Overview of the Second Shared Task on Automatic Minuting (AutoMin) at INLG 2023

Tirthankar Ghosal†, Ondřej Bojar*, Marie Hledíková*, Tom Kocmi‡, Anna Nedoluzhko*
†Oak Ridge National Laboratory, TN, USA
*Charles University, Faculty of Mathematics and Physics, ÚFAL, Czech Republic
‡Microsoft, Germany
†ghosalt@ornl.gov,*(bojar,marie.hledikova,nedoluzhko)@ufal.mff.cuni.cz

Abstract

In this article, we report the findings of the second shared task on **Automatic Minuting (AutoMin)** held as a Generation Challenge at the 16th International Natural Language Generation (INLG) Conference 2023. The second *Automatic Minuting* shared task is a successor to the first *AutoMin* which took place in 2021. The primary objective of the AutoMin shared task is to garner participation of the speech and natural language processing and generation community to create automatic methods for generating minutes from multi-party meetings. Five teams from diverse backgrounds participated in the shared task this year. A lot has changed in the Generative AI landscape since the last *AutoMin* especially with the emergence and wide adoption of Large Language Models (LLMs) to different downstream tasks. Most of the contributions are based on some form of an LLM and we are also adding current outputs of GPT-4 as a benchmark. Furthermore, we examine the applicability of GPT-4 for *automatic scoring* of minutes. Compared to the previous instance of AutoMin, we also add another domain, the minutes for EU Parliament sessions, and we experiment with a more fine-grained manual evaluation. More details on the event can be found at https://ufal.github.io/automin-2023/.

1 Introduction

Automatic Minuting (Shinde et al., 2022) refers to the task of automatically creating meeting minutes from multi-party meeting conversations. Since the pandemic, a significant portion of the global workforce, especially those in Information Technology (IT) and IT-enabled services, has gone virtual or preferring a hybrid mode of work.[1]

[1]96% of U.S. employees surveyed preferred to work on a hybrid basis as per https://www.forbes.com/sites/edwardsegal/2021/09/20/26-of-surveyed-employees-dont-plan-to-work-onsite-again-most-still-prefer-hybrid-arrangements/

Meetings have always been important to ensure smooth coordination and success of projects, but the proportion of sessions which combine remote and onsite workforce and collaboration between geographically distant members has grown manifold. Undeniably, frequent meetings create significant cognitive workload on people. To document the discussions, updates, tasks planned, *minuting* is an essential activity in meetings (be it online, in-person, or hybrid). Usually minutes are jotted down by some member(s) in the meeting but due to the lack of any fixed standards for minuting, different minute-takers may have different perspectives depending on their background. Despite that minutes of the same meeting by different authors may differ in certain aspects and content (Ghosal et al., 2022c), this activity can be automated to some extent.

There has been a body of research in this topic since the AMI (Mccowan et al., 2005), ICSI (Janin et al., 2003) and *Calo* Meeting Assistant (Voss and Ehlen, 2007) projects. Some interesting recent works on meeting and dialogue summarization include those by Zhu et al. (2020); Feng et al. (2021); Zhong et al. (2022); Prasad et al. (2023). We summarize our efforts on *Automatic Minuting* in Ghosal et al. (2022a); Singh et al. (2022, 2021). From the previous AutoMin shared task (Ghosal et al., 2021a), top-performing systems from Shinde et al. (2021); Yamaguchi et al. (2021) showed the usability of a BART-based (Lewis et al., 2020) system trained on SAMSum corpus (Gliwa et al., 2019) for the task. An exhaustive survey of abstractive meeting summarization research could be found in Rennard et al. (2022); Yang and Zhu (2023); Kumar and Kabiri (2022).

For long, resource creation for meeting summarization was difficult because of privacy reasons (AMI and ICSI were the only publicly available ones and later we introduced ELITR Minuting Corpus (Nedoluzhko et al., 2022)). However

Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges, pages 138–167

quite recently, we see few datasets are made available to support research in this topic, incl. e.g. Tardy et al. (2020); Zhong et al. (2021); Kim et al. (2023); Hu et al. (2023); Chen et al. (2021).

Recently, with the "generative AI revolution", pre-trained large language models (LLMs; Brown et al., 2020; Touvron et al., 2023; Devlin et al., 2019) have been employed for this task (Yang et al., 2023; Sándor, 2023; Chen et al., 2023), demonstrating amazing output quality. Based on the general public sentiment about the capabilities of LLMs, one could assume that *automatic minuting* belongs to one of the tasks that have suddenly become essentially solved. To verify the status in a rigorous way and to search for any open challenges that need to be addressed and also to assess how far we evolved since the last AutoMin (Ghosal et al., 2021b), we continued with the second iteration of the *AutoMin* shared task. A related effort along this direction was the DialogSum Generation Challenge (Chen et al., 2022; Bhattacharjee et al., 2022) at INLG 2022.

We proposed the second iteration of the *AutoMin* shared task as a *Generation Challenge* for INLG 2023 (Ghosal et al., 2022b). Essentially, with the current iteration of *AutoMin*, we wanted to find out:

- What are the current state-of-the-art approaches to minuting?
- What role LLMs play in these approaches; what benefits and risks they bring?
- Can we refine our manual evaluation of candidate minutes so that we have more reliable scoring techniques?
- What are the differences between different minuting domains? In addition to the same style of "project meetings" as used in AutoMin 2021, we included EU Parliament sessions in the task this year.

We describe our shared task and present our findings in the remainder of the paper.

2 Tasks Description

We offered four tasks (Task A, Task B, Task C, and Task D) to AutoMin participants (Ghosal et al., 2022b). Tasks A–C were known from the previous AutoMin instance, Task D was new and focused on evaluation of minutes.

In the end, all the teams decided to take part only in the first and most important task, namely minuting from diarized transcript (Task A). To compensate for the lack of participation in Task D, we experimented with automatic evaluation using LLMs, see Section 6.3.

2.1 Task A

*The **main task** consists of automatically generating minutes from multiparty meeting conversations* provided in the form of transcripts. The objective is to generate minutes as bulleted lists, summarizing the main contents of the meeting, as opposed to usual paragraph-like text summaries.

Task A was run in two domains. In English and Czech, we again relied on the meetings in the ELITR Minuting Corpus 1.0 (Nedoluzhko et al., 2022) but created a new test set for 2023 (meeting transcripts which were not previously published). For English, we added EuroParlMin 1.0,[2] a new resource we curated from the European parliamentary sessions, see Section 3 below.

Participants were free to submit their minutes for any selection of these test sets.

Note that the nature of meetings as well as the reference minutes are very different in the two datasets (technical project meetings vs. parliamentary sessions).

2.2 Task B

Given a pair of a meeting transcript and a manually-created minute, the task is to identify whether the minute belongs to the transcript.

During our data preparation from meetings on similar topics, we found that this task could be challenging due to the similarity of the discussed content and anchor points like named entities, e.g., in recurring meetings of the same project on the one hand, and the differences in the style of minuting, on the other hand. Another reason is that some minutes do not capture the central points in the meeting because the external scribes did not understand the context correctly and created minutes that miss significant issues discussed in the meeting or are simply too short.

2.3 Task C

Task C is a variation of Task B. *Given a pair of minutes, the task is to identify whether the two*

[2] https://github.com/ufal/europarlmin

	Lines	Words
ELITR Minuting Corpus		
Transcript (EN)	728.3 ± 389.9	7078.9 ± 3741.6
Minutes (EN)	45.8 ± 31.5	395.6 ± 388.5
Transcript (CS)	1198.7 ± 449.7	8748.6 ± 3023.2
Minutes (CS)	43.5 ± 26.9	277.9 ± 26.9
EuroParlMin		
Transcript (EN)	227.2 ± 257.0	8138.5 ± 10460.7
Minutes (EN)	48.6 ± 87.6	278.8 ± 534.2

Table 1: Summary across all data (training, development and test sets) used for AutoMin 2023. The figures correspond to mean±standard deviation.

minutes belong to the same meeting or to two different ones. This task is important as we want to uncover how minutes created by two different persons for the same meeting may differ in content and coverage.

2.4 Task D (New Task)

Given a meeting transcript, a candidate minute, and a set of one or more reference minutes, assign a score indicating the quality of the candidate minute.

The participating evaluation methods could focus on diverse aspects of minutes quality, such as the coverage of content discussed, the adequacy of the description, the readability, etc.

The original plan was to evaluate the submitted scores with respect to correlation with human judgements in terms of *adequacy*, *fluency* and *grammatical correctness* from AutoMin 2021 human evaluations, and possibly in terms of additional criteria.

3 Dataset Description

The datasets for AutoMin 2023 cover three types of data: project meetings in both English and Czech, as well as parliamentary sessions in English.

Basic statistics of the data are in Table 1.

The project meeting data was prepared from our own sources, while the parliamentary sessions were taken from the European Parliament mostly as-is, we merely selected data which was useful for our purposes.

3.1 Project Meeting Data

For the project meeting datasets, the partipants were advised to use ELITR Minuting Corpus (Nedoluzhko et al., 2022) as training data, with the option to use any other training data of their choice. We prepared new testsets for both languages, containing 12 new meetings for English and 11 new meetings for Czech. This new test set is again from technical project meetings and city planning meetings, the same as ELITR Minuting Corpus and the test set used in AutoMin 2021. The format is also identical.

	Train	Dev	Test-I	Test-II	Test-2023
ELMI EN	85	10	18	10	12
ELMI CS	33	10	10	6	11
EuroParl	2065	187	–	–	242

Table 2: Task A meeting counts used this year. ELMI stands for ELITR Minuting Corpus.

Table 2 shows our *train-dev-test* splits for Task A. Test-I and Test-II were made public already in 2021, Test-2023 reference minutes were not made available to anyone before the shared task was over.

The data preparation was completed in the following steps (same as in the past):

1. We obtained raw audio recordings of meetings and preliminary consent from their participants to process the data and publish it in a deidentified form.

2. The recordings were automatically transcribed using our ASR systems.

3. Our team of annotators was provided with the audio recordings and the automatic transcripts and was tasked with correcting the transcripts so as not to contain any mistakes. The next task was to break the transcript down into segments of similar length and to add speaker tags. The segments are approximately correspondent to sentences, although sentence boundaries are not always clearly defined in speech. Speaker tags are given at the beginning of each speaker's section in round brackets.

4. The same annotator who prepared the transcript was then asked to create reference minutes for the meeting.

5. One of the main concerns in our data preparation is privacy. Publicly released data must adhere to EU GDPR standards. Therefore in the next step, the annotators had to deidentify all personal identifiable information. This was mainly the names of per-

sons, projects and organizations. Each instance was replaced with a tag in the format of [PERSON*number*], [ORGANIZATION*number*] or [PROJECT*number*]. The speaker tags are deidentified as well, in the format (PERSON*number*). The numbers are consistent for each person, project and organization within one meeting, but are purposefully shuffled between different meetings, even if some of the participants are the same. Annotators also removed any potentially sensitive or offensive utterances completely and replaced them with a <censored/> tag.

6. The final step was to create alignments between the transcripts and their respective reference minutes. This is an annotation we were hoping would be useful especially during the evaluation process. The alignment maps utterances from the transcript onto the minutes line which summarizes them (if any), and/or assigns a remark for why the utterance is not suitable to be in minutes or otherwise problematic (e.g. small-talk or unintelligible). The utterances are aligned as to map whole conversations about a topic onto the appropriate minutes line, not just the part where the topic is introduced. Each utterance can be aligned to at most one line in the minutes. This is a simplification we chose for ease of annotation and processing, even though occasionally more lines would be more appropriate.

Generally, the whole process for a single meeting was carried out by the same annotator, but sometimes, the meeting had to be reassigned to a different annotator part-way through due to organizational complications.

A sample meeting summary in the ELITR Minuting Corpus style is provided in Appendix A.

3.2 Parliamentary Sessions

We created EuroParlMin 1.0 for the purposes of AutoMin 2023 and released it publicly at the beginning of the campaign.[3] EuroParlMin comes from the archives of the EU parliament.[4] We downloaded the transcripts and minutes and converted them to plaintext. Only very little text processing beyond dropping XML tags was done.

