NUTSHELL: A Dataset for Abstract Generation from Scientific Talks

Maike Züfle[1], Sara Papi[2], Beatrice Savoldi[2], Marco Gaido[2], Luisa Bentivogli[2], Jan Niehues[1]

[1]Karlsruhe Institute of Technology, [2]Fondazione Bruno Kessler

{maike.zuefle,jan.niehues}@kit.edu, {spapi,bsavoldi,mgaido,bentivo}@fbk.eu

Abstract

Scientific communication is receiving increasing attention in natural language processing, especially to help researches access, summarize, and generate content. One emerging application in this area is Speech-to-Abstract Generation (SAG), which aims to automatically generate abstracts from recorded scientific presentations. SAG enables researchers to efficiently engage with conference talks, but progress has been limited by a lack of large-scale datasets. To address this gap, we introduce NUTSHELL, a novel multimodal dataset of *ACL conference talks paired with their corresponding abstracts. We establish strong baselines for SAG and evaluate the quality of generated abstracts using both automatic metrics and human judgments. Our results highlight the challenges of SAG and demonstrate the benefits of training on NUTSHELL. By releasing NUTSHELL under an open license (CC-BY 4.0), we aim to advance research in SAG and foster the development of improved models and evaluation methods.[1]

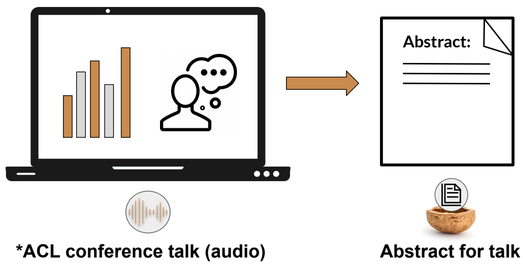


Figure 1: NUTSHELL, a dataset for Speech-to-Abstract Generation (SAG) from scientific talks.

1 Introduction

Abstracts are essential in scientific communication, allowing researchers to quickly grasp the key contributions of a paper. With the ever-growing number of publications, abstracts help researchers stay informed without reading full papers. Beyond their practical utility, abstracts also pose a significant challenge for natural language generation models: abstracts are a specialized form of summarization that not only condenses content but also promotes the work, often using domain-specific terminology and structured language.

Scientific summarization has been widely studied in natural language processing, including summarizing entire articles (Collins et al., 2017; Mao et al., 2022; Liu et al., 2024), particularly in the medical domain (Kedzie et al., 2018; Cohan et al., 2018; Gupta et al., 2021), generating abstracts from citations (Yasunaga et al., 2019; Zanzotto et al., 2020), summarizing specific paper sections (Takeshita et al., 2024), and leveraging knowledge graphs for abstract generation (Koncel-Kedziorski et al., 2019).

With the growing availability of recorded conference talks, a new challenge emerges: generating abstracts from spoken content or Speech-to-Abstract Generation (SAG). The abstracts offer researchers a quick way to assess relevant talks without watching entire recordings. Additionally, as conferences include more virtual content, automatically generated summaries enable efficient engagement with recorded talks (Murray et al., 2010).

While speech summarization has been explored in domains like news (Matsuura et al., 2024), YouTube videos (Sanabria et al., 2018), and meeting minutes (McCowan et al., 2005; Janin et al., 2003), large-scale datasets for scientific talk abstract generation are lacking. Existing work (Lev et al., 2019) aligns transcripts with the corresponding papers and extracts overlapping textual segments as summaries. However, these segments are drawn from the paper rather than the talk itself, failing to capture the distinct contributions, framing, and nuances conveyed in spoken presentations. Other studies have focused on summarizing TED

[1]https://huggingface.co/datasets/maikezu/nutshell

Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025), pages 19–32
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

Talks (Koto et al., 2014; Kano et al., 2021; Vico and Niehues, 2022; Shon et al., 2023), which target a broad audience and prioritize inspiration and engagement over technical content.

To bridge this gap, we introduce NUTSHELL a new multimodal dataset for abstract generation from scientific talks. Built from recorded presentations of *ACL conferences, the dataset pairs abstracts with their corresponding spoken content and video, offering a valuable resource for future research. To validate the quality of the abstracts as concise and well-structured summaries of the talks – i.e., capturing the essence of the presentations *in a nutshell* – we performed a human assessment, which confirmed their effectiveness and suitability for the SAG task.

To establish baselines for SAG using our dataset, we evaluate three model types: (1) a cascaded model combining automatic speech recognition (ASR) with text-based summarization, (2) a state-of-the-art speech-language model (SpeechLLM) without fine-tuning, and (3) a SpeechLLM fine-tuned on our dataset.

Our contributions are three-fold:

1. We introduce NUTSHELL, a novel dataset for abstract generation from scientific talks comprising 1,172 hours, which is released under CC-BY 4.0 License on HuggingFace;[1]
2. We provide baselines with different model types for comparison in future research, evaluated using both standard automatic metrics (e.g., ROUGE) and the emerging LLM-as-a-judge approach (Shen et al., 2023);
3. We conduct human evaluations to assess the quality of the abstracts and validate the suitability of automatic metrics for the SAG task.

2 The NUTSHELL Dataset

In this section, we introduce the new NUTSHELL resource. We chose to build our corpus upon the the ACL Anthology[2] since it provides a rich collection of multimodal resources (talks and abstracts) and open-access licensing. Starting from 2017, a significant number of papers published in the main *ACL conferences (ACL, EMNLP, and NAACL) include a video of the presentation, all released under the Creative Commons Attribution 4.0 license. This makes *ACL an ideal resource for building a multimodal dataset for the SAG task.

