Unsloth PEFT based Multilingual Meeting Summarization with Open-Source LLMs

A Comparative Analysis of LLaMA, DeepSeek, and Mistral Models in Zero-Shot and Few-Shot Settings

Dr. Nupur Giri
nupur.giri@ves.ac.in
Vivekanand Education Society's
Institute of Technology
Chembur, Mumbai, India

Manraj Singh Virdi
d2021.manrajsingh.virdi@ves.ac.in
Vivekanand Education Society's
Institute of Technology
Chembur, Mumbai, India

Sakshi Kirmathe
d2021.sakshi.kirmathe@ves.ac.in
Vivekanand Education Society's
Institute of Technology
Chembur, Mumbai, India

Deven Bhagtani
d2021.deven.bhagtani@ves.ac.in
Vivekanand Education Society's
Institute of Technology
Chembur, Mumbai, India

Piyush Chugeja
d2021.piyush.chugeja@ves.ac.in
Vivekanand Education Society's
Institute of Technology
Chembur, Mumbai, India

ABSTRACT

This paper presents a systematic approach to multilingual meeting summarization using open-source large language models. Three model families, LLaMA 3, Mistral, and DeepSeek, were evaluated under zero-shot, one-shot, and three-shot settings on a specially prepared dataset of career counseling meeting transcripts. The best models were fine-tuned using Unsloth's 4-bit quantization and Parameter-Efficient Fine-Tuning (PEFT) with Low Rank Adaptation (LoRA) methods. Experimental results showed that the fine-tuned LLaMA 3.1 (8B) model showed higher efficacy in both English and multilingual settings (English, Hindi, and Marathi), generating high-quality summaries, efficiency, and stable cross-lingual generalization. These findings provide valuable insights into scalable, domain-specific summarization approaches that well balance accuracy and computational cost, thereby enabling the construction of practical, AI-based tools for career counseling.

KEYWORDS

LLM Fine-Tuning, Multilingual Summarization, Open-Source LLMs, Unsloth Fine-Tuning

1 INTRODUCTION

The mass adoption of online meetings has reshaped professional communication, with platforms such as Zoom and Google Meet hosting millions of users daily [1, 2]. Career counseling, which assists people make informed career choices, has also been online, enhancing convenience for both clients and counselors. However, reading lengthy transcripts of these sessions to identify the most important points remains a significant challenge. It is time-consuming and not practical to analyze transcripts manually, so automated summarization is an essential solution.

Large Language Models (LLMs), built on transformer architectures [3], have significantly advanced natural language processing tasks, including summarization. Although LLMs have been widely explored for document summarization, research on summarizing conversations and meetings is relatively limited [4, 5]. Multilingual summarization, particularly for Indic languages like Hindi and Marathi, has received even less attention, despite a large user base relying on these languages for professional communication. Existing research on Indic language summarization focuses mainly on structured content such as news articles and formal reports [6]. However, career counseling conversations are more dynamic and require summarization techniques that can capture key insights from interactive dialogues. This gap highlights the need for models specifically optimized for multilingual meeting summarization.

In this paper, we evaluate open source LLMs from LLaMA 3 family [7], DeepSeek family[8], and Mistral family[9] to generate summary of career counseling sessions in English, Hindi and Marathi. We compare their performance in zero-shot, few-shot and fine-tuned settings using multiple evaluation metrics. Fine-tuning was performed using Unsloth [10], leveraging Parameter-Efficient Fine-Tuning (PEFT) to improve adaptability while minimizing computational overhead. In addition, we discuss challenges encountered during fine-tuning, including overfitting issues observed in some models.

2 RELATED WORK

With the rise of online meetings, researchers have been working on ways to summarize them effectively. Different methods exist, such as extractive summarization (which picks key sentences from the text), abstractive summarization (which rewrites the content in a shorter form), and hybrid approaches that combine both. However, challenges remain especially in handling multiple languages, summarizing in real time, and keeping the meaning clear in long, complex discussions.