For the purposes of AutoMin, it was necessary to select sessions with desirable properties. Some of the sessions in the original dataset had little content. For some others, the provided minutes contained a large amount of extra text not related to the transcript. We therefore filtered for sessions with (a) sufficiently long transcripts, and (b) a good compression rate from transcript to minutes. We then split this selection into training, development and test sets (generally choosing the sessions best in (a) and (b) for the test set) and provided the training and development sets publicly. Privacy was not a concern, since the data is public, so there was no need for de-identification, making this our only data which had real names.

It is important to mention that the reference minutes in EuroParlMin are the original texts as provided by the parliament. As such, they often lack the actual content or more details about the decisions met in the meetings and only focus on formalities like aspects and voting.

Also note in Table 1 the big standard deviations in EuroParlMin data and the low line count in EuroParlMin transcripts, despite comparable word count to ELITR Minuting Corpus. This documents the domain difference where EuroParlMin contains long and often prepared speeches whereas ELITR Minuting Corpus is much more interactive.

4 Shared Task Timeline

The second AutoMin followed this timeline:

- ELITR Minuting Corpus Training Data Available: well before
- EuroParlMin Training Data Available: March 3, 2023
- Test Data Release: March 3, 2023
- System Output Submission Deadline: May 1, 2023
- System Report Due: May 15, 2023
- Review Notification: July 7, 2023
- Camera-Ready for Reports: July 25, 2023
- Event Date: September 11-15, 2023

Registered participants were invited to access our private Github repository to access the test sets.

[3] https://github.com/ufal/europarlmin

[4] https://emeeting.europarl.europa.eu/emeeting/committee/en/archives

System outputs as well as system reports were submitted by e-mail to the organizers.

5 Evaluated Systems

We evaluate submissions from the participants, baselines we obtained using LLMs and also the reference minutes.

Kindly refer to Appendix B for samples of the automatically created minutes.

5.1 Participating Teams

Of the 10 teams who registered for AutoMin, 5 teams eventually took part in the shared task. We had participating teams from academia as well as industry.

We briefly discuss the approaches of our participating teams (ordered alphabetically):

- Team **Darbarer** (Rousseau et al., 2023) adopted a modular strategy (four modules) for the automatic minuting task. According to them, since each participant in a meeting communicates differently than others, they first use a text simplification model, *mBarthez* by Kamal Eddine et al. (2021) to standardize the utterances in the conversation and compress the input to focus on informative content. In the next module, they first do linear segmentation of the transcript followed by using a BART-model (Lewis et al., 2020) trained on the XSum (Narayan et al., 2018) and SAMSum (Gliwa et al., 2019) datasets for summarization. In the next step, they propose a titling module to add a short description for each summarized block. Finally in the post-processing module, they employ some rule-based heuristics to improve the readability of the minutes. Overall, team *Darbarer* followed the similar steps as Shinde et al. (2021) and Yamaguchi et al. (2021) from the First *AutoMin* (Ghosal et al., 2021a), with an added pre-processing step of Text Simplication.

- Team **Synapse** (Klesnilová and Elizabeth, 2023) followed a similar approach: pre-process→segment→summarize→post-process. They perform brute-force segmentation of the transcripts (into pre-defined token-lengths) to manage the length of the long transcripts for the subsequent summarizer module. In the summarizer module, they experimented with variants of BART trained on several summarization datasets: XSum, AMI, SAMSum, DialogSum (Chen et al., 2021), and CNN/DM (Nallapati et al., 2016).

- Team **Iterate** (Kmječ and Bojar, 2023) adopted an iterative approach where their summarization model is fed with a chunk of a transcript together with several previously generated minute points to both satisfy the input length constraint of Transformer models while providing the needed context for the minutes. With their "iterative" motivation to imitate human way of taking notes in a meeting (jotting minutes while keeping in mind previous points), they experimented with several models: BART, Longformer (Beltagy et al., 2020), and the large language model Llama-based Vicuna (Chiang et al., 2023). They found that even without fine-tuning, Vicuna shows promise to generate coherent minutes from zero-shot prompting.

- Team **NTR**'s (Borisov and Mikhaylovskiy, 2023) minuting pipeline consists of two stages: segmentation and summarization. They perform semantic segmentation of the meeting transcripts to assist the subsequent transformer-based summarization stage to receive the input in the desired token length range. However, they found that their semantic segmentation approach does not perform better than the naive segmentation technique. In the summarization stage, they experiment with prompting a large language model Dolly (Conover et al., 2023) and found comparable performance with their mBART (Liu et al., 2020) + BERTopic (Grootendorst, 2022) method.

- Team **Zoom** (Schneider and Turchi, 2023) used Content Vector Segmentation (CVS) (Alemi and Ginsparg, 2015) to segment the meeting transcripts. They used GPT-3's *text-da-vinci* model to generate additional data for training. Finally, they employed a `bart_large` model trained on XSum and SAMSum datasets for summarizing the meeting conversations. The authors claim that CVS significantly improved the downstream minuting task as opposed to using length-based segmentation.

5.2 GPT-4 and GPT-3 Baselines

In addition to submissions by participating teams, we decided to also evaluate outputs obtained using large language models, as we thought this could yield useful comparison and insight as to the usability of LLMs in the meeting minuting task. We used both OpenAI's GPT-4 and text-davinci-003 (sometimes referred to as GPT-3).[5] See Appendix C for sample GPT-4 and GPT-3 outputs.

Our transcripts are significantly longer than the maximum allowed length of prompts, we therefore had to split each one into smaller sections, have the LLMs process these sections separately, and then simply concatenate the results. We have not used GPT4-32k for this task.

For English project meetings, the prompt we used was "Summarize the following project meeting in the form of 5 - 10 bullet points: <meeting transcript section>". This prompt yielded results in a format very close to what we were looking for.

Similarly, for European Parliament data, we used the prompt "Summarize the following meeting in the form of 5 - 10 bullet points: <meeting transcript section>".

For Czech project meetings, however, we had to make a data-specific adjustment to ensure that GPT would keep our deidentification tags intact (i.e. wouldn't translate e.g. "PERSON4" into Czech). The prompt we used was therefore "The following project meeting contains identifiers in the format PERSONnumber. Summarize the meeting in 5 - 10 bullet points in Czech, keeping the original identifiers: <meeting transcript section>".

6 Evaluation

In this section, we describe the evaluation process carried out for AutoMin 2023. We used human evaluations, traditional automatic evaluations as well as evaluations using LLMs.

6.1 Manual Evaluation

We had our annotators carry out two kinds of evaluation – one at the document level and another alignment-based one at hunk level using our ALIGNMEET tool (Polák et al., 2022). All human evaluation was done using Likert scales from 1 (worst) to 5 (best). The submissions were anonymized so that annotators did not know which team's submission they were scoring, however, due to some obvious differences in the format, they were probably able to associate minutes by the same team across meetings.

See Appendix E for the full annotation instructions.

Document level evaluation assigned the following four scores to each candidate minute:

1. **Adequacy** assesses if the minutes adequately capture the major topics discussed in the meeting, also considering coverage (all such topics covered).

2. **Fluency** reflects if the minutes consist of fluent, coherent texts and are readable to the evaluator.

3. **Grammatical Correctness** checks the level to which the minutes are grammatically correct.

4. **Relevance** signifies the extent to which the minutes overall capture the important content from the source transcript (as opposed to summarizing useless parts).

The hunk-level evaluation was carried out in two steps: alignment and evaluation. In the alignment step, the annotator constructed a transcript to candidate minute alignment, as described in Section 3. Then, four scores were assigned to each so-called hunk, i.e. a section of the transcript aligned to a single minute line.

1. **Adequacy** assesses if the item in the minute adequately captures the content of the respective aligned segment(s) of the transcript.

2. **Fluency** reflects if the item in the minute consists of fluent, coherent text and is readable to the evaluator.

	Dist. Ann.	Judged	Doc-l.	Hunk-l.
Pr. Data (EN)	3	9	81	2048
Pr. Data (CS)	2	4	16	920
EuroParlMin	1	6	36	98
Total	3	19	133	3066

Table 3: Numbers of distinct annotators, total judged meetings and individual judgements by dataset. Doc-level judgements are one per meeting-submission (or reference) pair. Hunk-level judgements are the overall number of all evaluated minutes items.

[5] https://platform.openai.com/docs/model-index-for-researchers

	Document-level	Hunk-level
Grammaticality	4.53±0.76	4.76±0.44
Relevance	4.19±0.69	4.40±0.74
Fluency	3.80±0.74	4.06±0.87
Adequacy	3.47±0.84	4.04±0.84

Table 4: Averages of scores that individual systems reached in the manual evaluation at document and hunk level. Sorted by decreasing overall score.

3. **Grammatical Correctness** checks the level to which the item in the minute is free of grammatical errors.

4. **Relevance** signifies whether the item in the minute is relevant to be included in the meeting summary (e.g. a perfect summary of small talk is not relevant).

Compared to AutoMin 2021, we added the *Relevance* score. Our hope was that the scores would better reflect the quality of the overall structure of the minute and topic selection. This turned out to not necessarily be the case in the actual scores, see below for further discussion.

Due to time constraints and the relative complexity of the evaluation, only a subset of each test set was manually evaluated, see Table 3. The most reliable part of our evaluation therefore lies in ELITR Minuting Corpus in English. The Czech part of ELITR Minuting Corpus has a rather low number of document-level judgements and the Europarl domain has a relatively low number of hunk-level judgements.

Kindly note that the human evaluation was *reference-less*. In other words, our evaluators had access to only the transcript of the meeting to evaluate the candidate minutes (participant submissions, our baselines, and also the reference itself). We did this on purpose to avoid the bias of human annotators towards the reference and also to have the chance to evaluate the reference in the same conditions as the other systems.

Manual evaluation serves as the official scoring for AutoMin 2023 but we highlight that we do not focus on the ranking of the systems but rather on the takeaways from the best as well as the relatively poorer system outputs.

6.1.1 Comments on Human Evaluations

As we can observe in Table 4 and in detail in Table 6 below, annotators gave relatively high scores to most minute outputs. Especially Fluency and Grammatical Correctness scores are high, the average of averaged scores is over 4 for most settings except document-level Adequacy (average of 3.47±0.84) and Fluency (3.80±0.74). This scoring reflects the fact that the system outputs look very natural and fluent, almost indistinguishable from the human minutes from the linguistic point of view. At the same time, the high superficial quality can pose a challenge to the annotators; their attention can decrease and errors can go unnoticed.

Although human evaluations have been provided by experienced annotators, a number of discrepancies were observed. Many of them are about 'incorrect' judgements or inattention, which is natural given to the content of the minutes outputs. Minute items generated by systems are sometimes very close to what had been discussed in the meeting but still do not reflect the actual content. For example, two participants discussed their computer science courses, the possibility of failing a course being mentioned several times. A system outputs a minute saying that "they both failed their courses". This is not true but annotators did not notice and estimated that this information was correct.

Furthermore, we often observe that automatically generated minutes do not 'have a good sense' of the relevance of the discussion parts, which may also remain unnoticed by annotators. This may be exemplified on the same meeting topic (computer science courses) which had been summarized in much redundant detail by one of the systems, and evaluated as fully correct by the annotator. In reality, this discussion is not relevant for the meeting at all, see Figure 1.

6.2 Automatic Evaluation using Standard Metrics

For our automatic evaluation of Task A, we relied on the widely popular text summarization metric ROUGE (Lin, 2004) in its three variants: ROUGE-1, ROUGE-2, ROUGE-L, and we also added BART and BERT-based evaluations.