[2]https://aclanthology.org

In the following, we present a feasibility assessment of SAG through human evaluation (§2.1). Then, we describe the collection process performed to create NUTSHELL, together with the final dataset statistics (§2.2).

2.1 Are paper abstracts "good" talk summaries?

Before creating the corpus, we establish the validity of our data by investigating whether abstracts represent a good summary of the associated talk. To this aim, we conduct a qualitative check on a data sample of 30 talk-abstract pairs from the ACL Anthology. We involve a total of 5 annotators, who are all domain experts and thus familiar with scientific material.[3] To verify Inter-Annotator Agreement (IAA), a double annotation by different experts was carried out on 15 pairs.

Since we are interested in understanding whether paper abstracts are informative enough to represent a good summary of the talk, we asked evaluators to annotate: (1) Whether the information in the abstract is **all** uttered by the presenter in the talk; (2) The span of information present in the abstract that was not contained in the talk, if any; (3) Whether the abstract summarizes all **important** information presented in the talk. The human evaluation procedure, including the annotation template, is described in App. A.

The results indicate that 70.0% of the abstracts are considered good summaries by annotators as they contain important information about the talk. However, 63.3% of the abstracts also contain information not explicitly present in the talk itself. To better understand this, we conducted a qualitative analysis of the annotated spans corresponding to this missing information. We found that these spans typically involved dataset names, model names, shared task references (e.g., evaluation campaigns), or URLs (e.g., link to the resource or model being released). Notably, these elements are often displayed on slides but not explicitly verbalized by presenters.[4]

Despite this issue, the evaluation of automatic models against the same ground truth abstract can be considered fair, as models are equally penalized by this category of missing information. Moreover,

[3]Annotators include the paper authors and their colleagues.

[4]This issue could be overcome by exploiting the videos, as this information is typically shown in the slides. While out of scope for SAG, NUTSHELL includes the videos, making it a useful resource also for more complex multimodal tasks.

	conferences	year	# examples	total audio h	average audio min	average words per abstract
train	ACL,NAACL, EMNLP	2017-2021	4000	808.3	12.1 ± 11.2	142.8 ± 36.1
dev	ACL	2022	885	146.4	9.9 ± 3.6	141.9 ± 36.5
test	EMNLP, NAACL	2022	1431	217.1	9.1 ± 4.3	147.6 ± 37.4
total	ACL, NAACL, EMNLP	2017-2022	6316	1171.8	11.1 ± 9.9	143.7 ± 36.5

Table 1: Dataset statistics for NUTSHELL. The number of words is obtained by splitting the abstract at whitespaces.

it is worth noting that establishing a single ground truth for summarization tasks is still an open challenge (Zhang et al., 2024), given the inherent variability in human-produced summaries.

Both, questions (1) and (3) have an inter-annotator agreement of $\kappa = 0.466$, indicating moderate agreement (Landis and Koch, 1977), which can be regarded as acceptable given the subjective nature of evaluating summaries. While criterion (3) naturally involves subjective judgments about information importance, the lower agreement on criterion (1) can also be attributed to borderline cases, where small phrasing differences were sometimes overlooked by individual annotators. Such subtleties led to occasional discrepancies in annotator decisions, but were manually reviewed.

In summary, the manual evaluation confirmed both the feasibility of the SAG tasks and, despite the noted challenges, the overall reliability and usefulness of our resource.

2.2 Collection and Dataset Statistics

We collected talks from 16 ACL Anthology events: 6 ACL, 6 EMNLP, and 4 NAACL, including workshops, shared tasks and industry tracks. For each paper (both long and short format), we extracted the video and the associated abstract already available on the paper website. We exclude papers with invalid URLs, videos without audio, or abstracts missing from the paper page. Additional details on the data collection can be found in App. B.

Lastly, we split the dataset into training (years 2017 to 2021), dev (ACL 2022), and test (EMNLP/-NAACL 2022). These splits reflect a realistic evaluation setup, where models are trained on past data and tested on the most recent, unseen examples. In total, the corpus contains 1,172 hours of audio content corresponding to 6,316 different presentations. Full statistics are reported in Table 1.

3 Analysis

To demonstrate the quality and usability of our corpus, as well as provide baselines for future works, we develop and evaluate four different models using both automatic metrics and human evaluation.

3.1 Experimental Setting

3.1.1 Models

To establish baselines for the SAG task, we analyze the performance of four models described as follows. Prompts, model, generation, and additional training details are provided in App. C.

Whisper + LLama3.1-8B-Instruct. A cascaded solution, where the audio is first transcribed with `openai/whisper-large-v3` (Radford et al., 2022), and then `meta-llama/Llama-3.1-8B-Instruct` (Dubey et al., 2024) is prompted to generate the abstract from the generated transcript.

Qwen2-Audio-7B-Instruct. The `Qwen/Qwen2-Audio-7B-Instruct` (Chu et al., 2024) model, an existing SpeechLLM[5], which is used out of the box without any fine-tuning.

End2End Zero-Shot. A SpeechLLM composed of HuBERT (Hsu et al., 2021) as speech encoder, `meta-llama/Llama-3.1-8B-Instruct` as LLM, and a QFormer (Li et al., 2023) as adapter. The SpeechLMM is built to handle long audio inputs (App. C) and obtained by training only the adapter in two steps: (a) contrastive pretraining (Züfle and Niehues, 2024) to align the LLM representations for the speech and text modalities using MuST-C (Di Gangi et al., 2019) and Gigaspeech (Chen et al., 2021), and (b) fine-tuning on instruction-following tasks, including ASR, speech translation, and spoken question answering using MuST-C and Spoken-SQuAD (Lee et al., 2018). Therefore, the model is not trained or fine-tuned on NUTSHELL and operates in zero-shot for the SAG task.