Transformer-based models [3] have improved abstractive summarization. For example, Pointer Generator Networks [11] help avoid repetition and make summaries easier to read, but they depend on large, general-purpose datasets, making them less useful for specialized topics. Jotter [12], which combines BERT embeddings with sequence-to-sequence models, balances accuracy and fluency but requires a lot of computing power, making it less practical for real-time applications. Kumar and Kabiri [13] point out that most models use datasets like AMI and ICSI, which, while useful, do not always capture the specific details needed for fields like

Table 1: Open Source Large Language Models (LLMs) chosen for Inference Drawing

Model	Model Creator	#Parameters	Instruction Tuning
LLaMA 3	Meta	8B	✓
LLaMA 3.1		8B	✓
LLaMA 3.2		3.2B	✓
Mistral	Mistral	7B	✓
Mistral v0.3 Instruct		7B	✓
Deepseek LLaMA	DeepSeek	8B	✓

career counseling. For multilingual meeting summaries, AI-based methods have been developed. One such approach [14] uses Latent Semantic Analysis (LSA) to identify key points, but this method tends to oversimplify discussions. Transformer models perform better because they retain more context, making them more effective for summarizing conversations in different languages. Structured summarization methods have also been useful for specific fields. For example, ConSum [15], designed for mental health counseling, filters important speech patterns using PHQ-9-based scoring, showing that using specialized knowledge can make summaries more relevant.

Another key advancement is instruction tuning, which has been found to be more effective than traditional fine-tuning for text summarization. Zhang et al. [16] studied news summarization and found that instruction tuning helps models perform better without needing large, domain-specific datasets. Unlike fine-tuning, which requires a lot of training data, instruction tuning allows models to improve with high-quality prompts and well-structured instructions. This is particularly useful for summarizing meetings, where clear instructions can help models generate meaningful summaries even with limited training data. The growing popularity of open-source models has also led to new developments in efficient model training. While proprietary models like GPT-4 are strong at summarization without extra training, open-source models like LLaMA, DeepSeek, and Mistral [7–9] are becoming more popular because they offer competitive performance and better privacy. To make these models more efficient, researchers have developed Parameter-Efficient Fine-Tuning (PEFT) methods. One such method, QLoRA [17], helps fine-tune large models with minimal memory usage by adding small, learnable layers instead of retraining the entire model. Similarly, LoRA [18] injects lightweight layers into Transformers, making them more adaptable while keeping computing costs low.

Despite these advancements, challenges remain, especially in summarizing multilingual meetings, processing summaries in real time, and improving instruction tuning. Addressing these issues will lead to better AI-powered tools for summarizing professional discussions and improving decision-making.

3 OUR WORK

Existing research on meeting summarization often focuses on evaluating single models or relies on domain-agnostic datasets, which limits their effectiveness in more specialized contexts. Our study takes a different approach by evaluating multiple open-source LLM families. Table 1 lists the models used in this research, assessing their performance across zero-shot, one-shot, and three-shot scenarios with meeting transcripts. This comparison provides valuable insights into how different models handle the complexities of structured dialogue-based summarization. In contrast to previous studies that rely on generic benchmark datasets [13], we fine-tune the best-performing models from each family on our own dataset, specifically optimizing for English meeting summarization. This domain-specific fine-tuning improves the contextual coherence of the summaries. After identifying the best performing model, we extended its capabilities to handle multilingual summarization in English, Hindi, and Marathi, addressing a critical gap in non-English meeting summarization research [14].

To improve computational efficiency, we utilize Unsloth's 4-bit quantization [10], which reduces memory usage without compromising performance. This enables us to fine-tune large models with minimal computational overhead, making this approach more scalable for real world applications. This study presents a more adaptable and resource efficient pipeline for meeting summarization by combining structured evaluation, domain specific fine tuning and efficient quantization.

4 METHODOLOGY

This section explains the steps taken in this study. The work began by testing three families of models LLaMA, Mistral, and DeepSeek, as highlighted in Table 1 using a custom dataset of English meeting transcripts. The testing involved three different settings: zero-shot, one-shot, and three-shot prompts. In each case, a prompt was provided to the model with 0, 1, or 3 examples (shots) to help generate a summary in a structured JSON format.

