attributions) and precision (the proportion of attributions which contribute to the entailment of the response). Appendix C details the method implementation and discusses its limitations.

7 Baseline Models and Results

7.1 Models

As defined in Section 3, our task input contains the full meeting transcript and the dialog history, ending with the current user query. The output is a concatenation of the response and the set of attributions (indices of supporting segments within the transcript). In the rare occurrences where the context exceeds the model input size capacity, the beginning of the transcript is truncated. Results are compared for the following three model types.

Finetuned Encoder-Decoder LongT5 (Guo et al., 2021) is a T5 (Raffel et al., 2020) variant that uses transient global attention (windowing token averaging) to handle longer input contexts efficiently. We finetuned the open-source[5] LongT5 XL (3 billion parameters) on the MISeD training set, using a context length of 16 thousand tokens.

LLMs prompting we use the Gemini Pro model and the much larger Gemini Ultra model (Gemini Team Google, 2023),[6] without any additional tuning. Our prompt contains the transcript, an instruction, and the dialog ending with the user query (see Table 12 in the Appendix). The models 28 thousand tokens context limit makes it unfeasible to include few-shots examples (which would require providing additional transcripts with corresponding dialogs).

Finetuned LLM we finetune the Gemini Pro model[7] on the MISeD training set, using the same prompt and context length as for the prompting approach. The target format is given in Table 13.

7.2 Datasets

To assess the value of our dataset, in Section 7.3 we fintune the agent models on MISeD data, and test it on both MISeD and Wizard-of-Oz (§5.2) test sets. In Section 7.4 we compare training our best model on the semi-automatic MISeD data to training it on the manually created WOZ data. Section 7.5 reports results on the QMSum query-based summarization test set.

7.3 Results for Dialog Data

7.3.1 Response quality

Table 4 reports response quality results on the MISeD and WOZ test sets. Automatic evaluation covered the full test sets (628 MISeD queries, 700 WOZ queries), while manual evaluation was performed on a random subset of 100 queries from each. Our main takeaways are:

(1) The LongT5 model finetuned on MISeD performs similarly to much larger models (about 0.3 points difference in the 4-points human score), highlighting a key advantage of MISeD as it nearly closes the gap between the smaller 3B-parameter model and the much larger Gemini models.

(2) Finetuning Gemini Pro on MISeD significantly improves its performance, surpassing even the much larger Gemini Ultra model, demonstrating MISeD's effectiveness in boosting performance even for large models.

(3) The scores on WOZ test set exhibit the same trends, but are lower for all models. The lower performance for the non-finetuned models might suggest that fully-manual WOZ data is more challenging, being created freely by human annotators, compared to the templated methodology of MISeD. As expected, the finetuned models performed better on the MISeD test set, as it was created by the same protocol as the training data.

[5] https://github.com/google-research/longt5

[6] https://gemini.google.com/app

[7] https://ai.google.dev/docs/model_tuning_guidance

	Model	Response - Human scores		Response - Automatic scores			
		recall	precision	Rouge-1	Rouge-2	Rouge-L	BLEURT
MISeD	LongT5 Finetuned	2.38	2.52	44.59	27.30	37.62	0.47
	Gemini Pro	2.79	2.71	44.64	27.49	37.35	0.48
	Gemini Ultra	2.63	**2.87**	44.20	26.58	37.39	0.47
	Gemini Pro Finetuned	**2.96**	2.86	**51.02**	**33.38**	**43.03**	**0.52**
WOZ	LongT5 Finetuned	1.79	1.80	26.84	8.95	21.22	0.37
	Gemini Pro	2.10	2.08	27.82	10.08	22.03	0.38
	Gemini Ultra	1.98	**2.14**	28.57	10.99	23.64	0.38
	Gemini Pro Finetuned	**2.21**	2.13	**30.31**	**11.39**	**24.26**	**0.40**

Table 4: Response evaluation: average scores for the MISeD and WOZ test sets.