6.2.1 ROUGE Variants

ROUGE metrics are based on n-gram similarities with a given reference. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It works by comparing an automatically produced summary against a reference summary (usually

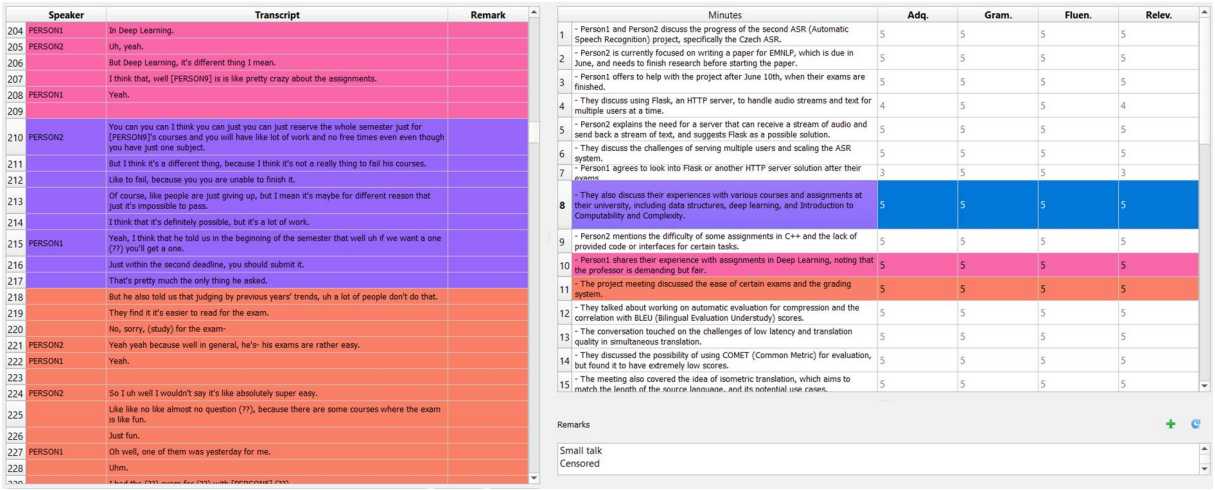

Figure 1: ALIGNMEET Evaluation interface with hunk-level scores. The left-hand side contains the transcript. The right-hand one contains one of the candidate minutes with each minute item asking for the four manual scores. (The document-level scores were reported at the end of candidate minutes.) The color background indicates the alignment between portions of the transcript and lines in minutes.

generated by a human). Different references thus inevitably lead to different ROUGE scores against each of them.

Recall in the context of ROUGE reflects how much of the reference summary the candidate summary is recovering or capturing:

$$\text{ROUGE}_{\text{Recall}} = \frac{\text{\# Overlapping n-grams}}{\text{Total n-grams in Reference Summary}} \quad (1)$$

Precision in the context of ROUGE means how much of the candidate summary was in fact relevant or needed:

$$\text{ROUGE}_{\text{Precision}} = \frac{\text{\# Overlapping n-grams}}{\text{Total n-grams in Candidate Summary}} \quad (2)$$

Despite the name ("Recall-Oriented..."), ROUGE actually commonly combines recall and precision using the harmonic mean to F-score. In our evaluation, we use ROUGE F1 scores for all ROUGE variants.

ROUGE-1 refers to the overlap of unigrams, ROUGE-2 is the overlap of bigrams, and ROUGE-L measures the longest matching sequence of words using Longest Common Subsequence (LCS).

As we mentioned earlier, proper evaluation metrics for meeting summarization are severely needed (Ghosal et al., 2021c), and text summarization metrics like ROUGE are only a poor alternative.

6.2.2 BERTScore

BERTScore (Zhang* et al., 2020) is an embedding-based metric that uses cosine similarity to compare each token or n-gram in the generated output with the reference sentence. There are three components to BERTScore:

- **Recall**: Average cosine similarity between each token in the reference and its closest match in the generated output.
- **Precision**: Average cosine similarity between each token in the generated output and its nearest match in the reference.
- **F1**: Harmonic mean of recall and precision

BERTScore is useful because it can account for synonyms and paraphrasing. Simpler metrics like BLEU and ROUGE can't do this due to their reliance on exact matches.

We used this[6] implementation of BERTScore.

6.2.3 BARTScore

BARTScore (Yuan et al., 2021) is another popular Natural Language Generation (NLG) metric which uses a pre-trained sequence-to-sequence model (BART in this case). The authors conceptualized the metric as *evaluation of generated text as a text generation problem* itself. The general idea is that models trained to convert the generated text to/from a reference output or the source text will

[6] https://pypi.org/project/bert-score/

achieve higher scores when the generated text is better.

$$\text{BARTSCORE} = \sum_{t=1}^{m} \omega_t \log p(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}, \theta) \quad (3)$$

where y is the generated text and x is the reference text. We use the original implementation[7] from the authors in terms of F-score.

6.3 Automatic Evaluation with LLMs

In the pursuit of Task D, we leveraged the capabilities of large language models (LLMs), particularly GPT (OpenAI, 2023), to assess the quality of system-generated meeting minutes.

The evaluation was structured around several parameters, which included the adequacy, relevance, topicality and fluency of the minutes. We instructed LLMs to rate each category for each set of minutes on a scale of 0 to 10. The prompt could be summarized as: "Given a specific meeting transcript, rate adequacy, relevance, topicality and fluency of the following system-generated minutes." We tried two different prompt structures, differing in the exact formulations and also in the set of qualities to be scored. The prompt (i.e. effectively an "evaluation method") called "GPT-ART" reports Adequacy, Relevance and Topicality. The prompt called "GPT-AFGR" reports Adequacy, Fluency, Grammaticality and Relevance and is thus aimed to mimic our manual criteria. The GPT-AFGR specification was also constructed to mimic annotators' instructions (Appendix E). The full prompts are provided in Appendix F.

A significant challenge in this study was the limitation of GPT's context window in dealing with extensive conversations. We attempted to overcome this by employing GPT-4-32k, with a context window of 32k subwords. However, this model still struggled to cover the entire conversation of several transcripts, and hence a more sophisticated approach was needed.

To counter this issue, we designed a technique that initially eliminates conversational turns that lack meaningful information. First, we removed all instances of the utterance "eh" and other non-speech items that are provided in the curated transcripts (e.g. <unintelligible>). Secondly, we asked GPT to mark each conversational turn as either meaning-bearing or filler content. The latter category referred to those segments that were merely iterations and could be excluded without the loss of any information. You can see the prompt in the Appendix F. For this filtering task, we used GPT "text-davinci-003" model, mainly due to the throttling limitations.

After this initial filtration process, we kept only the speaker identification from the filler turns. This methodology substantially reduced the length of the transcripts and their respective summaries, enabling them to fit within the context window of GPT-4-32k. This provided an effective solution to our problem and allowed for LLM-based evaluation of the system-generated meeting minutes.

[7] https://github.com/neulab/BARTScore

	Lines	Words
ELMI EN		
Transcripts	733.1±294.7	7198.1±2669.1
Ref. Minutes	59.8±29.6	480.3±251.3
Darbarer	49.8±18.4	358.0±158.8
davinci-003	48.5±15.3	524.1±191.3
GPT-4	34.3±12.7	551.1±219.9
Team Iterate	12.0±4.9	411.7±161.2
NTR	60.8±24.6	952.8±394.6
Synapse	38.2±9.9	484.9±169.7
Zoom-long	8.4±2.1	615.0±154.9
Zoom-short	6.2±1.3	117.2±25.7
ELMI CS		
Transcripts	1164.9±292.8	9897.6±2395.5
Ref. Minutes	86.0±36.9	435.6±194.6
Darbarer	69.3±14.2	495.7±121.0
davinci-003	89.5±22.2	905.6±247.5
GPT-4	96.0±20.1	1166.3±262.7
EuroParlMin		
Transcripts	33.2±59.2	873.8±2068.7
Ref. Minutes	37.7±59.3	185.0±323.6
Darbarer	9.2±6.2	65.0±72.1
NTR	8.1±12.0	155.1±310.8
Synapse	13.0±10.4	123.6±194.5
davinci-003	6.7±2.4	111.0±51.9
GPT-4	5.0±0.6	83.7±26.3

Table 5: Basic properties of manual transcripts, reference minutes and all participating team submissions of test set meetings. We report the average±standard deviation values for the number of lines and words.

6.4 Basic Statistics

We report basic test set statistics in Table 5: the average number of lines and words in each transcript, and reference minutes, as well as for the participant submissions (candidate minutes). This provides a first useful comparison of the participant minutes with respect to the reference minutes and transcripts.

Compared to the last year, there are no extreme outliers in terms of the numbers of lines (typically corresponding to summary points) across the sub-

	D Adeq	D Fluency	D Gram	D Relev	H Adeq	H Fluency	H Gram	H Relev
ELMI CS								
GPT-4	**5.00±0.00**	**5.00±0.00**	**4.50±0.71**	**4.50±0.71**	**4.84±0.56**	**4.84±0.56**	**4.77±0.57**	**4.68±0.56**
Reference	4.00±1.41	4.50±0.71	3.50±2.12	**4.50±0.71**	**4.81±0.59**	**4.81±0.59**	**4.36±0.87**	**4.34±0.84**
davinci-003	3.50±0.71	4.00±0.00	≀**4.50±0.71**	**4.00±1.41**	**4.71±0.73**	**4.71±0.73**	≀**4.54±0.74**	≀**4.44±0.85**
Darbarer	2.50±0.71	3.00±0.00	2.00±0.00	3.00±0.00	3.33±1.20	3.33±1.20	3.13±1.22	2.91±1.20
ELMI EN								
Zoom-long	**4.61±0.49**	**4.72±0.38**	4.81±0.35	4.47±0.48	**4.72±0.44**	**4.78±0.50**	4.85±0.34	**4.60±0.47**
GPT-4	**4.58±0.66**	≀**4.78±0.34**	≀**5.00±0.00**	≀**5.00±0.00**	**4.69±0.75**	**4.67±0.86**	≀**4.93±0.51**	≀**4.93±0.51**
Team Iterate	4.06±1.01	4.31±0.70	4.89±0.18	4.58±0.53	**4.46±0.83**	**4.67±0.69**	≀**4.96±0.17**	**4.80±0.45**
Reference	3.97±0.63	4.11±0.78	≀4.97±0.08	≀4.72±0.44	≀**4.60±0.70**	**4.64±0.68**	≀**4.98±0.24**	≀**4.89±0.38**
davinci-003	3.78±0.75	3.94±0.83	4.94±0.12	4.47±0.74	**4.50±0.80**	**4.49±0.94**	**4.97±0.33**	**4.75±0.71**
Synapse	3.50±0.48	3.61±0.73	4.69±0.43	4.11±0.92	**4.38±0.90**	**4.42±0.91**	**4.92±0.38**	≀**4.78±0.62**
Zoom-short	3.25±0.73	≀3.64±0.45	≀4.97±0.08	≀4.69±0.66	4.02±0.94	4.19±0.91	≀**4.99±0.07**	≀**4.82±0.47**
Darbarer	3.14±0.60	3.64±0.42	4.92±0.18	4.67±0.67	≀4.03±1.11	4.17±1.11	**4.93±0.41**	**4.76±0.71**
NTR	2.94±0.99	3.00±1.07	4.58±0.68	3.44±1.02	4.01±1.27	3.93±1.35	4.84±0.50	4.35±1.15
EuroParlMin								
Synapse	**3.17±1.17**	**3.33±0.52**	**5.00±0.00**	4.17±0.98	3.43±1.07	3.43±1.07	**5.00±0.00**	4.36±0.78
NTR	**2.67±1.03**	**3.33±0.82**	4.50±1.22	3.50±1.05	≀3.45±1.26	3.27±1.32	4.86±0.64	3.50±1.06
Darbarer	**2.33±1.21**	≀**3.50±1.05**	≀**5.00±0.00**	≀**4.83±0.41**	≀**4.44±0.81**	≀**4.44±0.81**	≀**5.00±0.00**	≀4.94±0.25
Reference	**2.00±0.63**	2.17±0.75	4.17±0.75	2.50±1.38	2.30±1.12	2.20±1.10	4.60±0.77	2.33±1.35
davinci-003	–	–	–	–	2.00±0.00	2.00±0.00	≀**5.00±0.00**	≀**5.00±0.00**

Table 6: Manual evaluation results: "D" means document-level, "H" means hunk-level. We report the average ± standard deviation. Sorted by decreasing quality according to document-level adequacy. The symbol "≀" highlights a disruption in the ordering in the given column. Due to an annotator's error, davinci-003 did not have the document-level scores provided. The top score and all scores that fall within its std. dev. bounds are in **bold**.

missions, although the variance across teams remains high, e.g. with 60.8±24.6 lines by NTR compared to 6.2±1.3 in the Zoom short submission on the ELMI EN test set.

The longest lines are, as expected, produced by Zoom-long (only available for ELMI EN), with the average of 73.2 words per line, followed by Team Iterate (34.3 words per line) and Zoom-short (18.9 words/line). GPT-4 comes next with 16.1 words/line on the ELMI EN domain.

We find it noteworthy that the reference minutes across all the domains have the fewest words per line an average (8.0 for ELMI EN, 5.1 for ELMI CS and 4.9 for EP). Only the Darbarer submission is at this level of line conciseness, producing even a little shorter lines than the reference on ELMI EN (7.2 vs 8.0).