End2End Finetuned. A SpeechLLM trained using the same contrastive pretraining procedure as End2End Zero-Shot but subsequently fine-tuned on

[5]By *SpeechLLM*, we refer to the combination of a speech encoder and an LLM through a learned modality adapter (Gaido et al., 2024).

Model	RougeL F1 ↑	BERTScore F1 ↑	Llama3.1-7B-Instruct Score with Expl. ↑	Plain Score ↑	Avg. Rank ↓	Human (on subset) Avg. Rank ↓
Whisper + LLama3.1-8B-Instruct	22.14	86.62	**77.84**	**82.47**	**1.24**	**1.53**
Qwen2-Audio-7B-Instruct	15.02	84.65	45.57	36.81	3.43	2.87
End2End Finetuned	**23.89**	**86.66**	68.78	73.53	1.98	1.6
End2End Zero-Shot	16.08	84.13	45.97	39.90	3.35	N/A

Table 2: We report results on the NUTSHELL test set for four models: a cascaded approach (Whisper+Llama-3.1-8B-Instruct), an existing SpeechLLM (Qwen2-Audio), and an end-to-end `HuBERT+QFormer+Llama3.1-8B-Instruct` model, either finetuned on our data (*End2End Finetuned*) or trained on audio instruction-following data (*End2End Zero-Shot*). Avg. Rank, assigned by an LLM judge or human annotators, reflects the mean ranking per model.

our NUTSHELL dataset. This not only evaluates the direct impact of task-specific datasets on the SAG performance, but it also ensures the feasibility of the task and the suitability of the collected data.

3.1.2 Evaluation

Metrics. We use standard (text) summarization metrics: **ROUGE** (Lin, 2004) – a text similarity metric that has been widely adopted for LM evaluation (Grusky, 2023) that focuses on n-gram overlap between the hypothesis and reference –, and **BERTScore** (Zhang et al., 2020) – a neural-based metric that measures the pairwise similarity of contextualized token embeddings between the summary and its reference. Also, we rely on **LLM-as-a-judge** (Shen et al., 2023; Zheng et al., 2024) where the LLM[6] is prompted to assign a score to each output, using the reference abstract as context (Score with Expl.). The score is based on four criteria: (1) relevance, (2) coherence, (3) conciseness, and (4) factual accuracy.[7] We also report results where the LLM judge provides a single score without explanations (Plain Score), as well as results where it ranks the given abstracts instead of scoring them individually (Avg. Rank).

All these metrics have known limitations and no metric is conclusively best for evaluating the SAG task: both ROUGE and BERTScore are known to fail to fully capture the extent to which two summaries share information (Deutsch and Roth, 2021) while LLM-as-a-judge is sensitive to prompt complexity and the length of input (Thakur et al., 2024) and struggle to distinguish similar candidates (Shen et al., 2023). For this reason, we complement the automatic scores with human evaluation.

[6]We use `Llama-3.1-8B-Instruct` (Dubey et al., 2024) as the judge using the prompts reported in Fig. 2 in App. D.2.

[7](1) *Does the predicted abstract capture the main points of the gold abstract?*, (2) *Is the predicted abstract logically organized and easy to follow?*, (3) *Is the predicted abstract free from unnecessary details?*, (4) *Are the claims in the predicted abstract consistent with the gold abstract?*

Human Evaluation. For the human evaluation, nine annotators – all experts in the field – were provided with the generated abstracts and the ground truth abstract. We use the same randomly sampled 30 test set examples as in Section 2.1 and validate their representativeness, which is discussed in App. E. Each sample is evaluated by three annotators. They follow the same criteria as the LLM evaluation but rank models instead of assigning scores. Detailed instructions are in App. E. As the End2End Zero-Shot model performance was comparable to that of Qwen2-Audio – also being a zero-shot model – and given that Qwen2-Audio is an established SpeechLLM with a distinct architecture, we exclude the End2End Zero-Shot from this analysis.

3.2 Results

Automatic Evaluation. Table 2 presents the performance of our models on the NUTSHELL test set. Among them, the cascaded model (Whisper + Llama3.1-8B-Instruct) achieves the highest scores across all LLM-based evaluation metrics. Instead, looking at both n-gram- and neural-based metrics, the End2End Finetuned model achieves the highest RougeL and BERTScore. In addition, Qwen2-Audio and our End2End Zero-Shot models demonstrate similar performance across all automatic metrics, showing a noticeable gap compared to the other two models. These results highlight the importance of our dataset for building high-performing end-to-end models, as the substantial gap between the cascaded and End2End Zero-Shot models is effectively bridged through fine-tuning on the NUTSHELL dataset.

For a more granular analysis, Table 3 in App. D.2 provides results for the LLM-based metrics. Given that all models except Qwen2-Audio rely on Llama3.1-8B-Instruct, one might question whether the Llama-based judge could introduce bias in favor of these models. To address this, we perform additional evaluations using `Qwen/Qwen2-7B` (Yang et al., 2024) as the judge (Table 4 in App. D.2), which confirm the same ranking, eliminating any concerns about evaluator bias.

Human Evaluation. As shown in Table 2, the human evaluation results closely align with the LLM-based judgments: the cascaded model ranks first, followed closely by the finetuned model while Qwen2-Audio ranks last. Notably, the gap between the first two models is small, whereas the difference between the second and third models is substantial – consistent with the LLM-based evaluation. This suggests that automatic metrics reliably capture both subtle and large performance differences between models. IAA, measured using pairwise rankings (Bojar et al., 2016) reached $\kappa = 0.53$, which is acceptable given the close ranking of the top two systems.