	Model	Attribution - Human scores			Attribution - Automatic scores		
		recall	precision	F1	recall	precision	F1
MISeD	Longt5 Finetuned	**0.76**	0.58	0.66	0.44	0.19	0.27
	Gemini Pro	0.20	0.69	0.31	0.26	**0.37**	0.31
	Gemini Ultra	0.19	**0.96**	0.32	0.12	0.34	0.18
	Gemini Pro Finetuned	0.68	0.69	**0.68**	**0.50**	0.27	**0.35**
WOZ	Longt5 Finetuned	**0.77**	0.54	0.63	**0.51**	0.17	0.26
	Gemini Pro	0.43	**0.85**	0.57	0.31	0.37	**0.34**
	Gemini Ultra	0.19	0.71	0.30	0.13	**0.40**	0.20
	Gemini Pro Finetuned	0.74	0.67	**0.70**	0.50	0.25	0.33

Table 5: Attribution evaluation: average scores for the MISeD and WOZ test sets.

	Response - Human scores		Response - Automatic scores				Attribution - Automatic scores		
Training Set	recall	precision	Rouge-1	Rouge-2	Rouge-L	BLEURT	recall	precision	F1
MISeD (n=2922)	1.91	1.97	29.73	10.90	23.64	0.40	0.28	0.25	0.26
MISeD (n=500)	**1.93**	**1.99**	30.52	11.10	24.10	0.40	0.26	0.16	0.20
WOZ (n=500)	1.91	1.94	**31.88**	**12.64**	**26.17**	**0.41**	0.02	0.36	0.04

Table 6: Evaluation using different training sets: the semi-automatic MISeD data and the WOZ manual data. Results on 200 WOZ test examples, using Gemini Pro finetuned.

Model	Rouge-1	Rouge-2	Rouge-L	BLEURT
QMSum paper (using retriever)	32.29	8.67	**28.17**	
Gemini Ultra	31.52	9.89	20.88	0.36
LongT5 Finetuned on QMSum	35.34	12.15	23.87	0.35
LongT5 Finetuned on MISeD	29.64	8.19	20.07	0.34
LongT5 Finetuned on both MISeD and QMSum	35.84	12.73	24.37	0.36
Gemini Pro	29.78	8.91	19.52	**0.37**
Gemini Pro Finetuned on QMSum	36.88	12.83	24.52	**0.37**
Gemini Pro Finetuned on MISeD	32.07	9.19	21.14	0.36
Gemini Pro Finetuned on both MISeD and QMSum	**36.98**	**13.62**	25.21	**0.37**

Table 7: Response evaluation for the QMSum test set.

Model	Rouge-1	Rouge-2	Rouge-L	BLEURT
Gemini Pro	27.82	10.08	22.03	0.38
Gemini Pro Finetuned on MISeD	**30.31**	**11.39**	**24.26**	**0.40**
Gemini Pro Finetuned on QMSum	27.72	8.89	21.60	0.38

Table 8: Average scores for response evaluation on the WOZ test sets.

7.3.2 Attribution quality

Attribution evaluation results are presented in Table 5. In line with previous work (Gao et al., 2023), our results suggest that pre-trained LLMs do not excel at finding attributions over meeting transcripts.[8]

Observing the more reliable human scores, we see that the zero-shot models suffer from low recall, due to their tendency to provide attribution less frequently. Finetuning with MISeD data improves performance, substantially increasing recall with a small drop in precision, notably increasing F1 scores for both test sets. The automatic scores show similar though weaker trends, but we regard them as less reliable (see Appendix C).

7.4 Comparison of Training Sets

In this section we compare the performance of models trained on our semi-automatic MISeD data to those trained on the manual WOZ data. We split our WOZ dataset (700 dialog turns) into training and test sets (500:200, respectively). We compare training with (i) 500 WOZ examples (ii) 500 semi-automatic MISeD examples (iii) the full MISeD training set (2922 examples). For response generation, Table 6 suggests that training on 500 semi-automatic MISeD examples yields comparable results to training on the same amount of fully manual examples. For attribution, the model benefits from a larger number of training examples, suggesting that 500 examples are insufficient for this subtask. Notably, the model trained on WOZ only predicted attributions for 6 out of 200 test examples; thus we didn't perform a manual attribution evaluation (as in Table 5). We also tried adding the WOZ data to the MiSeD data in finetuning, but, with respect to automatic evaluation, this did not improve notably the response quality while attribution quality was notably deteriorated, hence we did not pursue this data combination further. To summarize, training on our semi-automatic MISeD data yields response generation quality that is on par with training on the fully manual WOZ data, while also improving attribution, and saving time and effort.