6.5 Manual Evaluation Results

Table 6 presents average scores that individual systems received in the document as well as hunk-level manual annotations in each of the examined test sets.

As we see from the standard deviations, some evaluation settings are not discerning enough and any conclusions drawn from such areas should thus be treated with a big caution. This concerns primarily hunk-level judgements in project meetings (ELMI CS as well as ELMI EN), and also doc-level EuroParlMin.

In the document-level evaluations of project meeting minutes, we see that GPT-4 and also Zoom-long summaries achieve top scores. In addition to these, also Team Iterate scored better than the human reference. A detailed analysis of this result is desirable, to identify in which stage the human processing was sloppier than the automatic summary.

For the EuroParlMin, we attribute the worse score of human reference to the different style, not really appreciated by our evaluators. A similar situation is probably the case of Zoom-long vs. Zoom-short, where the annotators clearly prefer longer minutes.

The dominance of GPT-4 is apparent in terms of both phrasing (Fluency and Grammaticality) as well as content (Adequacy and Relevance) measures.

6.6 Automatic Evaluation Results

For automatic evaluation, we took the usual text summarization metric ROUGE (Lin, 2004) in its three variants (ROUGE-1, ROUGE-2, and ROUGE-L).

	ROUGE-1	ROUGE-2	ROUGE-L	BART-F1	BERT-F1
ELMI CS					
davinci-003	**0.33±0.16**	**0.09±0.04**	**0.13±0.06**	**-4.34±0.45**	**0.58±0.07**
Darbarer	**0.31±0.12**	**0.05±0.03**	**0.12±0.04**	**-4.55±0.38**	≀**0.59±0.03**
GPT-4	**0.30±0.15**	≀**0.08±0.05**	**0.12±0.06**	≀**-4.33±0.43**	**0.58±0.05**
ELMI EN					
GPT-4	**0.44±0.07**	**0.10±0.04**	**0.20±0.03**	**-4.40±0.42**	**0.60±0.03**
Synapse	**0.43±0.06**	≀**0.11±0.04**	**0.20±0.02**	**-4.56±0.42**	**0.59±0.03**
text-davinci-003	**0.41±0.07**	**0.10±0.02**	**0.19±0.02**	**-4.58±0.40**	0.55±0.01
Zoom-long	**0.41±0.09**	**0.10±0.03**	**0.18±0.02**	**-4.58±0.41**	0.55±0.02
davinci-003	**0.40±0.07**	**0.09±0.03**	**0.17±0.03**	≀**-4.45±0.42**	≀**0.58±0.02**
Team Iterate	**0.40±0.08**	≀**0.10±0.03**	≀**0.19±0.03**	**-4.63±0.42**	0.55±0.02
Darbarer	**0.39±0.06**	**0.10±0.03**	**0.19±0.03**	**-4.68±0.35**	≀**0.59±0.02**
NTR	**0.37±0.10**	**0.09±0.04**	0.16±0.03	≀**-4.66±0.49**	0.55±0.03
Zoom-short	0.29±0.08	0.06±0.02	0.15±0.04	**-4.82±0.37**	0.52±0.02
EuroParlMin					
Darbarer	**0.27±0.10**	**0.11±0.08**	**0.18±0.08**	-5.06±0.39	0.38±0.03
NTR	**0.27±0.11**	**0.09±0.07**	**0.17±0.07**	-5.23±0.39	0.29±0.02
Synapse	**0.26±0.10**	**0.08±0.07**	**0.16±0.08**	≀**-4.67±0.35**	≀**0.42±0.03**
davinci-003	**0.21±0.09**	**0.04±0.05**	**0.14±0.06**	-5.18±0.40	0.28±0.02
GPT-4	**0.20±0.09**	**0.04±0.05**	**0.13±0.06**	-5.22±0.40	≀0.29±0.03

Table 7: Automatic evaluation results using ROUGE, BART and BERT. We report the average ± standard deviation. Sorted by decreasing quality according to ROUGE-1. The symbol "≀" highlights a disruption in the ordering in the given column. The top score and all scores that fall within its std. dev. bounds are in **bold**.

Table 7 summarizes the ROUGE, BERT and BART scores across all our test sets.

For Czech minutes and BERT and BART, we first translate both the minutes by all participants and also the reference minutes into English using Lindat Translation.[8]

Best scores are in bold, again with all other scores that fall within the std. dev. band of the best one.

The automatic analysis using standard measures suffers even more from low statistical power. For the next year, we should clearly substantially increase the test set size, to gather the weak signal more reliably.

davinci-003 and text-davinci-003 are two variants of GPT-3 output. One of them was obtained by us, as discussed in Section 5.2, the other one was provided by Zoom. We did not score these two variants with manual or expensive automatic methods, so we only have ROUGE, BART and BERT to assess the effect of different (uncontrolled) prompt for the task. The comparison of these two outputs is mixed across the measures, and we would not overestimate the true value of the BERT-F1 difference where our prompt seems to win.

Synapse outputs on EuroParlMin stand out in BART and BERT scoring as the only system across the board. Rather likely, the pre- and post-processing heuristics were a good fit for the BART and BERT scoring.

6.7 LLM-Based Evaluation Results

Table 8 presents the scores that we obtained from our automatic scoring from GPT-4, as described in Section 6.3. We report the average and standard deviation across all the English meetings in the 2023 test set for ELITR Minuting Corpus. Note that not all systems were scored this way to fit the budget.

It is immediately apparent that GPT scorers prefer GPT produced outputs (GPT-4 and davinci-003), but GPT-4 scored exceptionally well also in the manual evaluation, so this cannot be taken as any bad sign.

What we see more as problematic is that there are only very few differences between the qualities that were supposed to be assessed. The rankings of Fluency, Grammaticality or Relevance according to GPT-AFGR are very much in line with its Adequacy (along which the table is sorted), and also the other prompt (GPT-ART) does not bring much difference. This is in some contrast with the manual document level judgements where Grammaticality and Relevance are not always in line with Adequacy.

[8] https://lindat.mff.cuni.cz/services/translation/

	GPT-AFRG				GPT-ART		
	Adeq.	Flu.	Gram.	Relev.	Adeq.	Relev.	Topic.
ELMI EN							
GPT-4	**8.75±0.45**	**8.83±0.39**	**9.00±0.00**	**8.75±0.45**	**8.17±0.39**	**9.00±0.00**	**8.75±0.45**
davinci-003	8.00±0.85	**8.58±0.67**	8.83±0.58	8.00±0.85	7.67±0.65	8.58±0.67	8.00±0.95
Zoom-long	7.83±0.39	8.42±0.51	8.75±0.45	7.83±0.39	7.50±0.67	8.17±0.58	7.50±0.67
Darbarer	7.58±0.67	≀**8.50±0.67**	≀8.83±0.39	7.58±0.67	7.08±0.79	7.92±0.67	7.17±0.94
Synapse	7.42±0.90	8.25±0.75	8.58±0.67	7.42±0.79	≀7.42±0.90	≀8.17±0.83	≀7.67±0.78
NTR	7.08±0.90	7.83±0.72	8.08±0.67	7.25±1.14	6.75±1.14	7.75±1.14	6.83±1.19
Team Iterate	6.58±1.38	7.67±0.98	≀8.17±0.72	6.75±1.22	6.75±1.06	7.67±0.98	6.83±1.11

Table 8: Automatic evaluation results using GPT with two types of prompt: AFGR and ART. We report the average ± standard deviation. Sorted by decreasing Adequacy estimated by GPT-AFRG The symbol "≀" highlights a disruption in the ordering in the given column. The top score and all scores that fall within its std. dev. bounds are in **bold**.

One striking outlier is Team Iterate which ended up third in manual evaluation but appears last according to GPT scoring.

7 Meta-Analysis of Automatic Evaluation

This section presents a comprehensive meta-analysis of the automatic evaluation in regard to human evaluation. The goal of this meta-analysis is to assess the usability of various automatic approaches, be it BART-score, BERT-score, variants of ROUGE, or GPT-based evaluation introduced in Section 6.3.

7.1 Pairwise Accuracy

Building on the methodologies used in the area of machine translation (Kocmi et al., 2021; Freitag et al., 2022), we use pairwise accuracy to explore how well automatic metrics align with human judgement. Pairwise accuracy is a simplification of Kendall's Tau.

In our setting, we focus on the system-level evaluation, where we first aggregate a score for each system (team submission) by averaging scores from all meetings. As the main unit, we use the difference in the score between a pair of systems:

$$\Delta = \text{score}(\text{System A}) - \text{score}(\text{System B}) \quad (4)$$

We define the pairwise accuracy as follows. For each system pair, we calculate the difference of the metric scores (metricΔ) and the difference in average human judgements (humanΔ). We calculate accuracy for a given metric as the number of rank agreements between metric and human deltas divided by the total number of comparisons:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|} \quad (5)$$

In other words, our pairwise accuracy reflects how often, across all pairwise comparisons, human ordering of the pair agrees with automatic ordering of the pair.

One of the advantages of pairwise accuracy resides in its interpretability. It demonstrates with what precision a given automatic evaluation can rank pairs of systems. The measure of accuracy is intuitively presented, where a 50% accuracy is equal to the flip of a coin. This provides insights into the potential improvements necessary in automated evaluation methods to make them correlate better with human evaluation, thus moving towards more reliable and accurate automatic minuting systems.

The results of the pairwise evaluation are in Table 9. The results suggest that ROUGE-1 is surprisingly the best performing technique when measuring Adequacy and Relevance (at both document and hunk level). For fluency and grammar, the ROUGE-L and BERTScore prevail. This result has a natural explanation in that ROUGE-1 is spotting certain individual words which are important for the meeting content. We speculate that due to the variance in possible reference summarizations, such a keyword match does not happen often, so the signal is weak and needs large test sets to be spotted, but it is there. ROUGE-2 and especially ROUGE-L measure overlap of longer sequences of words. Again, such a match can be rare, but if it happens, it reflects more some form of fluency rather than adequacy.

Contrary to expectations, GPT-based evalua-

	D Adeq	D Fluency	D Gram	D Relev	H Adeq	H Fluency	H Gram	H Relev
BART-score	66.7 %	47.6 %	42.9 %	61.9 %	85.7 %	71.4 %	76.2 %	71.4 %
BERT-score	57.1 %	76.2 %	71.4 %	61.9 %	66.7 %	90.5 %	76.2 %	52.4 %
GPT-AFGR A	61.9 %	61.9 %	57.1 %	66.7 %	81.0 %	76.2 %	71.4 %	66.7 %
GPT-AFGR F	52.4 %	66.7 %	61.9 %	61.9 %	71.4 %	81.0 %	71.4 %	57.1 %
GPT-AFGR G	52.4 %	71.4 %	66.7 %	66.7 %	71.4 %	85.7 %	71.4 %	57.1 %
GPT-AFGR R	61.9 %	61.9 %	57.1 %	66.7 %	81.0 %	76.2 %	71.4 %	66.7 %
ROUGE-1	85.7 %	66.7 %	61.9 %	81.0 %	85.7 %	61.9 %	66.7 %	81.0 %
ROUGE-2	76.2 %	66.7 %	61.9 %	81.0 %	76.2 %	61.9 %	57.1 %	71.4 %
ROUGE-L	61.9 %	81.0 %	76.2 %	66.7 %	61.9 %	76.2 %	61.9 %	57.1 %
GPT-ART A	66.7 %	57.1 %	52.4 %	61.9 %	85.7 %	71.4 %	66.7 %	71.4 %
GPT-ART R	61.9 %	52.4 %	47.6 %	57.1 %	81.0 %	71.4 %	66.7 %	66.7 %
GPT-ART T	61.9 %	52.4 %	47.6 %	57.1 %	81.0 %	76.2 %	71.4 %	66.7 %
GPT-ART avg	66.7 %	57.1 %	52.4 %	61.9 %	85.7 %	71.4 %	66.7 %	71.4 %
GPT-AFGR avg	57.1 %	61.9 %	57.1 %	61.9 %	76.2 %	76.2 %	71.4 %	61.9 %

Table 9: Pairwise accuracy, where each column represent one manual human evaluation attribute used as a gold standard (Doc and Hunk level scores for Adequacy, Fluency, Grammaticality and Relevance). Grey background highlights highest value for each manual setting. There are only 21 system pairs, meaning that 4.8% absolute difference is a single mislabeled system pair.

tion did not perform well, sometimes staying very close to the 50% coin-flip baseline. A possible explanation could be that GPT doesn't understand each category (adequacy, fluency, etc) the same way as humans, therefore we tried to aggregate them into a single score. However, neither this aggregated score performed well.