4 Conclusion

In this work, we introduce NUTSHELL, a novel dataset for SAG from recorded *ACL conference talks. By releasing this dataset under an open license, we hope to foster further advancements in SAG research and encourage the development of more effective models and evaluation techniques. Future work could explore the integration of the video content provided in the corpus, offering an additional modality for enriching the generation process and further improving abstract quality.

5 Limitations

While the current study provides a new resource and offers valuable insights about the SAG task, two main limitations should be noted:

- The analysis focused on the speech-to-text abstract generation task. However, our dataset also provides access to the corresponding videos, which were not utilized here. Future research could explore the integration of video content as an additional modality to enhance the generation process and improve the quality of the abstracts.

- The human evaluation was limited in scope, involving only a small set of models and samples. Future work could expand this evaluation to include more models and a larger number of samples to better assess the performance of different metrics and determine which is most effective in various contexts.

Potential Risks Generating automatic summaries for scientific talks carries the risk that automatic summaries may misrepresent key findings or lack scientific accuracy. However, we hope that by providing more high-quality training data, summarization models can be improved and lead to more reliable and accurate summaries.

Acknowledgments

This work has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People). We gratefully acknowledge Poland's high-performance Infrastructure PLGrid ACC Cyfronet AGH for providing computer facilities. Beatrice Savoldi and Marco Gaido are supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

We thank Leonard Bärmann, Béni Egressy, Mia Fuß, Lukas Hilgert, and Danni Liu for their contributions to the data annotation process.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 4376–4380. International Speech Communication Association. Publisher Copyright: Copyright © 2021 ISCA.; 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021 ; Conference date: 30-08-2021 Through 03-09-2021.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 195–205, Vancouver, Canada. Association for Computational Linguistics.

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online. Association for Computational Linguistics.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech translation with speech foundation models and large language models: What is there and what is missing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14760–14778, Bangkok, Thailand. Association for Computational Linguistics.

Max Grusky. 2023. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. SumPubMed: Summarization dataset of PubMed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 292–303, Online. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. 2021. Attention-based multi-hypothesis fusion for speech summarization. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 487–494. IEEE.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text Generation from Knowledge Graphs with Graph Transformers. 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 2284–2293, Minneapolis, Minnesota. Association for Computational Linguistics.

Fajri Koto, Sakriani Sakti, Graham Neubig, Tomoki Toda, Mima Adriani, and Satoshi Nakamura. 2014. The use of semantic and acoustic features for open-domain ted talk summarization. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–4.

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

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proc. Interspeech 2018*, pages 3459–3463.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. TalkSumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2125–2131, Florence, Italy. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

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

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024. SumSurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9632–9651, Bangkok, Thailand. Association for Computational Linguistics.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022. CiteSum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10922–10935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Masato Mimura, Takatomo Kano, Atsunori Ogawa, and Marc Delcroix. 2024. Sentence-wise speech summarization: Task, datasets, and end-to-end modeling with lm knowledge distillation. *Preprint*, arXiv:2408.00205.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, Dennis Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140. Noldus Information Technology.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *Preprint*, arXiv:1811.00347.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937, Toronto, Canada. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

Gianluca Vico and Jan Niehues. 2022. Ted talk teaser generation with pre-trained models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8067–8071.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the Thirty-Third AAAI Conference* *on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Fabio Massimo Zanzotto, Viviana Bono, Paola Vocca, Andrea Santilli, Danilo Croce, Giorgio Gambosi, and Roberto Basili. 2020. Gasp! generating abstracts of scientific papers from abstracts of cited papers. *ArXiv*, abs/2003.04996.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Maike Züfle and Jan Niehues. 2024. Contrastive learning for task-independent speechllm-pretraining. *Preprint*, arXiv:2412.15712.

A Human Evaluation: Are abstracts good summaries of the talk?

We aim to assess whether paper abstracts can serve as effective abstracts for *ACL talks. To this end, we conducted a human evaluation by randomly sampling 30 examples from our dataset. The annotation team consisted of five individuals (four women and one man), including the paper authors and their colleagues. All annotators were already familiar with the NLP domain, scientific presentation and writing, and the task itself. They are experts in Natural Language Processing, holding at least a master's degree in NLP or a related field, with some holding PhDs or professorial positions. Their ages ranged from 25 to 55.

The annotation guidelines were initially developed by the authors and subsequently refined in collaboration with the annotators to ensure a shared and well-defined set of evaluation criteria. Detailed instructions for the human annotators are provided in Fig. 3. The annotation template also included a comment section for uncertain cases, though no comments were submitted.

B Dataset Details

We include all *ACL conferences from 2017 to 2022 in NUTSHELL, covering main conferences, Findings, industry tracks, and workshops. As not all conferences are held every year, the number of talks varies accordingly. Table 5 provides a detailed overview.

C Baseline Details

Generation Settings We evaluate four different models to establish baselines for abstract generation from spoken ACL talks. The evaluations were conducted on a single NVIDIA A100-SXM4-40GB GPU.

For all models, we use the default generation parameters and apply greedy search, following the usage instructions for `meta-llama/Llama-3.1-8B-Instruct`[8] (Dubey et al., 2024), `Qwen/Qwen2-Audio-7B-Instruct`[9] (Chu et al., 2024) and the contrastively pretrained models from Züfle and Niehues (2024)[10].

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[9] https://github.com/QwenLM/Qwen2-Audio

[10] https://github.com/MaikeZuefle/contr-pretraining

Cascaded Model For the cascaded model, we segment the audio into 30-second chunks and transcribe them using `openai/whisper-large-v3` (Radford et al., 2022). The transcribed chunks are then concatenated and processed by `meta-llama/Llama-3.1-8B-Instruct` (Dubey et al., 2024) to generate the abstract. Inference took 5:40 hours on a single NVIDIA A100-SXM4-40GB GPU, including transcribing and summarizing.