7.5 Results for the QMSum Data

For comparison with existing results, we report results on the QMSum single-turn query-based summarization test set. We compare the model response quality when finetuned on (i) MISeD training data (ii) QMSum training data (iii) both MISeD and QMSum training data together (attribution evaluation is not included in the original work).

Our results (Table 7) indicate that finetuning with MISeD has additive benefits: when the model learns on both MISeD and QMSum data, it surpasses zero-shot models, as well as models trained on QMSum alone or MISeD alone. Our best model, Gemini Pro finetuned on both MISeD and QMSum data, surpasses the performances reported in the original QMSum study for Rouge-1 and Rouge-2. As expected, models trained solely on MISeD do not outperform QMSum-trained models on their own test data: MISeD data is out-of-distribution as it deviates significantly from the QMSum schema, designed for fully manual annotation.

Additionally, Table 8 shows that training on the single-turn QMSum data yields a lower response quality compared to training on MISeD dialog data, while testing on WOZ dialogs. Furthermore, finetuning with QMSum does not improve at all over the non-finetuned model. This demonstrates that non-dialog data composed of standalone question-answer pairs has a limited value when training models for dialog tasks.

8 Conclusion

In this paper, we investigate LLM automation to generate source-grounded information-seeking dialog datasets. We introduce a method to partially automate the WOZ process using targeted user and agent prompts, followed by human attribution, verification, and potential editing. We apply this method to create MISeD, the first dataset for information-seeking dialogs over meetings.

Baseline models and experiments demonstrate the value of MISeD: finetuninig with MISeD data brings the same improvement as a comparable amount of the more expensive fully-manual WOZ data. MISeD enables the creation of modestly sized finetuned encoder-decoder models with performance approaching much larger pre-trained LLMs, and further improves the performance of such LLMs through finetuning. We suggest that our work yields valuable insights about the feasibility of LLM-based data generation in the important and challenging case of source-grounded information-seeking dialogs.

[8] This is consistent with our findings when developing the MISeD annotation methodology (§4.2.2), which led us to leave attribution generation fully manual at this point.

9 Limitations

Attribution Our work successfully automates query and response generation, but highlights the challenge of attribution in both modeling and automatic evaluation. Future research on improving attribution generation models could enable fuller automation of dialogs generation, leading to an even more efficient process.

Long texts depend on LLM context length Our method's reliance on LLM prompting poses challenges for long meeting transcripts that exceed the maximal context length of the model. The same problem occurs in the response generation agent task. Possible solutions include using LLMs with larger context lengths (which may be computationally expensive) or integrating a retrieval system as a first stage in the response generation, to condense the transcript and provide the most relevant information to the LLM.

Manually crafted prompt templates Our method currently depends on manually crafted prompt templates to represent a broad spectrum of potential queries in the given domain. Research into more flexible and diverse query generation strategies could reduce the effort of applying our method in new domains while increasing the generality of the obtained data.

Acknowledgements

We wish to thank David Karam, Michelle Tadmor Ramanovich, Eliya Nachmani, Benny Schlesinger, David Petru and Blaise Aguera y Arcas for their support and feedback on the research. We also would like to express our gratitude to Victor Cărbune, Lucas Werner and Ondrej Skopek for their help with the annotation framework, and to Riteeka Kapila for leading the annotators team.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting.

Reut Apel, Tom Braude, Amir Kantor, and Eyal Kolman. 2023. Meeqa: Natural questions in meeting transcripts.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa – accessing domain-specific faqs via conversational qa.

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

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andrew Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Z. Hakkani-Tür. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. *ArXiv*, abs/2210.14169.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16409–16423. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, pages 364–367. IEEE.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.

Hyun Kim, Minsoo Cho, and Seung-Hoon Na. 2023. ExplainMeetSum: A dataset for explainable meeting summarization aligned with human intent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13079–13098, Toronto, Canada. Association for Computational Linguistics.