For the poor result of GPT-based evaluation in doc-level Fluency and Grammaticality, we do consider a possible problem with the annotation. With very good outputs in general and with non-native speakers, GPT may be actually more careful and better, thus departing from the human judgement.

The largest limitation of our pairwise accuracy assessment is the total number of system pairs, which is equal to 21. A possibility to increase the number of pairs would be to evaluate each minutes separately instead of aggregating them to system-level scores. We evaluated it, but ran into the issue of ties, where two minutes get the exact same score, are penalized under pairwise accuracy. This issue is not found in automatic scores such as BARTScore or ROUGE, which use continuous scale for ranking. However, GPT and humans use discrete scale of 5 or 10 points, which often result in score ties. This problem with pairwise accuracy and Kendall Tau was described earlier this year in Kocmi and Federmann (2023) and possible solutions are suggested in Deutsch et al. (2023).

7.2 Correlation between Automatic and Human Evaluation

Figure 3 presents Pearson correlations of each pair of manual and automatic metrics of minutes qual-

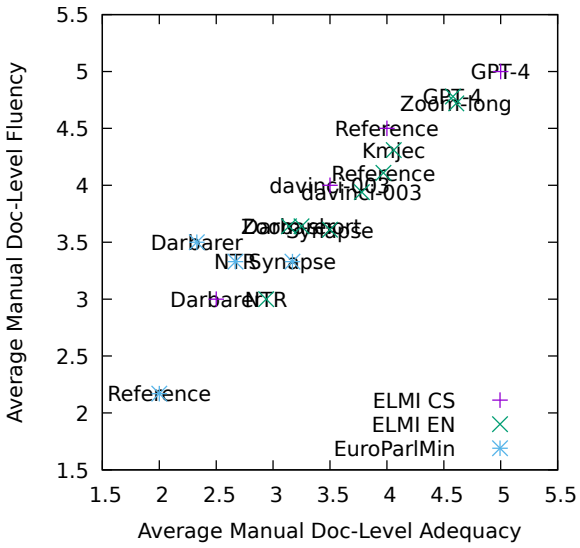

Figure 2: Average manual scores for Adequacy vs. Fluency at the document level in our three test sets

ities across all the datasets. Figures 4 to 6 provide the same information for ELITR Minuting Corpus EN, CS and for EuroParlMin, respectively.

It is important to keep in mind that these correlations are for system-level scores and that there are rather few systems in our collection. Consider the reported correlation of manual doc-level Adequacy vs. Fluency, see Figure 2. In total, there are 17 points, based on which the overall Pearson of 0.94 is calculated. Considering only the 4 EuroParlMin points, we observe a substantially softer correlation of 0.64. The test set with the most participating system, ELMI EN, on the other hand, shows a stronger correlation of 0.97. Pearson cor-

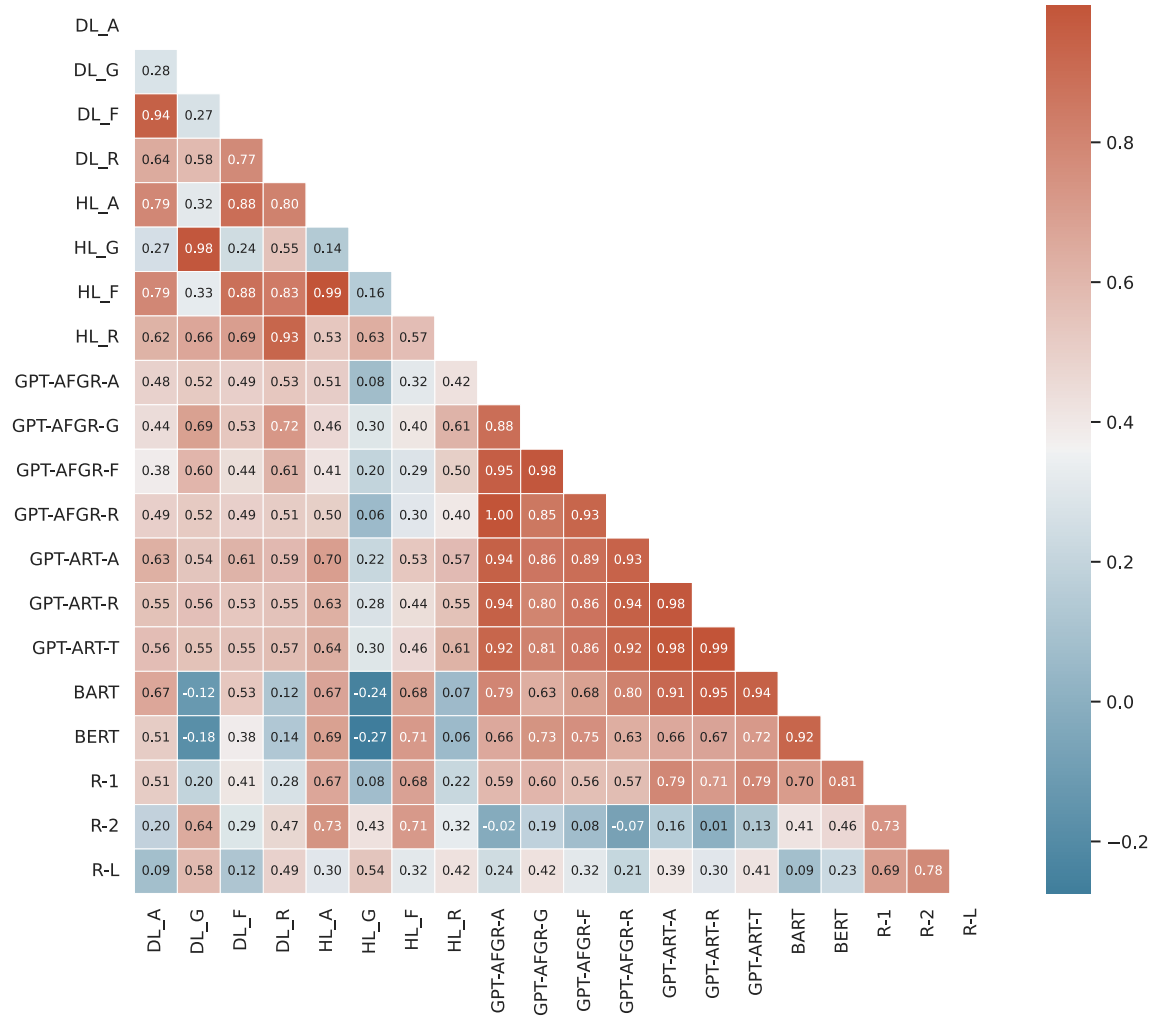

Figure 3: Heatmap showing the average correlation of the different metrics across the three different datasets in our shared task

relations are also very influenced by outliers and, in the case of shared tasks like ours, the underlying set of systems that took part in the task.

Let us discuss the correlations observed in the best covered (in terms of systems as well as manual judgements) test set of ELMI-EN (see Figure 4).

The most striking area is the red triangle of GPT-based metrics (rows and columns GPT-AFGR-G till GPT-ART-T). It shows that GPT essentially ignored the specific quality it should be evaluating (as described in the prompt).

We already mentioned the high correlation for manual Adequacy and Fluency and the "chequered pattern" in the upper left triangle confirms that this holds not just for doc-level but also for hunk-level scores. Relevance, on the other hand, seems to correlate well with Grammaticality on ELMI EN (Pearson of 0.95 for the doc-level scores, see the think crosses in Figure 7) but this result could be attributed also to the rather low discerning power of Grammaticality (most systems around 5 on the x axis) and the two systems (Synapse and NTR) setting the direction. Across all the test sets, Pearson is 0.58.

Pearson correlations also show that hunk-level scores are typically in line with their doc-level counterparts.

Looking at the lower rows of the heatmaps, BART and ROUGE-1 seem to correlate well with Adequacy at both document and hunk level (Figure 3), although this is not confirmed for BERT on the ELMI EN dataset (Figure 4). Relevance, on the other hand, seems very hard to predict for BART, BERT and also ROUGE, with Pearsons typically under 0.2.

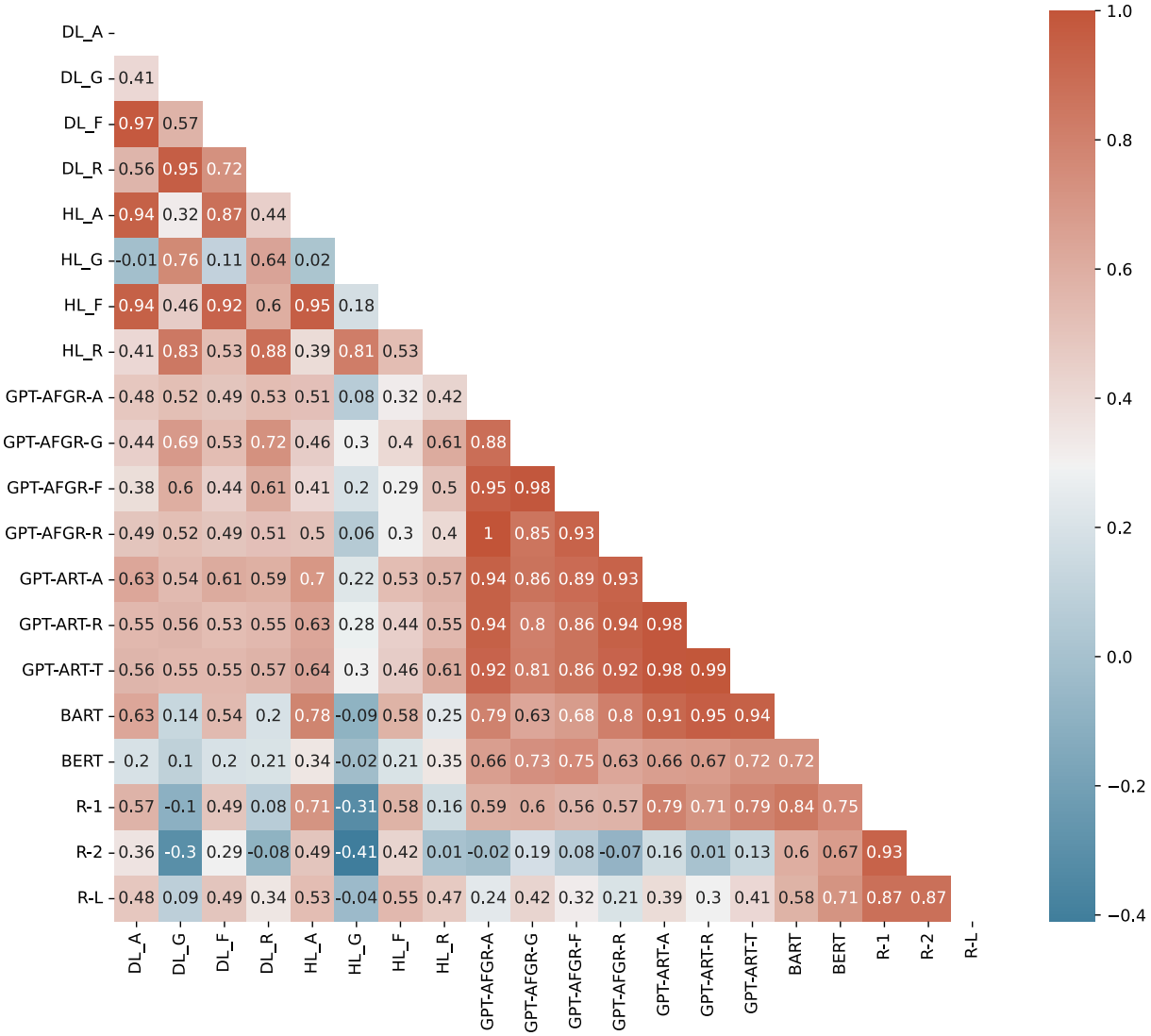

Figure 4: Heatmap showing the average correlation of the different metrics on the ELMI-EN dataset

8 Findings from 2nd AutoMin

A lot has changed since the previous instance of AutoMin in 2021. Last time, we were contemplating about one or two teams reaching perfect or close to perfect scores in Fluency and Grammaticality. This time, most of the submitted systems have this property.