The testing followed an approach similar to that in [16], where models are given different numbers of examples to understand how well they can summarize a news article. For each model, several performance metrics were calculated. These metrics include:

- **ROUGE:** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap of words or phrases between the generated summary and a reference summary [19]. Higher ROUGE scores indicate that the summary has captured more of the important content.
- **BERT Score:** This metric uses BERT embeddings to compare the similarity between the generated and reference summaries [20]. It is helpful because it considers the meaning of the text, not just exact word matches.
- **BLEU:** BLEU (Bilingual Evaluation Understudy) measures how many words in the generated summary match the

Table 2: Testing pre-trained LLMs on a custom dataset of English transcripts to identify the best performing model from each model family

Setting	Model	ROUGE			BERT Score				
		R1	R2	RL	Precision	Recall	F1	BLEU	GLEU
0-shot inference	Mistral (7B)	0.3823	0.1791	0.3099	0.9243	0.8819	0.9025	0.0506	0.1315
	LLaMA 3 (8B)	0.4318	0.1863	0.3432	0.9277	0.8926	0.9098	0.0901	0.1657
	LLaMA 3.2 (3B)	0.3377	0.1410	0.2503	0.9187	0.8759	0.8967	0.0274	0.1045
	LLaMA 3.1 (8B)	0.4830	0.2374	0.3677	0.9341	0.9004	0.9169	0.1324	0.2058
	Mistral v0.3 Instruct (7B)	0.4920	0.2439	0.3891	0.9232	0.9082	0.9155	0.1557	0.2221
	Deepseek LLaMA (8B)	0.2115	0.3528	0.3528	0.9230	0.9112	0.9169	0.1382	0.2105
1-shot inference	Mistral (7B)	0.5059	0.2321	0.3916	0.9116	0.9108	0.9111	0.1354	0.1892
	LLaMA 3 (8B)	0.5037	0.2646	0.4130	0.9349	0.9057	0.9200	0.1540	0.2224
	LLaMA 3.2 (3B)	0.5369	0.2557	0.4114	0.9456	0.9239	0.9346	0.2114	0.2744
	LLaMA 3.1 (8B)	0.5503	0.3145	0.4611	0.9381	0.9170	0.9274	0.2213	0.2758
	Mistral v0.3 Instruct (7B)	0.6041	0.3596	0.5182	0.9389	0.9272	0.9329	0.2642	0.3118
	Deepseek LLaMA (8B)	0.5321	0.2902	0.4343	0.9322	0.9149	0.9234	0.2117	0.2665
3-shot inference	Mistral (7B)	0.5554	0.3250	0.4759	0.9431	0.9200	0.9313	0.2310	0.2935
	LLaMA 3 (8B)	0.5554	0.3250	0.4759	0.9431	0.9200	0.9313	0.2309	0.2935
	LLaMA 3.2 (3B)	0.5369	0.2557	0.4114	0.9456	0.9239	0.9346	0.2114	0.2744
	LLaMA 3.1 (8B)	0.5194	0.2926	0.4414	0.9369	0.9114	0.9239	0.2023	0.2593
	Mistral v0.3 Instruct (7B)	0.5964	0.3428	0.4963	0.9287	0.9319	0.9302	0.2629	0.3086
	Deepseek LLaMA (8B)	0.5691	0.3144	0.4646	0.9366	0.9246	0.9305	0.2176	0.2822

reference summary [21]. It is commonly used in machine translation and summarization.

- **GLEU:** GLEU is similar to BLEU but is designed to be more sensitive to fluency and grammatical correctness [22].

Table 2 shows the performance of each model in the three settings. In simple terms, higher numbers in these metrics mean better performance. For example, a higher ROUGE score means more important parts of the transcript are captured in the summary.