Siheng Li, Yichun Yin, Cheng Yang, Wangjie Jiang, Yiwei Li, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. Newsdialogues: Towards proactive news grounded conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yen Ting Lin, Alexandros Papangelis, Seokhwan Kim, and Dilek Hakkani-Tur. 2022. Knowledge-grounded conversational data augmentation with generative conversational networks. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 26–38, Edinburgh, UK. Association for Computational Linguistics.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. Lad: Language models as data for zero-shot dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 595–604.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language*

Resources and Evaluation Conference, pages 3174–3182, Marseille, France. European Language Resources Association.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15000–15025, Toronto, Canada. Association for Computational Linguistics.

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–64.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Kai Sun, Seungwhan Moon, Paul A Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2022a. QAConv: Question answering on informative conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5389–5411, Dublin, Ireland. Association for Computational Linguistics.

Qingyang Wu, Song Feng, Derek Chen, Sachindra Joshi, Luis Lastras, and Zhou Yu. 2022b. DG2: Data augmentation through document grounded dialogue generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 204–216, Edinburgh, UK. Association for Computational Linguistics.

Qinglin Zhang, Chong Deng, Jiaqing Liu, Hai Yu, Qian Chen, Wen Wang, Zhijie Yan, Jinglin Liu, Yi Ren, and Zhou Zhao. 2023. Mug: A general meeting understanding and generation benchmark.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2022. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *Annual Meeting of the Association for Computational Linguistics*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *CoRR*, abs/2104.05938.

A Prompts additional information

Our prompt structure consists of generic instructions that provide guidance to the LLM, helping it formulate user queries based on the provided meeting transcript and the conversational history (Table 9), followed by the user query template to use in the specific turn (Table 10).

Adapted from QMSum, we include both *General* queries for overall meeting themes and key takeaways, and *Specific* queries which focus on particular topics or individuals discussed. We expanded the QMSum schema based on a preliminary user study, which allowed volunteers to upload their meeting recordings and inquire a prototype agent model about them. Analyzing this interaction data, we added an *Unanswerable* and *Context-dependent* query templates.

Unlike human annotators who can create tailored queries for a given transcript, our model is guided to always generate queries from any prompt. Thus, we adapt the schema towards more generic prompts to ensure the model always produces relevant queries.

For example, QMSum authors suggest the following query in their schema: *"Why did A agree / disagree with B when discussing X?"*. This assumes a disagreement between speaker A and B around some topic X. This assumption will not necessarily be true for any given meeting. Therefore we changed the prompt to be: *"The question should be a rephrase of asking if anyone disagreed with <Speaker> about <Topic>"*.

B Annotation details

MISeD was annotated by 3 professional annotators, fluent in English, employed in a commercial organization which provides data annotation services. The training process included detailed instructions and a pilot session, over held-out data which included 525 dialog turns, following which we clarified misunderstandings raised in the process. MISeD instances were divided between the 3 annotators, so that each instance was annotated by a single annotator.

C Automatic Evaluation of Attribution Implementation

Automatic evaluation of attribution is a recent research area with no established methods in the literature. We follow the methodology proposed recently by (Gao et al., 2023) which applies an NLI model to approximate human entailment judgements, automatically computing recall (the proportion of response sentences fully entailed by the attributions) and precision (the proportion of attributions which contribute to the entailment of the response). While they utilized the True NLI model (Honovich et al., 2022), we use the newer NLI model TrueTeacher (Gekhman et al., 2023).

We compute precision and recall scores for attribution in the following way.

Recall: recall measures to what extent the generated text is entailed by the attribution. For each sentence s in the response S, and given an attribution set A (the concatenation of all transcript spans a provided by the model as attribution), we compute $NLI(A, s)$ as the TrueTeacher-computed score of A entailing s (either 1 or 0). We then average this score over all sentences (micro-average).