We observed that the excellent superficial quality can, to some extent, hinder manual evaluation because errors can go unnoticed.

One of our attempts to improve manual evaluation was to introduce the division into hunks and hunk-level annotation. We have to conclude that this approach was not successful. On the positive side, the more fine-grained scoring provides more points of measurements. The critical drawback is, however, that the minutes get fragmented for the annotator. Assessed in such a partial isolation, more hunks seem flawless and the overall scores do not allow us to separate good vs. bad systems. It is conceivable that the hunk-level annotation would be informative for spotting problems, but its aggregate interpretation is problematic. For the next instance, we need to refocus such detailed manual annotation into spotting errors. The results, with some variance across the test sets, indicate that manual Fluency and Adequacy are strongly correlated. Relevance and Grammaticality differ. Please keep in mind the relatively low number of points behind this analysis. A good sign is that the hunk-level judgements correlate with the document-level ones.

We confirm that according to current manual measures, LLMs deliver excellent results. GPT-4, Zoom and also Team Iterate scored better than the human reference in terms of Adequacy of project

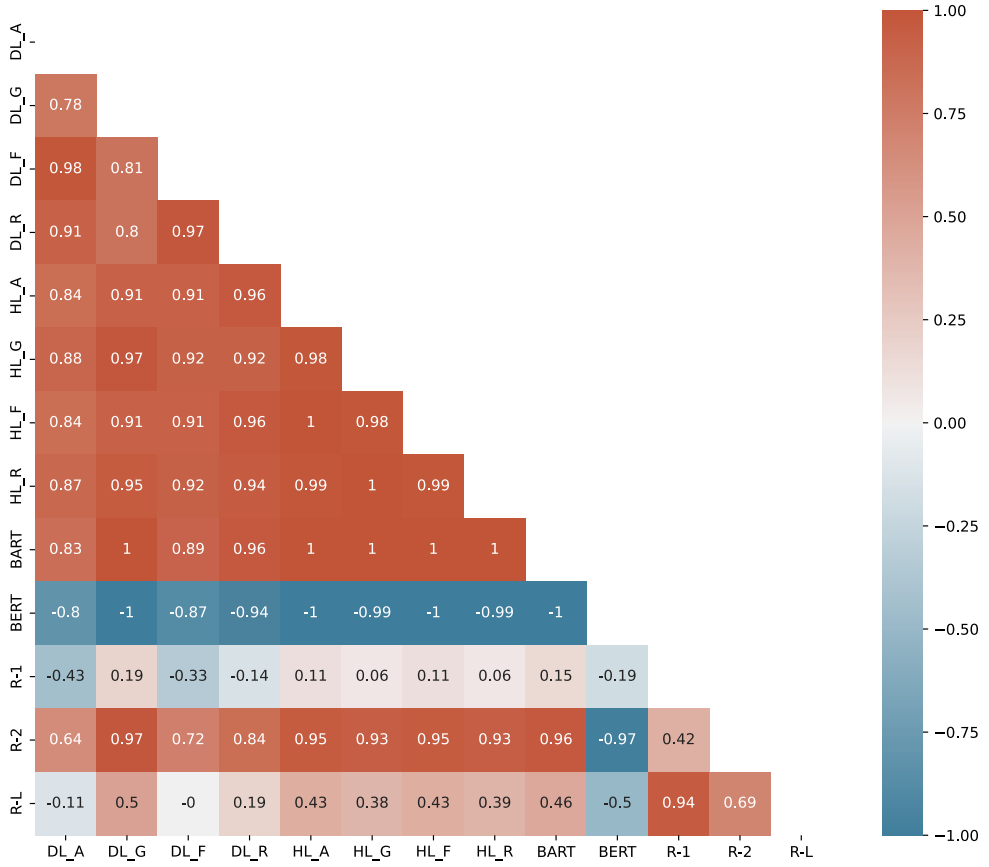

Figure 5: Heatmap showing the average correlation of the different metrics on the ELMI-CS dataset

meeting minutes.

In our meta-evaluation, we used Pearson correlation coefficients and pairwise accuracy to figure out which automatic metrics, including GPT prompting, are most similar to the different styles of our manual judgements. Our analysis revealed very interesting observations. GPT confidently prefers GPT-generated summaries and it is not really able to distinguish among the evaluation criteria. Its pairwise accuracy compared to human judgements reaches only 50–66%. Traditional metrics like ROUGE are more reliable, with ROUGE-1 obtaining 85.7% for predicting document-level Adequacy and ROUGE-L obtaining 81.0% for Fluency. We explain this by weak but reliable signal (infrequent but important keyword and key phrase matches).

9 Conclusions and Future Plans

This paper presented the results of AutoMin 2023, the second instance of our shared task on automatic summarization of meeting transcripts into meeting minutes. The data of the shared task (inputs, references, submissions, scores etc.) are available in this repository:

https://github.com/ufal/automin-2023-data

As in the previous instance, the task was run on two languages (English and Czech), with English receiving more attention from the participants.

The submissions were scored manually at the level of full minutes (document level) and also averaging the scores assigned to individual minutes points (hunk level). We concluded that the more fine-grained hunk-level annotation is less useful; the most important question we would like to learn from our annotators is whether the minutes properly reflect the content and overall impression of the meeting. With hunk-level annotation, the annotation process is fragmented and even if each of the fragments is of a high quality, the aggregation of these scores does not answer the key question.

This year, we added the domain of EuroParl sessions and observed that our preferred style of bulleted minutes is in sharp contrast with the officially released summaries. Our annotators liked our style better and the official references did not score well.

AutoMin 2023 also responded to the emergence of large language models, applying them both to

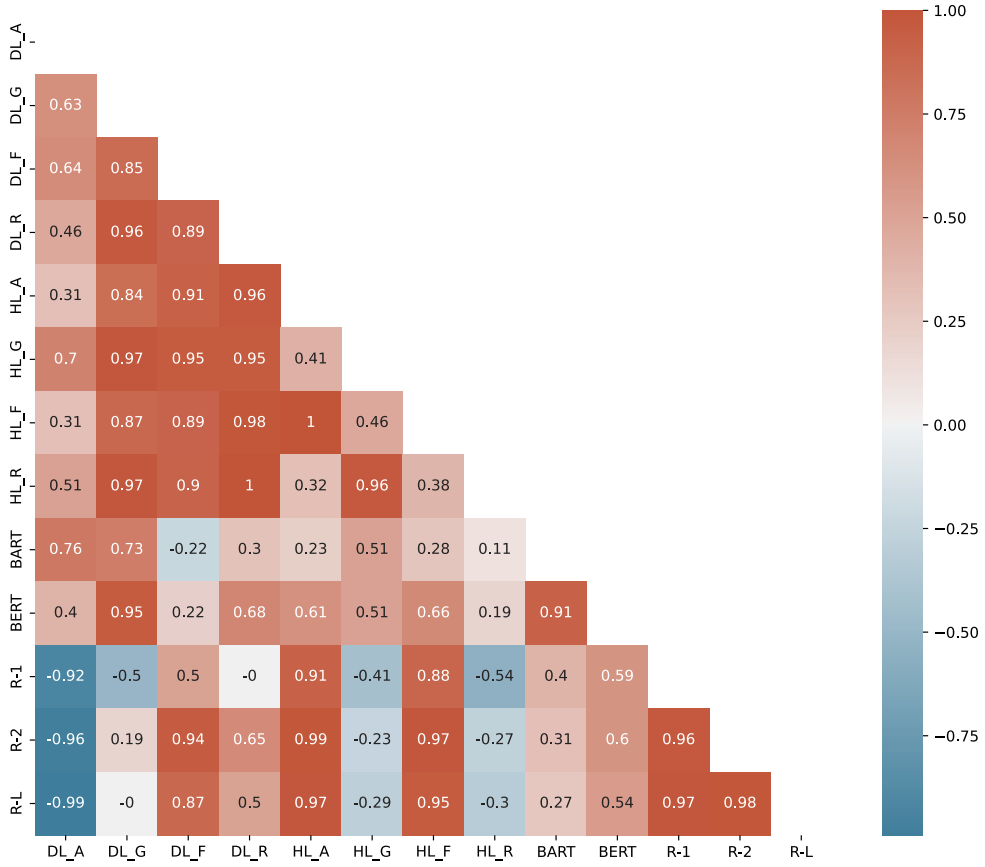

Figure 6: Heatmap showing the average correlation of the different metrics on the EuroParlMin dataset

the task of summarization itself, as well as to the task of assigning scores to the minutes we received from the participants. We confirmed that GPT-4 excels at summarization, surpassing the reference in manual document-level scoring. On the other hand, the automatic evaluation carried out by asking GPT-4 to evaluate the minutes is unreliable. GPT-4 does not distinguish between the different qualities (Adequacy, Fluency, Grammaticality, Relevance, Topicality) and reflects too much the grammaticality; it also prefers its own outputs.

A rather surprising result this year is that the very simple ROUGE is very much in line with manual ranking of system pairs. (ROUGE-1 for Adequacy and Relevance, ROUGE-L for Fluency and Grammaticality). The complex metrics did not provide any substantial benefit or performed simply worse.

We are again trying to secure funding for a future round of AutoMin, aiming primarily at a reliable assessment on whether the gist of the meeting has been well preserved in the minutes, and on a strategy for discovery and scrutiny of summarization errors.

10 Limitations

The main limitations behind AutoMin arise from the complexity of the task, which leads to demanding (and thus costly) annotation effort. We would prefer to have far more judgements, and we would have preferred to be able to run, e.g., multiple independent manual evaluations of the same meeting, in order to increase the discovery of errors, but our budget was limited.

Another serious limitation comes from the subjectivity of the minuting task as such. With so varied opinions on what is important in a meeting, it is difficult to assess minutes qualities reliably.

11 Acknowledgement

We would like to thank the participants for their enthusiastic participation in AutoMin and for bearing with our delays.

AutoMin 2023 was supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

12 Ethical Considerations

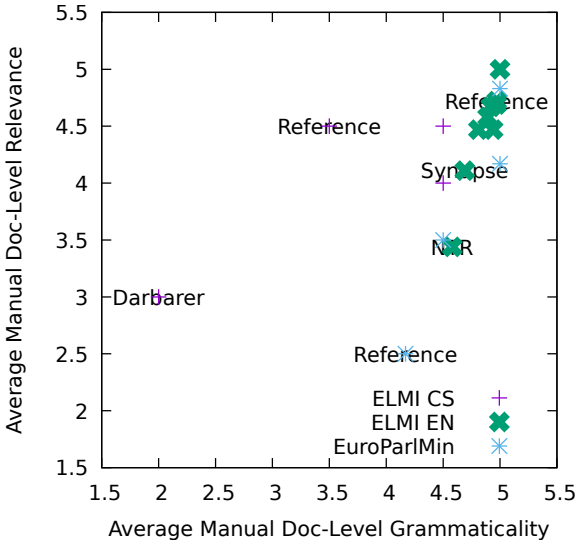

Figure 7: Average manual scores for Grammaticality vs. Relevance at the document level in our three test sets. System names shown only for the reference and a few selected systems.

We note that our annotators were real humans, and they carried their annotation tasks themselves, not delegating it to AI. The annotators were paid the standard hourly wage for this type of work in the Czech Republic.

References

Alexander A. Alemi and Paul H. Ginsparg. 2015. Text segmentation based on semantic word embeddings. *ArXiv*, abs/1503.05543.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Saprativa Bhattacharjee, Kartik Shinde, Tirthankar Ghosal, and Asif Ekbal. 2022. A multi-task learning approach for summarization of dialogues. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 110–120, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Eugene Borisov and Nikolay Mikhaylovskiy. 2023. Team ntr @ automin 2023: Dolly llm improves minuting performance, semantic segmentation doesn't. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yulong Chen, Naihao Deng, Yang Liu, and Yue Zhang. 2022. DialogSum challenge: Results of the dialogue summarization shared task. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 94–103, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Ruochen Xu, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2023. UniSumm and SummZoo: Unified model and diverse benchmark for few-shot summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12833–12855, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation. *arXiv preprint arXiv:2305.14324*.

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.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021a. Overview of the First Shared Task on Automatic Minuting (AutoMin) at Interspeech 2021. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–25.