Since the model's outputs often included a title and category for the talk, we explicitly prompt it to generate only the abstract. This adjustment was not necessary for the other models.

We use the following prompt:

System Prompt:

```
A chat between a curious user and an
artificial intelligence assistant. The
assistant gives helpful, detailed, and
polite answers to the user's questions.\
n
```

Prompt:

```
Summarize the following talk to create
an abstract for an ACL Paper, don't
include the title or other information,
only the abstract:\n<transcription>\n
```

Qwen2-Audio For `Qwen/Qwen2-Audio-7B-Instruct` (Chu et al., 2024), inference took 50 minutes on a single NVIDIA A100-SXM4-40GB GPU. We use the system prompt as provided in the code documentation[9].

System Prompt:

```
You are a helpful assistant.
```

Prompt:

```
Summarize this talk to create an
abstract for an ACL Paper:\n
```

Contrastively Pretained Models For the contrastively pretrained model, we follow Züfle and Niehues (2024) and adopt their settings[10], including training configurations, hyperparameters, and system prompts. The SpeechLLM consists of HuBERT (Hsu et al., 2021) as speech encoder, `meta-llama/Llama-3.1-8B-Instruct` as LLM, and a QFormer (Li et al., 2023) as adapter. We choose HuBERT as an encoder in contrast to the bigger and more powerful `openai/whisper-large-v3` (Radford et al., 2022), as it needs less memory and is therefore more suitable for the summarization task of longer audio. However, due to the extended duration of the audio inputs, we additionally introduce two modifications:

Model	Llama-3.1-8B-Instruct						
	Relevance ↑	Coherence ↑	Conciseness ↑	Factual Accuracy ↑	Avg. Score ↑	Plain Score ↑	Avg. Rank ↓
Whisper + LLama31-Instruct	**77.12**	**86.00**	**61.13**	**87.13**	**77.84**	**82.47**	**1.24**
Qwen2-Audio	37.21	52.52	45.91	46.63	45.57	36.81	3.43
End2End Finetuned	66.41	78.24	50.25	80.22	68.78	73.53	1.98
End2End Zero-Shot	40.28	48.02	37.69	57.89	45.97	39.90	3.35

Table 3: Results using Llama-3.1-8B-Instruct as a judge. We report results on the NUTSHELL test set for four models: a cascaded approach (`openai/whisper-large-v3 + meta/Llama-3.1-8B-Instruct`), `Qwen/Qwen2-Audio-7B-Instruct`, and an end-to-end `HuBERT+QFormer+Llama3.1-7B-Instruct` model, either finetuned on our data (*End2End Finetuned*) or trained on audio instruction-following data (*End2End Zero-Shot*). Avg. Rank reflects the mean ranking per model.

Model	Qwen2-7bInstruct						
	Relevance ↑	Coherence ↑	Conciseness ↑	Factual Accuracy ↑	Avg. Score ↑	Plain Score ↑	Avg. Rank ↓
Whisper + LLama31-Instruct	**79.61**	**83.54**	72.08	**86.07**	**80.33**	**74.60**	**1.66**
Qwen2-Audio	56.99	75.35	**75.91**	59.28	66.88	49.55	3.18
End2End Finetuned	75.13	81.78	75.04	81.16	78.28	70.83	2.12
End2End Zero-Shot	57.93	68.02	69.34	66.65	65.49	53.61	3.04

Table 4: Results using Qwen2-7bInstruct as a judge. We report results on the NUTSHELL test set for four models: a cascaded approach (`openai/whisper-large-v3 + meta/Llama-3.1-8B-Instruct`), `Qwen/Qwen2-Audio-7B-Instruct`, and an end-to-end `HuBERT+QFormer+Llama3.1-7B-Instruct` model, either finetuned on our data (*End2End Finetuned*) or trained on audio instruction-following data (*End2End Zero-Shot*). Avg. Rank reflects the mean ranking per model.

Split	**Conference**	**Year**	**Talks**
Train	ACL	2017	140
		2018	185
		2019	244
		2021	849
	EMNLP	2017	93
		2018	221
		2021	1480
	NAACL	2018	120
		2019	114
		2021	554
Dev	ACL	2022	885
Test	EMNLP	2022	465
	NAACL	2022	966

Table 5: Number of talks per conferences in the NUTSHELL dataset.

1. We segment the audio into one-minute chunks, encode each chunk using the encoder and then concatenate the encoded representations before passing them through the adapter and LLM backbone.

2. We use a batch size of 1 for fine-tuning with NUTSHELL.

Despite these adjustments, we encountered memory limitations for audio files exceeding 35 minutes. In such cases, we truncate the audio to 35 minutes, which affects one example in the test set.

The training of the models was conducted on four NVIDIA A100-SXM4-40GB GPUs. The contrastive pretraining took 33 hours on four GPUS. Finetuning on ASR, speech translation, and spoken question answering data took 30 hours, finetuning on the NUTSHELL dataset took 2:10 hours. Generating the outputs of the test set (on a single NVIDIA A100-SXM4-40GB GPU) took 2:35 hours.

System Prompt:

```
A chat between a curious user and an
artificial intelligence assistant. The
assistant gives helpful, detailed, and
polite answers to the user's questions.\
n
```

Prompt:

```
Summarize this talk to create an
abstract for an ACL Paper:
```

D Evaluation Details

We evaluate the results of our models using automatic metrics including ROUGE, BERTScore, and LLM-as-a-judge.