Precision: precision measures whether the attribution only includes citations needed to entail the response. We compute citation precision for each citation a of the attribution set A. The precision of a is either 0 or 1. We first determine if a is relevant or irrelevant with respect to every sentence s in the response S: the citation a should be considered irrelevant wrt s if it does not entail s by itself ($NLI(a, s) = 0$) and the overall NLI score of A vs s does not change removing a from A ($NLI(A \setminus \{a\}, s) = NLI(A, s)$). a has a precision of 1 if a is relevant for at least a sentence s which has recall=1 ($NLI(A, s) = 1$). We then average the precision of all attributions (micro-average).

As pointed out in (Gao et al., 2023), a limitation of this citation precision evaluation is that it cannot detect a citation that partially supports the statement, therefore resulting in lower scores compared to human evaluation. We see this effect in Table 5 which shows a gap between human-computed and automatically-computed attribution precision.

As also noted in (Gao et al., 2023), not all model responses require attributions. While in their settings responses not requiring attribution were rare and ignored, in the meeting domain we find this case to be more prevalent, particularly when the response indicates that no answer can be found in the transcript (see Section 4), an example being "*The meeting participants did not discuss specific ideas for the remote control's shape and size* .".

In the human evaluation we filter out, for each model, from attribution evaluation those responses for which it was judged (manually) that attribution is not needed (20% of all responses). Unlike the manual evaluation process, there is no available automatic method to filter out responses that do not need attribution, which remains a challenge for future research. Because the automatic method includes also responses which do not require attribution, we can see a gap between the automatic and manual scores.

Finally we note that while the attribution format is consistent for fine-tuned models, zero-shot models present some inconsistency in how they list attribution and thus they require more complex parsing logic.

User query full prompt

<Meeting transcript>

Instruction: Generate a multi turn dialog between a user and a chatbot about the provided meeting. Every turn the user asks a question about the meeting and the chatbot respond with an answer based on the meeting

<Dialog history>

<Query template>

Table 9: Our query prompt includes generic instructions that guide the LLM in formulating user queries, given the meeting transcript and dialog history. This is followed by a randomly chosen query template, from a pool of templates that are based on the QMSum schema and our edits and additions.