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021b. Overview of the first shared task on automatic minuting (automin) at interspeech 2021. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–25.

Tirthankar Ghosal, Marie Hledíková, Ondřej Bojar, and Muskaan Singh. 2022a. Final report on summarization.

Tirthankar Ghosal, Marie Hledíková, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022b. The second automatic minuting (AutoMin) challenge: Generating and evaluating minutes from multi-party meetings. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 1–11, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. 2022c. Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial). In *ACM SIGIR Forum*, volume 55, pages 1–17. ACM New York, NY, USA.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. 2021c. Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial). In *ACM SIGIR Forum*, volume 55, pages 1–17. ACM New York, NY, USA.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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)*, pages 16409–16423, Toronto, Canada. 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. pages 364–367.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Kristýna Klesnilová and Michelle Elizabeth. 2023. Team synapse @ automin 2023: Leveraging bart-based models for automatic meeting minuting. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.

František Kmječ and Ondřej Bojar. 2023. Team iterate @ automin 2023 - experiments with iterative minuting. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *EAMT 2023*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Lakshmi Prasanna Kumar and Arman Kabiri. 2022. Meeting summarization: A survey of the state of the art. *arXiv preprint arXiv:2212.08206*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

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

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. 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.

OpenAI. 2023. Gpt-4 technical report.

Peter Polák, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022. ALIGNMEET: A comprehensive tool for meeting annotation, alignment, and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1771–1779, 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.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2022. Abstractive meeting summarization: A survey. *arXiv preprint arXiv:2208.04163*.

Ismaël Rousseau, Loïc Fosse, Youness Dkhissi, Géraldine Damnati, and Gwénolé Lecorvé. 2023. Team darbarer @ automin2023: Transcription simplification for concise minute generation from multi-party conversations. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.

Kristóf Sándor. 2023. Evaluating the effectiveness of large language models in meeting summarization with transcript segmentation techniques: How well does gpt-3.5-turbo perform on meeting summarization with topic and context-length window segmentation?

Felix Schneider and Marco Turchi. 2023. Team zoom @ automin 2023: Utilizing topic segmentation and llm data augmentation for long-form meeting summarization. In *Proceedings of the 16th International Conference on Natural Language Generation: Generation Challenges*. Association for Computational Linguistics.

Kartik Shinde, Nidhir Bhavsar, Aakash Bhatnagar, and Tirthankar Ghosal. 2021. Team ABC @ AutoMin 2021: Generating Readable Minutes with a BART-based Automatic Minuting Approach. In *Proc. First Shared Task on Automatic Minuting at Interspeech 2021*, pages 26–33.

Kartik Shinde, Tirthankar Ghosal, Muskaan Singh, and Ondřej Bojar. 2022. Automatic minuting: A pipeline method for generating minutes from multi-party meeting proceedings. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 691–702, Manila, Philippines. De La Salle University.

Muskaan Singh, Tirthankar Ghosal, and Ondřej Bojar. 2021. An empirical performance analysis of state-of-the-art summarization models for automatic minuting. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 50–60, Shanghai, China. Association for Computational Lingustics.

Muskaan Singh, Rishu Kumar, Tirthankar Ghosal, Ondřej Bojar, Chiara Canton PV, Andrea Sosi PV, Adelheid Glott AV, and Franz C Krüger AV. 2022. Demonstrator of automatic minuting.

Paul Tardy, David Janiszek, Yannick Estève, and Vincent Nguyen. 2020. Align then summarize: Automatic alignment methods for summarization corpus creation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6718–6724, Marseille, France. European Language Resources Association.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

L. Lynn Voss and Patrick Ehlen. 2007. The CALO meeting assistant. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 17–18, Rochester, New York, USA. Association for Computational Linguistics.

Atsuki Yamaguchi, Gaku Morio, Hiroaki Ozaki, Ken ichi Yokote, and Kenji Nagamatsu. 2021. Team hitachi @ automin 2021: Reference-free automatic minuting pipeline with argument structure construction over topic-based summarization. In *Proceedings of the First Shared Task on Automatic Minuting at Interspeech 2021*, pages 1–8.

Diyi Yang and Chenguang Zhu. 2023. Summarization of dialogues and conversations at scale. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 13–18, Dubrovnik, Croatia. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

A Sample Reference Minutes

```
Date: 2021-01-11
Attendees: [PERSON1], [PERSON2], [PERSON3], [PERSON4], [PERSON5]
Purpose of meeting: Progress report


- [PERSON1] and [PERSON2] discuss recent progress on ACL paper.

- [PERSON1] points out the automatic metrics seems not to be sufficient to evaluate the performance system (
    it performs too good).
- More challegning experiments are discussed by [PERSON1], [PERSON2] and [PERSON3].

- The available test set does not contain enough less common and rare words, there is a need for bilingual
    vocabulary and additional sources for system training.
  -[PERSON3] is requested to provide more data for Portugese language model
  -[PERSON7](not attending the meeting) is expected to deliver additional training data
  -Options to find bilingual texts are being discussed, namely to search for aricles and texbooks, check
    university repositories for master theses and additional sources of terminology words.

- [PERSON1] Proposed training of the system on different style of speech/writing (colloquial, male vs.
    female, scientific vs. non-scientific)
  The only available model of style transfer was regarding simplification or generalization of the text.
- Sub-part of elitr test set will be created to cover language style specifics, such as gender
    transformations.

- [PERSON4] showed how to search for named entities in the available bilingual dataset.
  Some manual work will be necessary to compare the outputs with those listed in there.
- This might give an overview what are the common problems with terminology and make a basis for markable(?)
     experiments
- Examples of related mistranslated words are mentioned (tenant vs lessee) and quality of writted audits is
    being discussed
- For further work, [PERSON2] will proceed with WMT elitr dataset for spoken language and (sao? domain for
    written texts
- Specific terms evaluation will be performed by [PERSON5] based on WMT testsuite markables.

- [PERSON1] briefly mention the opportunity to multi source, however other people involved in this are busy,
     thus compatibility of their implementation will be asssessed later

Minutes submited by: [ANNOTATOR1]
```

Figure 8: A sample minute taken by our external annotator

B Sample Minutes from AutoMin participants

We present some minute samples from our participants' submissions to show the variety of automatically generated minutes by different methods. For fair comparison we include the participant's generated minutes from the same meeting.

```
Date: 13.12.2022
- Attendees: PERSON2, PERSON3, PERSON1


- Replaning the satellite event
        - [PERSON1] and [PERSON2] are at the office
        - [PERSON3] is at a satellite event
        - They have to replan everything now.

- Preparation of the workshop
        - The proposal for a statistical workshop this year is over, so it would be too late for the next
            one in 2022
        - The participants have to submit a proposal for the workshop
        - They have to write a paper on the topic
        - The deadline for submitting the proposal is March fifteenth (!)
        - They need to prepare the event in three months.

- Preparation of shared task
        - [PERSON3] and [PERSON2] are preparing a shared task
        - They have to get an endorsement
        - They need to take membership to submit the proposal and pay some fees for their work- uh workshop.

- Editing the journal
        - [PERSON1], [PERSON2] and [PERSON3] are interested in publishing their work in a journal
        - The journal was published in 1964
        - It is not indexed on the website
        - The proceedings would be indexed with Google.

- Working on the satellite event
        - [PERSON2] and [PERSON3] will meet tomorrow to discuss the additional works they need to do for the
            satellite event.

- Submitted by: Darbarer
```

Figure 9: A sample minute by team Darbarer. This text was edited to replace some emoticons with hyphens to avoid encoding issues.

```
PERSON1 is at the office today, in the office. PERSON2 is waiting for the person who was supposed to join
    the group, but he didn't come. The person has returned the last e-mail.
(PERSON1) and (PERSON3) are going to make a shared task for a statistical workshop. The proposal for this
    year's workshop was made in September, which was too early for them. The next proposal for 2022 would
    be too late for them, so they have
PERSON1 will be back in two minutes.
PERSON3 and PERSON2 have to go to both workshop now. They need to take membership to submit the proposal.
    They have to make a website for their shared task and pay some fees for their work workshop.
PERSON1, PERSON2, PERSON3 and PERSON4 are planning to publish their research in a journal. The journal was
    published in 1964. The core papers were in red. It is indexed in Google Color. The proceedings would be
     indexed with ethylanthology.
PERSON2 and PERSON3 are looking for the conference proceedings from El Severe journal. They need to revise
    the proposal for the satellite event and then meet tomorrow to work on the additional works.
(PERSON2) and (PERSON3) have to do double work tomorrow.
```

Figure 10: A sample minute by team Iterate.

```
Meeting KeyPoints:


* PERSON2 invited PERSON1 to the meeting.
* PERSON1 accepted the invitation and attended the meeting.
* PERSON2 is aware that PERSON1 is in the office and is enjoying her walk.
* PERSON1 is enjoying the walk and deciding to enjoy it more by visiting an unfamiliar location.
* PERSON2 wanted to mention that PERSON1 is joining ORGANIZATION1 for
- There will be 3 papers for the satellite event
- The 3 papers should follow the same submission criteria as other papers
- If it's for a satellite event, the authors should follow the rules
- Authors should not submit the papers to the Interspeech proceedings
1. The last time a shared task was held at (sick dial) was in August of 2022.
2. The workshop for statistical theory in 2021 was held in September of 2022.
3. The workshop for statistical theory in 2022 would be too late for Thumbtack.
4. P3 preferred to apply for the special session for the workshop for statistical theory in 2021 instead of
   2022.
5
- PERSON1 wants to conduct a workshop, rather than a special session, for their shared task
- The workshop should be advertised as such - "Please take part in the shared task. We will provide you with
    the necessary tools and guidelines."
- PERSON1 wants to conduct a 3-month long programme, during which time they will provide the workshop
    participants with a pre-prepared programme, the workshop data and the workshop's results
- The deadline to submit the workshop proposal is March 15th, 2022
- In order to get an ISBD endorsement, the organizers need to submit a proposal to conduct a workshop at the
    interspeech conference by January 15th, 2022
- To submit the workshop proposal, the organizers will need to do the following steps:
  - Submit the proposal and get the endorsement from the speech community
  - Write the call for
- The person who will do the coding/data extraction is unsure. They think that they have to code/extract the
    data but they don't know how to do it. They need someone who knows the system and who will extract the
    data for them.
- Three people are going to the online workshop but they need to settle down first.
- The person who responded to the email also agreed to do the workshop.
-
So the meeting participants decided to seek advice from a senior colleague regarding the following:
*       Which aspects of the organizational structure are unclear to PERSON1
*       How much it will cost to register for the workshop
*       How to approach (german name) about financial aspects
*       What are the main tasks of the program committee
*       How to make a website for the shared task
1. P4 is the technical lead for the project.
2. P1 is the project manager.
4. P2 is aware of P3's interest and suggests that P4 can inform whether the journal is indexed.
5. P4 is able to provide some information regarding the journal
bblt?
it?
is not indexed..
keaboard, distant noise
Less work -
Less work and intent- more intensive output.
P
1. The proceedings from El Severe will be indexed in Cosmopile and Google Color
2. Also, the quality of papers published in the pvl journal is very good
3. The pvl journal is open access and is directly indexed by journals that offer open access to publications
1. PERSON3 and PERSON2 disagree on the feasibility of the proposed project.
2. PERSON3 will double the work required compared to the original proposal.
```

Figure 11: A sample minute by team NTR-NLP

```
DATE : 2023-05-01
ATTENDEES : PERSON1, PERSON2, PERSON3


SUMMARY
- PERSON1 is in the office today.
- PERSON2 is waiting for the person who was supposed to join the group.
- The person has not turned up yet.
- There was a technical communication meeting.
- The participants were told to prepare their papers for special sessions according to the same criteria as
    PERSON3, PERSON2 and PERSON1 are planning a mini conference.
- The proposal is going to be up there.
- It is entirely independent of interspeech to program committee.
- They will promote their workshop in their webpage.
- There is a possibility to do both sectile and inter PERSON1, PERSON2 and PERSON3 explain to each other how
    the process of organising a workshop works.
- PERSON3 and PERSON2 have prepared the document for the shared task.
- They have to take institutional membership to submit the proposal.
- They can do both.
- PERSON1, PERSON2, PERSON3 and PERSON4 are organizing a workshop on Interspeech.
- They have to study some satellite events that they got accepted to last year and make a website for their
    shared task.
- They need to pay for institutional membership, manage the project and PERSON2 has read a lot of papers on
    machine translation in her researcheration.
- The journal was published in 1964.
- The core papers were in this journal.
- It is already written in there.
- The proceedings would be indexed with ethylanthology.
- PERSON2 and PERSON3 are planning to present the severe conference proceedings from El Severe on January 15
    th.
- They are going to use pvl format.
- The paper is open access and indexed in Cosmopile Google color.
- PERSON3 wants to differ the discussion with PERSON4.
- They need to revise the proposal and leave down the thing that they need to do for the satellite event.
- They will meet tomorrow.


Minuted by: Team Synapse
```