D.1 ROUGE and BERT Score

As automatic metrics, we use ROUGE[11] (Lin, 2004) and BERTScore (Zhang et al., 2020).

11

Model	RougeL F1 ↑	BERTScore F1 ↑	Llama3.1-7B-Instruct Score with Expl. ↑	Plain Score ↑	Avg. Rank ↓
Whisper + LLama31-Instruct	23.26	86.81	**77.75**	**84.30**	**1.23**
Qwen2-Audio	16.26	84.94	48.42	39.50	3.47
End2End Finetuned	**24.47**	**86.71**	70.67	75.73	1.83

Table 6: Baseline Results, the finetuned model is a HuBERT + Qformer + LLama31Instruct model on the subset used for human annotation (30 examples).

Concretely, we compute ROUGE-L, which focuses on the longest common subsequence, with `DD/sacrerouge` (Deutsch and Roth, 2020), as recommended by Grusky (2023) and for BERTScore, we use the `bertscore` implementation from HuggingFace[12] and report the F1-score.

D.2 LLM as a judge

To evaluate the model outputs, we also use an LLM *as a judge*, specifically `meta-llama/Llama-3.1-8B-Instruct` (Dubey et al., 2024). The LLM assigns a score to each output using the reference abstract as context, based on four criteria: (1) relevance (*Does the predicted abstract capture the main points of the gold abstract?*), (2) coherence (*Is the predicted abstract logically organized and easy to follow?*), (3) conciseness (*Is the predicted abstract free from unnecessary details?*), and (4) factual accuracy (*Are the claims in the predicted abstract consistent with the gold abstract?*). Additionally, we report results where the LLM provides a single overall score without explanations and results where it ranks the given abstracts instead of scoring them individually. The prompts are given in Fig. 2. If the model fails to return a valid json dictionary, we instead take the first number after the score name in the output. We present the results for all four criteria, the average score, the score without explanations, and the ranking in Table 3. One potential concern is that this LLM might be biased, as all our models except Qwen2-Audio are based on Llama-3.1. However, we find this is not the case. When using `Qwen/Qwen2-7B` (Yang et al., 2024) as the judge, we obtain the same ranking as with Llama. The results with Qwen-as-a-judge can be found in Table 4.

E Human Evaluation for Model Outputs

We evaluate the models using ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and LLM-as-a-judge. However, it is known that automatic evaluation metrics can come with limitations. Namely, the first two metrics may not fully capture semantic overlap (Deutsch and Roth, 2021), while LLM-as-a-judge is sensitive to prompt phrasing (Thakur et al., 2024) and struggles to distinguish between closely similar candidates (Shen et al., 2023). To validate the reliability of our automatic evaluation scores and better understand model behavior, we complement these metrics with a human evaluation. This allows us also to verify the robustness of our findings.

Specifically, we asked nine domain experts (four women and five men) to rank model outputs relative to the reference abstract, with each example annotated by three independent annotators. All annotators were already familiar with the NLP domain, scientific writing and presentation, and the task itself. They are experts in Natural Language Processing, holding at least a master's degree in NLP or a related field, with some holding PhDs or professorial positions. Their ages ranged from 25 to 55. The annotation instructions are provided in Fig. 4.

We conduct this human evaluation on a randomly selected subset of 30 test examples. We consider this subset representative, as the model rankings based on automatic metrics remain consistent with those on the full test set. The corresponding automatic scores for this subset are reported in Table 6. We want to include three diverse models in our human evaluation: a zero-shot model, a cascaded model, and a model finetuned on our dataset. Since we have two zero-shot models (Qwen2-Audio and our contrastively pretrained zero-shot model) that perform similarly, we decided to exclude one for efficiency in the human evaluation. We keep the Qwen2-Audio model as this is an already existing and widely used SpeechLLM.

[12] https://huggingface.co/spaces/evaluate-metric/bertscore

System Prompt for Score with Explanation:

```
You are an expert AI trained to evaluate scientific abstracts. Your task is to
compare a predicted abstract with a gold standard (reference) abstract and provide a
 detailed evaluation based on the following criteria:\n\n
1. **Relevance**: Does the predicted abstract capture the main points of the gold
abstract?\n
2. **Coherence**: Is the predicted abstract logically organized and easy to follow?\
n
3. **Conciseness**: Is the predicted abstract free from unnecessary details?\n
4. **Factual Accuracy**: Are the claims in the predicted abstract consistent with
the gold abstract?\n\n
For each criterion:\n
- Assign a **score** between 1 and 10 (1 = very poor, 10 = excellent).\n"
- Provide a **brief explanation** for the assigned score.\n\n"
Your output must be in the following JSON format:\n\n"
{\"relevance\": {\"score\": int, \"explanation\": \"string\"},
 \"coherence\": {\"score\": int, \"explanation\": \"string\"},
 \"conciseness\": {\"score\": int, \"explanation\": \"string\"},
 W\"factual_accuracy\": {\"score\": int, \"explanation\": \"string\"}}\n\n
```

Prompt for Score with Explanation:

```
### Gold Abstract:\n<reference abstract>\n\n### Predicted Abstract:\n<predicted
abstract>\n\nPlease evaluate the predicted abstract based on the criteria mentioned.
```

System Prompt for Score without Explanation:

```
You are an expert AI trained to evaluate scientific abstracts. Your task is to
compare a predicted abstract with a reference abstract. Evaluate how well the
prediction aligns with the reference using a score from 0 (lowest) to 100 (highest).
 Your output must only be in the following JSON format: {\"prediction\": int}. Do
not provide any explanation or additional text.
```