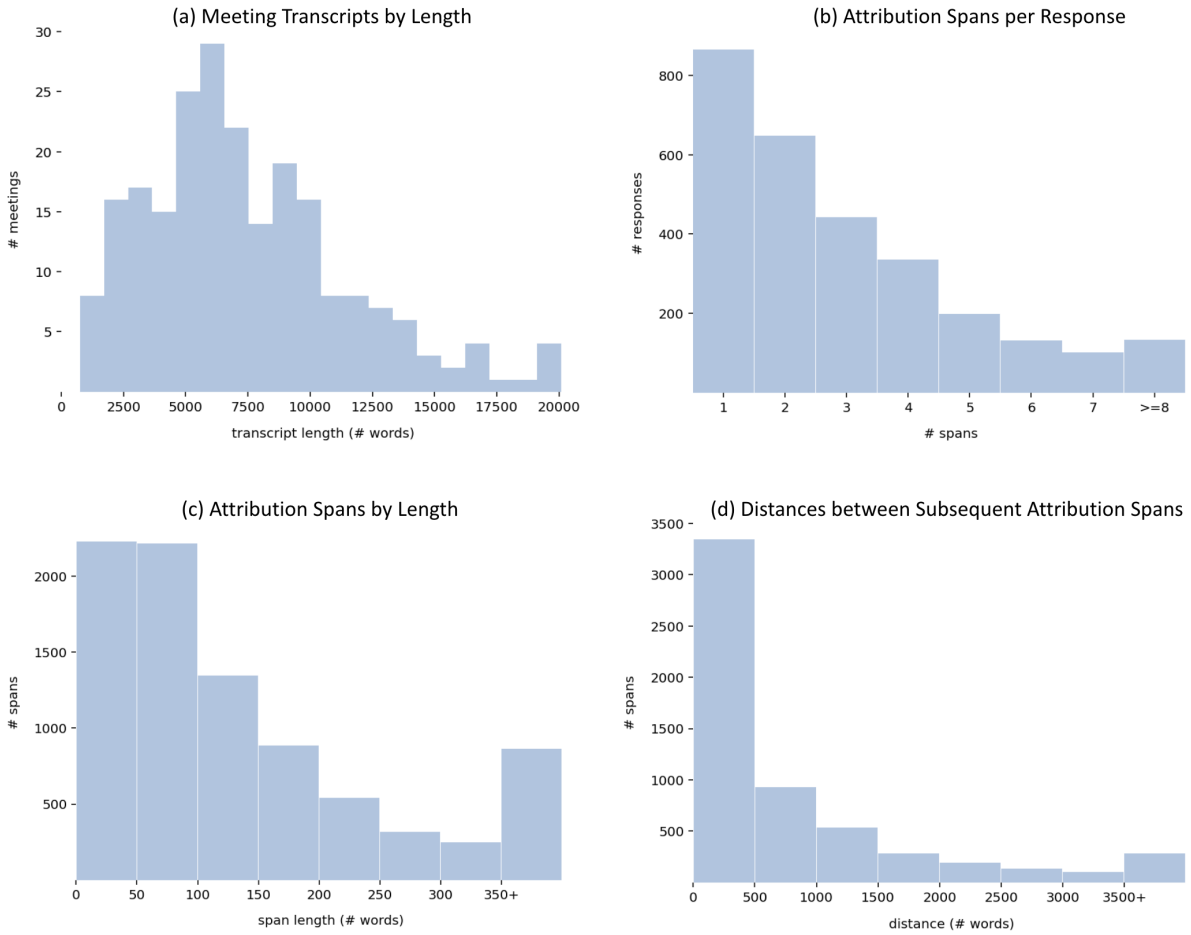

Figure 3: Transcript and attribution statistic. (a) Distribution of transcript length across the meetings used for MISeD. (b) Number of attribution spans in MISeD responses (among responses with attribution). (c) Distribution of attribution span length. (d) Distances between subsequent response attribution spans.

General query templates
1. The question should be a rephrase of asking for a summary of the meeting.
2. The question should be a rephrase of asking for a summary of the things <Speaker> said in the meeting.
3. The question should be a rephrase of asking what was the conclusion of the meeting.
4. The question should be a rephrase of asking what was the purpose of the meeting.
5. The question should be a rephrase of asking what were the action items of the meeting.
6. The question should be a rephrase of asking to identify questions raised during the meeting that were left unresolved.
Specific query templates
1. The question should be a rephrase of asking for a summary of <Topic>. <Topic> should be a topic that was discussed within the meeting.
2. The question should be a rephrase of asking why <Decision> was made.
3. The question should be a rephrase of asking what did <Speaker> say regarding <Topic> in the meeting.
4. The question should be a rephrase of asking what was the advantage of <Solution>.
5. The question should be a rephrase of asking why <Speaker> held an <Opinion>.
6. The question should be a rephrase of asking what was decided regarding <Topic>.
7. The question should should be a rephrase of asking if anyone disagreed with <Speaker> about <Topic>.
8. The question should be a rephrase of asking what did <Speaker> recommend to do when discussing <Topic>.
9. The question should be a Yes/No question where the answer is "Yes" based on the meeting.
10. The question should be a Yes/No question where the answer is "No" based on the meeting.
11. The question should be a Yes/No question where the answer can not be found within the meeting.
Unanswerable query templates
1. The question should be a rephrase of asking for a summary of <Topic>. <Topic> should be a topic that was never discussed within the meeting.
2. The question should be a rephrase of asking what <Speaker> said regarding to <Topic> in the meeting. <Topic> should be a topic that was never discussed by <Speaker> within the meeting.
3. The question should be a rephrase of asking what <Speaker> said regarding <Topic> in the meeting. <Speaker> should be a name of a person that did not participate in the meeting.
4. The question should be a rephrase of asking what was the advantage of <Solution>. <Solution> should be something that was never discussed within the meeting.
5. The question should be a rephrase of asking what was decided regarding <Topic>. <Topic> should be a topic that was never discussed within the meeting.
6. The question should be a rephrase of asking what was decided regarding <Topic>. <Topic> should be a topic that was discussed within the meeting but never got to any conclusion.
Context-dependent query templates
1. The question should be an implicit follow-up question regarding the previous message in the dialog using demonstrative pronouns such as "It", "He", "She", "They", "That". Example: "What did he say about it?", "What was their conclusion?"
2. The question should include a word with the same meaning as "else"/"other"/"besides".