Figure 12: A sample minute by team Synapse

```
PERSON1 and PERSON3 are discussing the proposal for a satellite event for a spoken language technology
    workshop.
They discuss the need to follow the same criteria as other special sessions and the timelines for the
    submission of papers.
They agree to submit two papers, one for the satellite event and one for a full fledged workshop.
PERSON2 offers to help with some potential things, but PERSON3 reminds him to do double work.
They decide to go for sigdial for the machine translation corpus linguistics and morphologies workshop in
    Texas, as the proceedings would be indexed with ethylanthology.
```

Figure 13: A sample minute by team Zoom - short version

PERSON2 and PERSON3 are discussing the meeting meeting at the ORGANIZATION1. PERSON1 is at the office and is waiting for PERSON3 to join, but he is not answering the call. PERSON3 has missed the call and PERSON1 will translate the last mail for them, which is a czech text. They have to replan everything now.
PERSON1 and PERSON3 are discussing the proposal for a satellite event for a spoken language technology workshop. They discuss the need to follow the same criteria as other special sessions and the timelines for the submission of papers. They also discuss the possibility of doing both a shared task at sick dial and submitting a proposal for the next conference. They agree that it is better to submit two papers, one for the satellite event and one for a full fledged workshop, as it gives them more flexibility and freedom to choose their event or accept papers from the Proceedings of Intra Speech. However, they also need to figure out how to make sure the participants are not deterred from submitting their own papers.
PERSON3 and PERSON1 are discussing the organizational points of their proposal for their workshop. PERSON3 explains that they have to take institutional membership to submit the proposal to the ORGANIZATION2 workshop application portal and that they need to contact Andre and Hero to discuss the proposal content. PERSON1 has no idea about the administrative plans for the workshop, but PERSON3 suggests asking PERSON4.
PERSON3 and PERSON1 are discussing a paper published in 1964 by noise-Linguistics. PERSON1 is not interested in all the points of the paper, but he has published it once or twice. PERSON2 has read a lot of papers about machine translation in his researcheration, but it depends on the participants.
PERSON2 and PERSON3 discuss the possibility of doing a machine translation corpus linguistics and morphologies workshop in Texas. They decide to go for sigdial, as the proceedings would be indexed with ethylanthology. PERSON1 suggests that sigdial is better than interspeech workshop, as it has less work and more intensive output. PERSON3 suggests that they should look at where the proceedings are indexed to see if there is any specific charge for the journals.
PERSON3 and PERSON2 discuss the proposal for the satellite event. PERSON2 offers to help with some potential things, but PERSON3 reminds him to do double work. They agree to meet tomorrow and say goodbye.

Figure 14: A sample minute by team Zoom - long version

C Sample GPT-Generated Minutes

```
- Meeting held to discuss organization of a shared task or workshop for [ORGANIZATION1]
- [PERSON1] attended the meeting from the office
- [PERSON3] joined the meeting late, but everything was fine
- Discussed the possibility of organizing a satellite event or special session at Interspeech
- [PERSON1] translated an email regarding the requirements and timelines for special sessions at Interspeech
- Discussed the possibility of submitting a shared task proposal to both Interspeech and SigDial
- Concerns raised about the workload involved in organizing both events and finding enough participants
- [PERSON3] mentioned that they have experience organizing such events and believes it is possible to manage
    both
- Discussed the need for institutional membership to submit a workshop application to [ORGANIZATION2]
- [PERSON2] mentioned receiving an email reply agreeing to some discussed points
- Discussed the need for membership to [ORGANIZATION4] to submit a proposal
- Considered paying fees for a workshop and contacting Andre and Hero for more information
- Discussed preparing content for the workshop proposal and studying other satellite events
- Talked about the need for management and creating a website for the shared task
- Discussed potential publication options, including [PROJECT2] Linguistics, archive, and [PROJECT1] from
    Elsevier
- Mentioned the possibility of publishing as part of the () community and deferring the discussion to [
    PERSON4]
- Agreed to revise the proposal and list additional tasks needed for the satellite event
- [PERSON1] offered to help with writing chapters for the proposal and reviewing it
- Agreed to meet again tomorrow to discuss further
```

Figure 15: A sample minute generated with GPT-4

```
- Project meeting was held between [PERSON1], [PERSON2], and [PERSON3]
- [PERSON1] was at the office and had decided to enjoy walking there
- [PERSON3] joined the meeting late
- Discussed the possibility of submitting a proposal for a special session or a satellite event at [
    ORGANIZATION1]
- [PERSON1] suggested submitting a shared task at [ORGANIZATION2]
- [PERSON3] proposed doing both the special session and the satellite event
- [PERSON3] suggested that the proposal for the workshop should be submitted as soon as possible and get
    endorsement
- [PERSON3] also suggested that the workshop should be entirely independent and the program committee,
    keynotes, and program should be ready before submitting the proposal
- Deadline for the satellite event is February 28th
- Deadline for submitting the full papers of participants is January 15th
- [PERSON3] suggested doing both the special session and the satellite event
- Need to contact [PERSON4] for membership to [ORGANIZATION4]
- Need to pay fees for workshop
- Need to discuss proposal content for workshop
- Need to create website for shared task
- Need to study satellite events accepted to Interspeech last year
- Need to do management and contact program committee
- Need to decide between SIGDIAL and Interspeech workshop
- SIGDIAL has more intensive output and indexed with Ethylanthology
- [PERSON4] can provide more information about [PROJECT2]-Linguistics
- [PERSON2] suggests [PROJECT1] from Elsevier
- [PERSON2] suggests [PROJECT2] from PVML
- [PERSON2] suggests [PROJECT3] from BBLT
- [PERSON2] suggests [PROJECT4] from Open Access Journals
- Need to discuss with [PERSON4] for better proposal
```

Figure 16: A sample minute generated with text-davinci-003

D Annotation Instructions

Instructions for Evaluation of Minutes

Install/update ALIGNMEET and Populate it with Minutes

- Open command line and run *pip install --upgrade alignmeet* if updating or *pip install alignmeet* if installing for the first time
- In case of issues talk to Marie Hledíková (*email removed*).

Find out which meetings are for you

- See this sheet *(link to sheet removed)*.
- Whenever you start or finish a meeting (all its minutes), please enter it in the appropriate cell.

Annotation Instructions

- **Do not forget to count how many hours you have spent annotating!**
- You are assigned a **set of meetings** (mostly English, some in Czech).
- Each meeting comes with:
    - the **transcript**
    - a set of **several minutes**, each created by a different system.
- You need provide the following annotations to **each of the minutes** (independently of other minutes):
    - **alignment** between the minutes and the transcript
    - **quality scores** for each alignment "hunk" (line in minutes)
    - **quality scores** for the minutes as a whole.
- You *may want* to also use the "Remarks" area (bottom right pane, it used to be called "Problems" in previous ALIGNMEET versions) for your convenience in annotation:
    - You may use the remark "Small talk" to indicate a portion of the transcript which you do not expect to appear in the minutes. However, if the system does include this in the minutes, you **need to primarily align it with the minutes**. (It is allowed to use both for any line in the transcript, to have a remark and be aligned to minutes.)
    - You may want to tell us some extra observations. For this use the remark "See separate comment" and write this comment to the table for assigning annotators.
    - The remarks you make will be copied over to the next minutes of the same transcript if the minutes have not been processed yet. You may switch between the different minutes of the same transcript as you like but as soon as the minutes have such a remark, no other remarks will be copied to them. Sequential processing of the minutes one by one is thus the best option.
- Detailed instructions:
    - **Alignment**:
        - Try to cover all items in the minutes and all text in the transcript but:

1. Not all parts of the transcript have to be aligned to an item in the minutes; e.g. if the system decided to exclude the given piece of information or topic altogether.
2. Occasionally, some items in the minutes also end up non-aligned; e.g. items in the minutes which are fully hallucinated will not be linked to any segment in the transcript.

        - Do try to make use of the new **autoalign feature:** in the top toolbar, set the threshold (the lower, the fewer alignments will be made) and click Autoalign. Alignmeet will suggest an alignment hint which will be displayed as color only under the speaker names. You will then need to manually confirm the alignments. It generally tends to help the most if you set a low threshold just to get a rough idea of where things are, the finer suggestions do not tend to be very good.
    - **Quality scores** are in the range 1 (worst) to 5 (best).
    - For individual "hunks" (i.e. the colored alignments) the scores should reflect:
        - **Adequacy** assesses if the item in the minute adequately captures the content of the respective aligned segment(s) of the transcript.
        - **Fluency** reflects if the item in the minute consists of fluent, coherent text and is readable to the evaluator.
        - **Grammatical Correctness** checks the level to which the item in the minute is free of errors in the grammar.
        - **Relevance** signifies whether the item in the minute is relevant to be included in the meeting summary (e.g. a perfect summary of small talk is not relevant).
    - For the whole meeting minutes, the scores should reflect:
        - **Adequacy** assesses if the minutes adequately capture the major topics discussed in the meeting, also considering coverage (all such topics covered).
        - **Fluency** reflects if the minutes consist of fluent, coherent texts and are readable to the evaluator.
        - **Grammatical Correctness** checks the level to which the minutes are grammatically correct.
        - **Relevance** signifies the extent to which the minutes overall capture the important content from the source transcript (as opposed to summarizing useless parts).
- Remark on **minutes styles**:
    - There are two major types of meetings in our collection, you may be able to notice this difference.
    - The style of minutes can however vary a lot depending on which system prepared the minutes.
    - The quality scores are designed so that they **should not be affected by the style** differences too much.
    - If you cannot avoid considering the style of the minutes in your scoring, **consider the style across all the different minutes** that are provided for the given meeting. (I.e. minutes departing seriously in their style from the rest may suffer slightly worse scores, but try to avoid this effect as much as possible.)

E GPT prompts

```
Given the following meeting transcipt and minutes, evaluate the minutes of the meeting for it's adequacy (
    the judgment if summary sentences represent conclusions clearly visible in the transcripts of the
    discussions), relevance (how well the summary sums up the main idea of the meeting), and topicality (
    whether summary sentences cover topics that are discussed in the transcript).
_____
Transcript:
{transcript}
_____
Minutes:
{system_generated_minutes}
_____
Evaluate minutes for it's adequacy (the judgment if summary sentences represent conclusions clearly visible
    in the transcripts of the discussions), relevance (how well the summary sums up the main idea of the
    meeting), and topicality (whether summary sentences cover topics that are discussed in the transcript).
     Give each score separately on a scale 0 to 10, where 10 is the best:
```

Figure 17: The prompt asking to rate each minutes for adequacy, relevance, and topicality. We label this prompt as "GPT-ART".

```
Given the following meeting transcript and minutes, evaluate the minutes for their adequacy (to what extent
    the minutes adequately capture the major topics discussed in the meeting, also considering coverage, i.
    e. all such topics covered), fluency (if the minutes consist of fluent, coherent texts and are readable
     to the evaluator), grammatical correctness (the level to which the minutes are grammatically correct)
    and relevance (the extent to which the minutes overall capture the important content from the source
    transcript (as opposed to summarizing useless parts).
_____
Transcript:
{transcript}
_____
Minutes:
{system_generated_minutes}
_____
Now evaluate the minutes for their adequacy, fluency, grammatical correctness and relevance. Give each score
     separately on a scale 0 to 10, where 10 is the best:
```

Figure 18: The prompt asking to rate each minutes for adequacy, relevance, and grammatical correctness. We label this prompt as "GPT-AFGR".

```
The following conversational turn is from a meeting transcript. Classify the turn into 'Filler' (not
    relevant outside of the meeting) or 'Content' (contains relevant information).
_____
{conversation_turn}
_____
Classification of the turn as 'Filler' or 'Content':
```

Figure 19: A prompt used to mark conversational tun as containing information or being a filler.