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 in 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. The experimental results showed that the fine-tuned LLaMA 3.1 (8B) model showed greater efficacy in both English and multilingual settings (English, Hindi, and Marathi), generating high-quality summaries, efficiency, and stable cross-lingual generalization. These findings show that using a low learning rate (1×10^{-5}), small batch sizes with gradient accumulation, and a maximum sequence length of 4096 tokens combined with Unsloth's 4-bit quantization and PEFT with LoRA helps the model achieve high accuracy while keeping computational costs low. Evaluation using metrics like ROUGE-L, BERT Score, BLEU, and GLEU, along with fast inference on a GPU P100, confirms that this approach delivers clear and high-quality summaries. This balance of performance and efficiency makes the solution scalable and practical for creating AI-based tools for career counseling.

KEYWORDS

LLM Fine-Tuning, Multilingual Summarization, Open-Source LLMs, Unsloth Fine-Tuning, LLaMA, Mistral, DeepSeek

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 helps 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 that relies 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 a 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

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	

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 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.

Recent efforts in Hindi summarization have produced helpful assessment tools guiding this paper. Singh et al. [19] presented the HindiSumm dataset together with measures like redundancy, conciseness, novel n-grams ratio, and abstractivity, which help evaluate the quality and diversity of produced summaries. Similarly, Daisy et al. [20] proposed the ICE-H metric to evaluate how well a summary covers key information in low-resource settings.

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 details the steps taken in this study. The methodology is organized into the following subsections:

4.1 Dataset Construction

A custom dataset was created to support career counseling meeting summarization. Since publicly available datasets for this domain are scarce, 35 meeting transcripts per language (English, Hindi, and Marathi) were manually curated. Each transcript underwent careful cleaning and annotation to ensure consistency. The transcripts include structured summaries, key action items, insights, and speaker details, all formatted in JSON. This structured dataset forms the reference for evaluating model performance.

4.2 Model Evaluation

For model selection, open-source large language models (LLMs) from three families Mistral, LLaMA 3, and DeepSeek as explained in Table 1, were evaluated on a custom dataset of career counseling transcripts. The evaluation was carried out under three inference settings: zero-shot, one-shot, and three-shot.

In the **zero-shot setting**, no examples are provided in the prompt. This tests the model's innate capability to generate a summary without any specific guidance. However, because no task-specific context is given, the generated summary may lack the detailed structure or clarity required for an accurate summarization.

The **one-shot setting** introduces a single example in the prompt. By providing a demonstration, the model is given a clear idea of the desired output format and content. This often improves the coherence of the summary and ensures that key details are better captured, as the model can align its output with the provided example.

In the **three-shot setting**, three examples are provided. With more demonstrations, the model can learn from multiple instances of the expected structure and content, which typically results in more consistent and accurate summaries. The use of multiple shots is particularly helpful when the task is complex or when the domain (in this case, career counseling) requires nuanced understanding.

The selection of these shot settings follows a methodology similar to that described in [16], where the advantages of in-context learning are highlighted. Using different numbers of examples allows for a systematic assessment of the model's performance under varying degrees of guidance, ultimately helping to identify the best-performing model from each family. The results of this evaluation are summarized in Table 2.

4.3 Fine-Tuning Setup

After selecting the best candidates from the initial evaluation, the next step was to fine-tune these models for the specific task of summarizing career counseling transcripts. Fine-tuning was performed on English transcripts using Unsloth's 4-bit quantization framework and Parameter-Efficient Fine-Tuning (PEFT) with LoRA [10, 18].

The fine-tuning process employed a Supervised Fine-Tuning (SFT) approach, where a pre-trained model is further adapted on task-specific data. The SFT method was chosen despite its limitations, such as the potential sensitivity to the amount of labeled data and sometimes requiring careful hyperparameter tuning because it offers a straightforward way to align the model's outputs with the structured requirements of the task. SFT has been widely adopted in recent large language model research (as discussed in [21]) and provides a reliable means to improve model performance for downstream tasks.

Key hyper-parameters for fine-tuning were set as follows:

- **Learning Rate:** 1×10^{-5}, to ensure small and precise updates.
- **Batch Size and Gradient Accumulation:** A per-device batch size of 2 with gradient accumulation over 8 steps, allowing for efficient memory use.
- **Maximum Sequence Length:** 4096 tokens, to accommodate the full context of the transcripts.
- **Optimizer and Scheduler:** The cosine learning rate scheduler and an 8-bit variant of the AdamW optimizer were used to balance performance and computational efficiency.