The results show that, across different testing conditions, some models perform better than others. For instance, in the zero-shot setting, the LLaMA 3.1 (8B) model produced higher ROUGE scores (with ROUGE-1 at 0.483 and ROUGE-L at 0.3677) compared to its peers, indicating that it captured more key content from the transcripts. In the one-shot setting, Mistral v0.3 Instruct (7B) achieved the highest scores, with a ROUGE-1 score of 0.6041 and improved BLEU and GLEU values, showing that it generated summaries that closely matched the reference texts. Similarly, for the three-shot setting, Mistral v0.3 Instruct (7B) maintained strong performance. Based on these findings and considering all evaluation metrics, the best models from each family were selected for further fine-tuning:

- **LLaMA Family:** LLaMA 3.1 (8B) was chosen because it consistently produced high ROUGE and BERT Score values, reflecting its ability to retain important content and context.
- **DeepSeek Family:** DeepSeek LLaMA (8B) was selected as it showed reliable performance in multiple metrics even though its overall scores were slightly lower than some others.
- **Mistral Family:** Mistral v0.3 Instruct (7B) was the top performer in one-shot and three-shot scenarios, demonstrating a strong ability to generate fluent and accurate summaries.

Following this model selection, the next phase involved fine-tuning these three models on the dataset, focusing first on English transcripts. The fine-tuning process was carried out using Unsloth's 4-bit quantization framework [10, 17], which significantly reduces the memory requirements while maintaining performance. After fine-tuning on English data, the best model was then extended to support multilingual summarization for English, Hindi, and Marathi.

Each metric used in the evaluation plays an important role in understanding model performance:

- **ROUGE** measures the overlap between the generated summary and the reference summary. It is widely used to evaluate the recall of important information [19].
- **BERT Score** uses semantic similarity, which helps in capturing the meaning even if the exact words differ [20].
- **BLEU** and **GLEU** measure the similarity in wording and fluency between the generated and reference texts [21, 22].

5 RESULTS AND DISCUSSION

Three models were fine-tuned on a dataset of career counseling meeting transcripts using Unsloth's 4-bit quantization framework. The fine-tuning process used Parameter-Efficient Fine-Tuning (PEFT) with LoRA [10, 18]. Key hyperparameters were set as follows: a learning rate of 1e-5, a per-device batch size of 2, gradient accumulation steps of 8, and a maximum sequence length of 4096 tokens. Training was carried out with the cosine learning rate scheduler and an 8-bit variant of the AdamW optimizer.

Before fine-tuning, the pre-trained models generated summaries based on an input prompt that included an example transcript and the expected JSON output. This pre-fine-tuning phase helped to understand the initial performance of the models. After fine-tuning, the models were expected to generate more accurate and consistent

Table 3: Performance Comparison of Fine-tuned Models on English Transcripts

Model	ROUGE-L	BERT F1	BLEU	GLEU	Inference Time
Mistral v0.3 Instruct (7B)	0.4796	0.6932	0.2745	0.3496	16.5s
LLaMA 3.1 (8B)	0.5178	0.9378	0.3253	0.3334	12.3s
DeepSeek LLaMA (8B)	0.4527	0.7903	0.3103	0.2396	19.2s

summaries. The evaluation was carried out using metrics such as ROUGE-L, BERT Score F1, BLEU, and GLEU, along with inference time, measured on a P100 Graphics Processing Unit (GPU) in a Kaggle notebook environment.

5.1 Fine-Tuning Results

Table 3 summarizes the performance of the three fine-tuned models on English meeting transcripts. The LLaMA 3.1 (8B) model achieved a ROUGE-L score of 0.5178 and a BERT Score F1 of 0.9378. It also recorded a BLEU score of 0.3253 and a GLEU score of 0.3334, while requiring only 12.3 seconds of inference time. In comparison, the Mistral v0.3 Instruct (7B) model had lower scores and a longer inference time of 16.5 seconds, and DeepSeek LLaMA (8B) showed lower overall performance with an inference time of 19.2 seconds.

The comparison shows that fine-tuning has markedly improved the ability of the models to capture the key content of transcripts. In particular, LLaMA 3.1 (8B) demonstrated a significant improvement in both the ROUGE-L and BERT Score F1 metrics. The pre-fine-tuning outputs were less structured and less consistent, whereas post-fine-tuning outputs exhibit better coherence, higher accuracy, and reduced inference time. In contrast, although the Mistral model performed better than DeepSeek on some metrics, neither could match the balance of quality and efficiency seen with LLaMA 3.1 (8B).