Table 10: All query templates. The bracketed placeholders (e.g., <Topic>) signal where relevant meeting-specific topics should be inserted. Empirical evidence indicates that the LLM demonstrates a capacity to interpret these bracketed instructions, successfully generating contextually appropriate queries.

Agent response full prompt
<*Meeting transcript*>
<*Dialog history*>
<*User query*>
Instructions: The response should answer the user's question based on the meeting.
The response should follow the rules:

- The response should be always directly derived from the meeting. It should not include any opinion or fact that was not presented within the meeting.
- The response should take one of two forms:
  1. Up to 3 sentences of free text.
  2. Up to 2 intro sentence and between 3 and 5 bullet points. Mark each bullet point with a single asterisk ('*'). for example:
     "Here are the topics discussed in the meeting:
     * <Topic1>
     * <Topic2>
     * <Topic3>",
     "The participants raised few concerns regarding the timeline:
     * <Concern1>
     * <Concern2>
     * <Concern3>"
     Choose the form that best suit the answer.
- When referencing the participants of the meeting as a group, the response should refer to them as "The participants".
- When referencing to a single person, use gender neutral pronouns such as "They" or "Them".
- When referencing to the meeting, refer to it as "the meeting" and NOT "the meeting transcript".

Table 11: Full system's response prompts

<*Meeting transcript*>	T#0 Grad C said: Nice. T#1 Grad D said: OK. T#2 Grad A said: to to handle. T#3 Grad D said: Is that good? T#4 Grad C said: Right. Yeah, I've have never handled them. ...
<*Instructions*>	<Markers instructions> Generate the next response in the dialog between user and bot, adding a reference to the indices the answer comes from.
<*Dialog history*>	user: What did the meeting participants decide to do to move the project forward? bot: The participants decided to create a middle layer in their belief-net model...
<*User query*>	user: What did Grad C suggest to do when discussing the middle layer for the belief net model? bot:

Table 12: Prompt for LLM agents at training and inference time.

(T#734,T#759) Grad C suggested that values could be expanded...

Table 13: Example target format for fine-tuning. It includes the attribution spans and the expected response.

	Train	Validation	Test	Overall
# meetings	157	34	34	225
# dialogs	303	63	66	432
# query-response pairs	2922	611	628	4161
context-dependent	388 (13.28%)	81 (13.26%)	79 (12.58%)	548 (13.17%)
unanswerable	927 (31.72%)	149 (24.39%)	198 (31.53%)	1274 (30.62%)
query type: general	644 (22.04%)	88 (14.40%)	138 (21.97%)	870 (20.91%)
query type: specific	1515 (51.85%)	358 (58.27%)	308 (49.04%)	2179 (52.37%)
query type: yes/no	763 (26.11%)	167 (27.33%)	182 (28.98%)	1112 (26.72%)

Table 14: Statistics of the MISeD dataset by split.

	WOZ	MISeD Test Set
# meetings	35	34
# dialogs	70	66
# query-response pairs	700	628
overall vocabulary size	2012	2492
avg. response length	28.52	41.71
avg. response vocabulary size	23.18	30.53
avg. % of response vocabulary overlapping with the transcript	84.86%	78.43%

Table 15: Statistics of the WOZ test set compared to the MISeD test set. MISeD responses tend to be longer, and to use a broader vocabulary, overlapping less with the transcript vocabulary. Vocabulary size is calculated as the number of unique lemmas.