Prompt for Score without Explanation:

```
### Reference Abstract:\n<reference abstract>\n\n### Predicted Abstract:\n<predicted
 abstract>\n\nPlease evaluate the predicted abstract with respect to the reference
abstract and assign a score from 0 to 100.
```

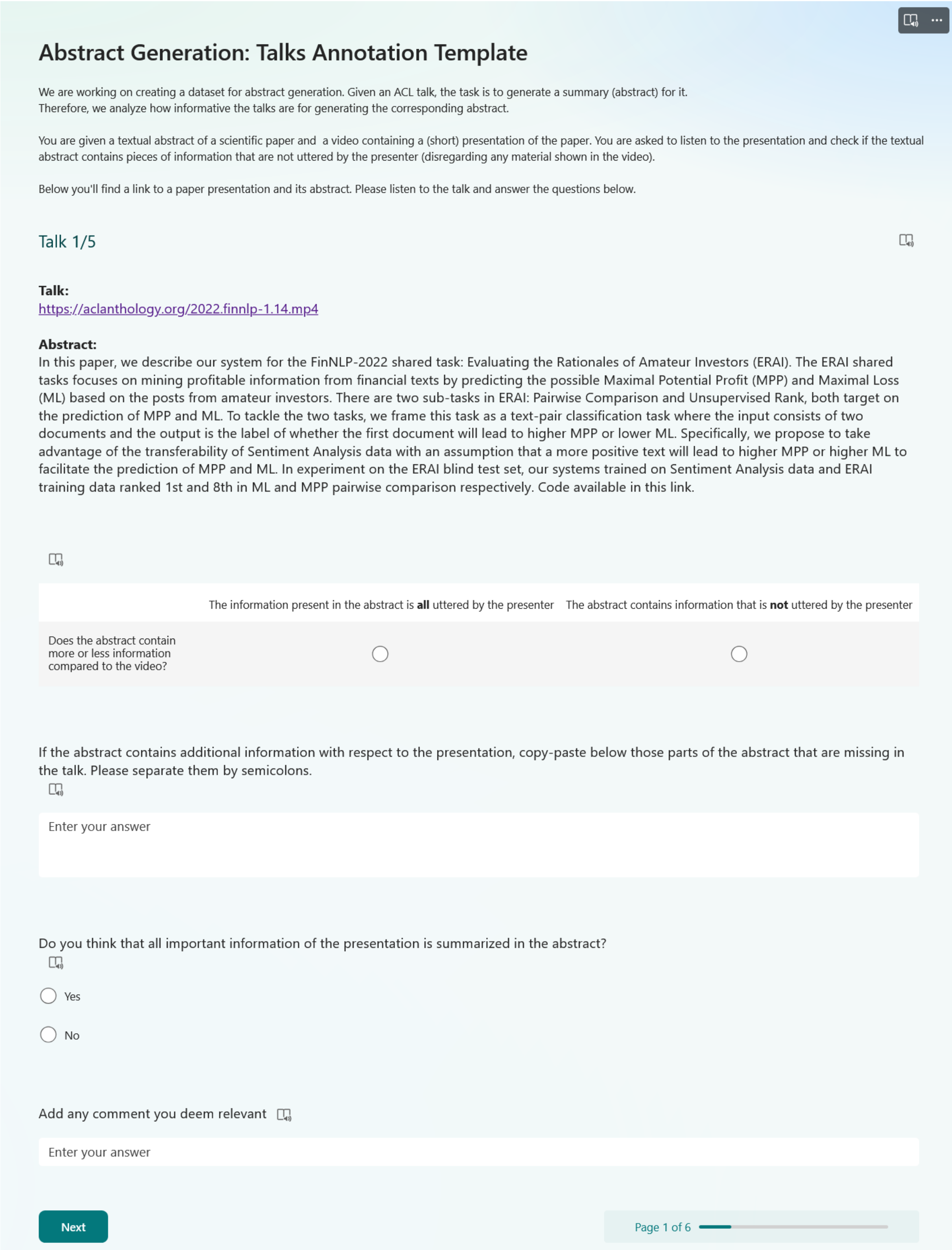
System Prompt for Ranking:

```
You are an expert AI trained to evaluate scientific abstracts. Your task is to rank
four different abstracts based on a reference abstract. Your output must only be in
the following format: <Model A, Model B, Model C, Model D> where the first model is
the best model, and the last model the weakest. Do not provide any explanation or
additional text.
```

Prompt for Ranking:

```
### Reference Abstract:\n<reference abstract>\n\n
### Model A Predicted Abstract:\n<predicted abstract 1>\n\n
### Model B Predicted Abstract:\n<predicted abstract 2>\n\n
### Model C Predicted Abstract:\n<predicted abstract 3>\n\n
### Model D Predicted Abstract:\n<predicted abstract 4>\n\n
Please rank the four predicted abstracts.
```

Figure 2: Prompts for LLM as a judge. We use the same prompt for both, Qwen2-7bInstruct and Llama 3.1 8B Instruct. <reference abstract> and <predicted abstract> are replaced with the actual abstracts. For ranking, we shuffle the predicted abstracts, so that the LLMs sees the abstracts of different models in a different order every time to avoid position bias.



Abstract Generation: Talks Annotation Template

We are working on creating a dataset for abstract generation. Given an ACL talk, the task is to generate a summary (abstract) for it.
Therefore, we analyze how informative the talks are for generating the corresponding abstract.

You are given a textual abstract of a scientific paper and a video containing a (short) presentation of the paper. You are asked to listen to the presentation and check if the textual abstract contains pieces of information that are not uttered by the presenter (disregarding any material shown in the video).