DeepSeek LLaMA (8B) encountered challenges during fine-tuning. While it performed reasonably on transcripts it had seen during training, it struggled to generate meaningful summaries for unseen transcripts. Adjustments to training parameters, such as reducing the maximum sequence length and modifying dropout rates, did not resolve this issue. As a result, DeepSeek was not selected for further testing. This discovery underlines the importance of a model's ability to generalize beyond the training data, a factor that is critical for real-world applications.

5.2 Multilingual Analysis

After fine-tuning on English transcripts, the best model, LLaMA 3.1 (8B), was further evaluated on a multilingual dataset covering English, Hindi, and Marathi. This phase included additional metrics such as ICE, Redundancy, Abstractivity, N-gram Ratio, and Conciseness to provide deeper insights into summary quality. Table 4 presents the multilingual performance. The multilingual evaluation indicates that LLaMA 3.1 (8B) performs consistently well across all three languages. For instance, the BERT Score for Hindi (0.6991) is slightly higher than for English (0.6510) and Marathi (0.6505), suggesting that the model can capture semantic content effectively even when dealing with different linguistic structures. The ICE, Redundancy, and other metrics further confirm that the model produces clear, concise, and relevant summaries without unnecessary repetition.

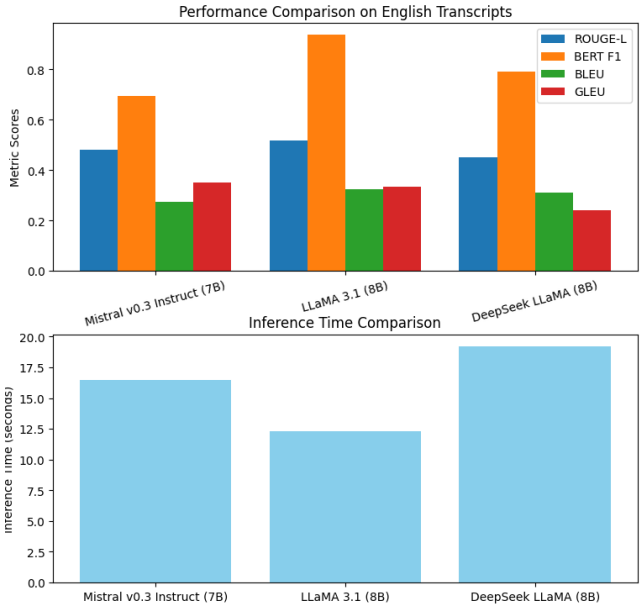

Figure 1: Performance comparison of fine-tuned models on English transcripts.

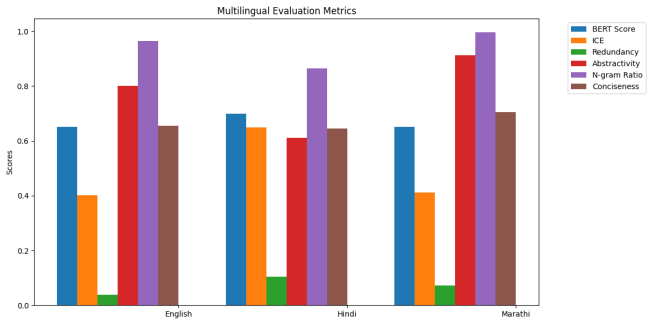

Figure 2: Multilingual Evaluation Metrics for the Selected Model across English, Hindi, and Marathi.

Figures 1 and 2 provide visual comparisons of the performance on English transcripts and the multilingual evaluation, respectively.