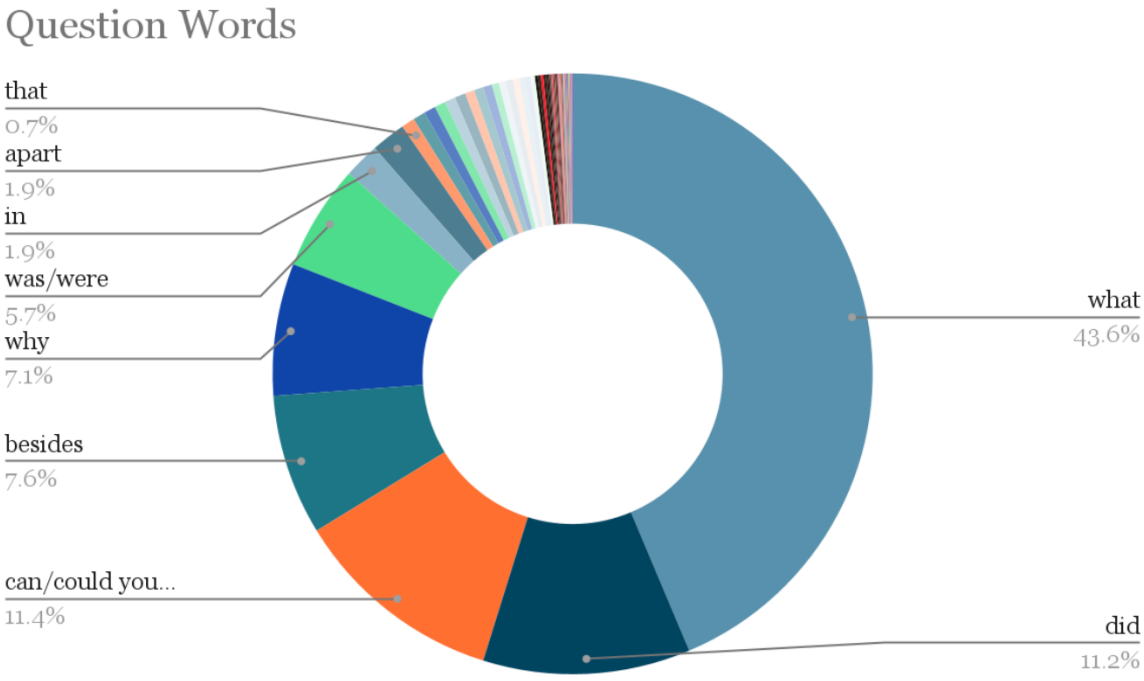

Figure 4: Distribution of question words (query prefix) within the dataset.

MISeD dialog example

query: "Can you give me a summary of the group's battery life concerns?"
answer: "The participants were concerned about the device's energy consumption, as it has a number of features that could quickly drain the battery. They discussed using both batteries and solar cells to power the device."
relevant text spans: 55, 68, 140-149, 177
metadata: query type: general

query: "And what devices could it support?"
answer: "The device is compatible with batteries and solar cells as a power source. The participants determined that the device could support an L_C_D_ screen, an infra-red unit, and speech recognition."
relevant text spans: 68, 140-141, 170-177, 266-270
metadata: query type: specific | context-dependent

query: "What was the advantage of using a push-button design on the remote?"
answer: "The advantage of using a push-button design on the remote was not discussed in the meeting."
metadata: query type: specific | unanswerable

(... continue of the dialog)

query: "Did the participants agree to use a banana as the base model for the device's design?"
answer: "Yes, the participants agreed to use a banana as the base model for the devicés design. They felt that it would be a good way to incorporate the fruit and vegetable motif that was suggested by the fashion team."
relevant text spans: 224-226
metadata: query type: yes-no

QMSum query-based summarization examples

General query:
query: "Summarize the whole meeting.",
answer: "The meeting was mainly about the conceptual design of the remote control. Firstly, marketing introduced market research and fashion trends around the world. While the user interface emphasized that easy to use was more important. (...) After that, they discussed the shape of the remote control and decided to make it like a banana. Based on this shape, they talked about the detailed design of the appearance and user interface."

Specific query:
query: "Summarize the discussion about the result of marketing research and fashion trends."
answer: "From the marketing's point of view, fancy was the highest priority, followed by technologically innovativeness and ease of use. The marketing expected the feel of the material to be spongy."
relevant text spans: 17, 57

(... more specific standalone queries)

Table 16: Example of a MISeD dialog vs. QMSum single-turn query-based summarization, over the same transcript. MISeD includes queries of additional types and characteristics, such as yes/no, unanswerable, and context-dependent queries.