Below you'll find a link to a paper presentation and its abstract. Please listen to the talk and answer the questions below.

Talk 1/5

Talk:
https://aclanthology.org/2022.finnlp-1.14.mp4

Abstract:
In this paper, we describe our system for the FinNLP-2022 shared task: Evaluating the Rationales of Amateur Investors (ERAI). The ERAI shared tasks focuses on mining profitable information from financial texts by predicting the possible Maximal Potential Profit (MPP) and Maximal Loss (ML) based on the posts from amateur investors. There are two sub-tasks in ERAI: Pairwise Comparison and Unsupervised Rank, both target on the prediction of MPP and ML. To tackle the two tasks, we frame this task as a text-pair classification task where the input consists of two documents and the output is the label of whether the first document will lead to higher MPP or lower ML. Specifically, we propose to take advantage of the transferability of Sentiment Analysis data with an assumption that a more positive text will lead to higher MPP or higher ML to facilitate the prediction of MPP and ML. In experiment on the ERAI blind test set, our systems trained on Sentiment Analysis data and ERAI training data ranked 1st and 8th in ML and MPP pairwise comparison respectively. Code available in this link.

	The information present in the abstract is **all** uttered by the presenter	The abstract contains information that is **not** uttered by the presenter
Does the abstract contain more or less information compared to the video?	○	○

If the abstract contains additional information with respect to the presentation, copy-paste below those parts of the abstract that are missing in the talk. Please separate them by semicolons.

Enter your answer

Do you think that all important information of the presentation is summarized in the abstract?

○ Yes

○ No

Add any comment you deem relevant

Enter your answer

Next

Page 1 of 6

Figure 3: Instructions for annotators to evaluate whether the paper abstracts are good and informative abstracts for the ACL talks.

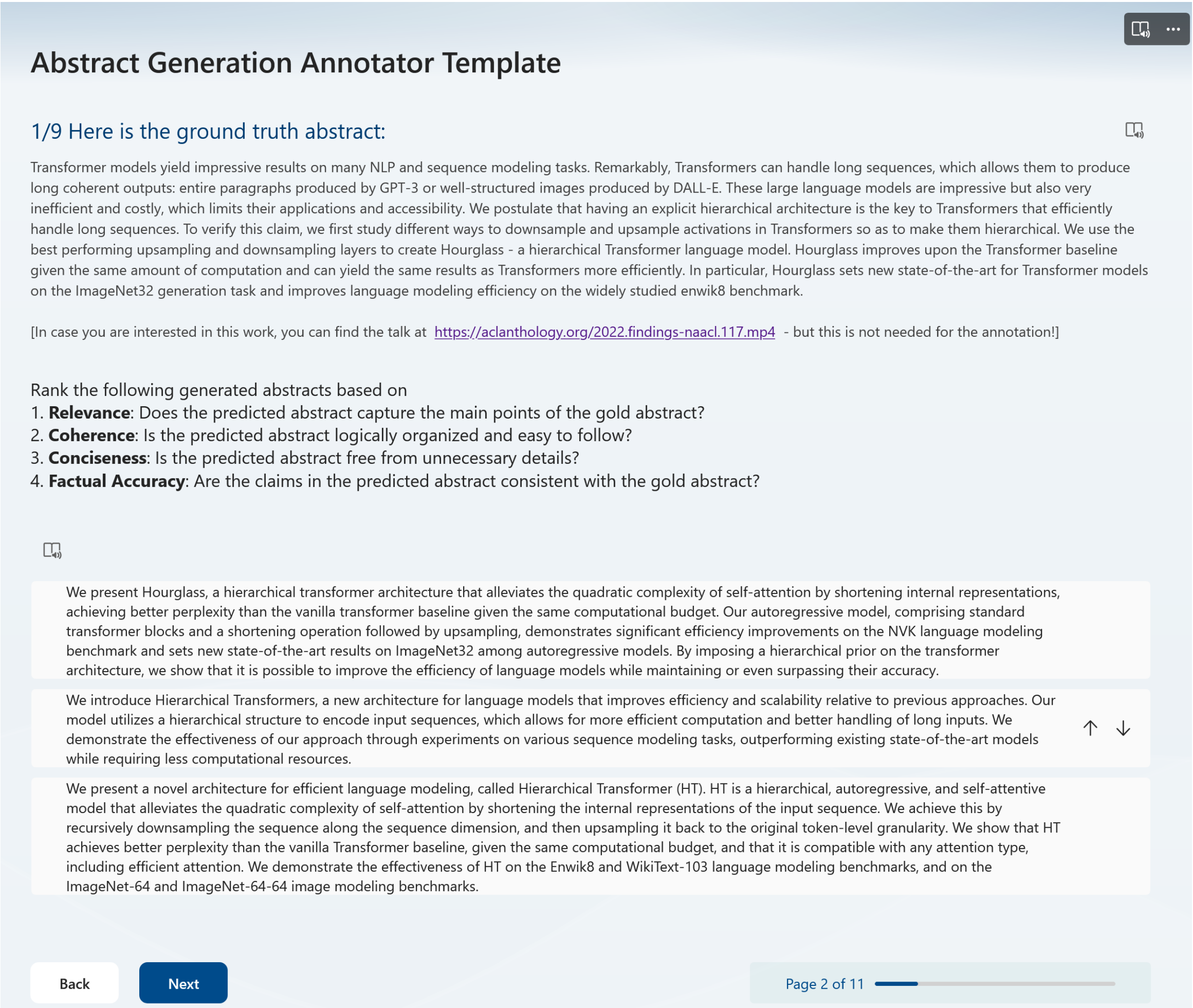


Figure 4: Instructions for human annotators for ranking model outputs.