The improvement observed from pre-fine-tuning to post-fine-tuning is evident. The fine-tuned LLaMA 3.1 (8B) model not only generated more accurate and coherent summaries but also reduced processing time. This study confirms that fine-tuning, supported by efficient quantization techniques and targeted hyperparameter tuning, results in a significant performance boost for domain-specific

Table 4: Multilingual Evaluation of the Fine-tuned Model

Language	BERT Score	ICE	Redundancy	Abstractivity	N-gram Ratio	Conciseness
English	0.6510	0.4011	0.0389	0.8014	0.9654	0.6559
Hindi	0.6991	0.6493	0.1051	0.6119	0.8650	0.6456
Marathi	0.6505	0.4111	0.0721	0.9118	0.9960	0.7043

tasks such as career counseling transcript summarization. The successful extension to multilingual evaluation further demonstrates the model's robustness and practical applicability in diverse linguistic settings.

6 CONCLUSION

The study in this paper outlines a systematic assessment and optimization process for meeting summarization models in a multilingual setting. Three types of models were exhaustively tested using zero-shot, one-shot, and three-shot prompting methods on a specially designed dataset of career counseling meeting transcripts. ROUGE-L, BERT Score F1, BLEU, and GLEU were used as metrics to assess and conclude which model performed optimally for each type. Fine-tuning experiments with Unsloth's 4-bit quantization framework and Parameter-Efficient Fine-Tuning (PEFT) with LoRA proved that the LLaMA 3.1 (8B) model not only generated accurate summaries but also executed with outstanding efficiency and speed.

Further, using the best-performing model to work with multiple languages proved its ability to extract critical information in English, Hindi, and Marathi transcripts. The findings of the experiment highlight the need for domain-specific fine-tuning for domain-related tasks and reveal the potential of open-source LLMs in creating working and resource-efficient AI tools for career counseling.

The research leaves multiple doors open for additional work, such as additional investigation into instruction tuning versus fine-tuning, optimization for real-time summarization, and larger applications across other professional fields. Overall, the research shows a clear path from model selection to successful multilingual summarization, providing a sound basis for next-generation AI-assisted communication tools.

REFERENCES

[1] Jay Peters. Google's Meet teleconferencing service now adding about 3 million users per day — theverge.com. https://www.theverge.com/2020/4/28/21240434/google-meet-three-million-users-per-day-pichai-earnings. [Accessed 14-10-2024].

[2] Tom Warren. Zoom grows to 300 million meeting participants despite security backlash — theverge.com. https://www.theverge.com/2020/4/23/21232401/zoom-300-million-users-growth-coronavirus-pandemic-security-privacy-concerns-response. [Accessed 14-10-2024].

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[4] Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. Building real-world meeting summarization systems using large language models: A practical perspective, 2023. URL https://arxiv.org/abs/2310.19233.

[5] Fei Ge. *Fine-tune Whisper and transformer large language model for meeting summarization.* PhD thesis, UCLA, 2024.

[6] Aatman Vaidya, Tarunima Prabhakar, Denny George, and Swair Shah. Analysis of indic language capabilities in llms, 2025. URL https://arxiv.org/abs/2501.13912.

[7] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, and et. al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[8] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, and et. al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

[10] Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

[11] Nima Sadri, Bohan Zhang, and Bihan Liu. Meetsum: Transforming meeting transcript summarization using transformers!, 2021. URL https://arxiv.org/abs/2108.06310.

[12] Sumedh S Bhat, Uzair Ahmed Nawaz, Sujay M, Nameesha Tantri, and Vani Vasudevan. Jotter: An approach to summarize the formal online meeting. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, pages 1–6, 2023. doi: 10.1109/AIKIIE60097.2023.10390455.

[13] Lakshmi Prasanna Kumar and Arman Kabiri. Meeting summarization: A survey of the state of the art, 2022. URL https://arxiv.org/abs/2212.08206.

[14] Medha Wyawahare, Madhuri Shelke, Siddharth Bhorge, and Rohit Agrawal. Ai powered multilingual meeting summarization. In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 86–91, Jan 2024. doi: 10.1109/Confluence60223.2024.10463307.

[15] Aseem Srivastava, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3920–3930, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539187. URL https://doi.org/10.1145/3534678.3539187.

[16] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, K. McKeown, and Tatsunori Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2023. doi: 10.1162/tacl_a_00632.

[17] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

[19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[22] Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. Learning to score system summaries for better content selection evaluation. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4510. URL https://aclanthology.org/W17-4510/.