Training was performed on Kaggle's P100 GPU, providing the necessary computational power for fine-tuning. Although SFT has known challenges, such as sometimes not capturing long-range dependencies as effectively as other methods, it was chosen because it is well-established, relatively simple to implement, and effective for domain-specific tasks. Future work may explore alternative fine-tuning strategies to further enhance performance.

4.4 Training and Evaluation

During training, the model was evaluated using the same metrics as during model selection as explained in 4.2 along with inference time, which was measured in seconds. The training process involved generating structured JSON outputs that captured the summary, key action items, insights, and speaker names. This approach ensured the model not only learned to summarize accurately but also produced outputs that could integrate seamlessly with a dashboard for real-time use. The results of this evaluation are presented in Table 3.

4.5 Multilingual Evaluation Metrics

In addition to conventional metrics such as ROUGE [22], BLEU [23], GLEU [24], and BERT Score [25], the evaluation of summary quality in a multilingual setting (English, Hindi, and Marathi) involved several additional metrics as discussed by Singh et. al. and Daisy et. al. [19, 20]. These metrics were used to capture various aspects of the summaries and to ensure that they are informative, diverse, and succinct.

Information Coverage Estimate (ICE): ICE measures how well the generated summary captures the key information from the original transcript. It is calculated by encoding both the reference and generated summaries using Sentence-BERT and computing the cosine similarity between these embeddings. A higher ICE indicates better retention of important information.

Redundancy: Redundancy quantifies the amount of repeated content within the summary. It is defined as:

$$\text{Redundancy} = 1 - \frac{\text{Number of unique } n\text{-grams}}{\text{Total number of } n\text{-grams}} \tag{1}$$

A lower redundancy score means that the summary is more concise and free from unnecessary repetition.

Abstractivity: Abstractivity evaluates the extent to which the summary is generated using new, rephrased content rather than copying segments of the original text. It is calculated as:

$$\text{Abstractivity} = \frac{\text{Number of novel words}}{\text{Total words in summary}} \tag{2}$$

A higher abstractivity score reflects the model's ability to effectively paraphrase and generate novel expressions.

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

N-gram Ratio: The N-gram Ratio measures lexical diversity by comparing the number of novel n-grams in the summary to the total number of n-grams:

$$\text{N-gram Ratio} = \frac{\text{Number of novel } n\text{-grams}}{\text{Total } n\text{-grams in summary}} \tag{3}$$

A higher ratio indicates greater linguistic variety, showing that the model uses a richer vocabulary.

Conciseness: Conciseness is determined by comparing the length of the summary to that of the original transcript:

$$\text{Conciseness} = \frac{\text{Number of words in summary}}{\text{Number of words in original text}} \tag{4}$$

A lower value indicates that the summary is succinct, retaining only the most important content.

These extra measures offer a thorough system for assessing summary quality in a multilingual setting. They make sure that the summaries are varied, clear, and able to convey all vital information across English, Hindi, and Marathi in addition to correctness and fluency. The findings and metric values are presented in Table 4

5 OUTCOMES

The outcomes of this study are discussed in three main parts: the initial model evaluation, the fine-tuning results, and the multilingual performance analysis.

5.1 Performance of Model Families

Table 2 presents the evaluation of pre-trained large language models on English transcripts across zero-shot, one-shot, and three-shot inference settings. The evaluation metrics ROUGE, BERT Score, BLEU, and GLEU are used to determine how well each model captures essential content and maintains fluency in the generated summaries.

- ***Mistral Family*** Within the Mistral family, the base Mistral (7B) model achieves moderate scores in the zero-shot setting, while the Mistral v0.3 Instruct (7B) variant shows noticeable improvements in both one-shot and three-shot scenarios. This suggests that instruction tuning has a positive impact on its summarization capabilities, yielding higher overlap with reference summaries and better semantic alignment.

- ***LLaMA 3 Family*** For the LLaMA 3 family, three variants were tested. LLaMA 3 (8B) and LLaMA 3.2 (3B) deliver competitive results; however, LLaMA 3.1 (8B) consistently stands out. It produces the highest ROUGE scores, indicating superior content retention, and achieves the best BERT Score F1, reflecting strong semantic similarity with the reference summaries. LLaMA 3.1 (8B) shows improved BLEU and GLEU scores, which imply that the summaries are both fluent and well-structured.

- ***DeepSeek LLaMA (8B)*** It reaches competitive ROUGE and BERT Score values in the three-shot setting. Although its performance is notable, its overall scores are slightly lower compared to the top variants from the LLaMA and Mistral families.

Based on the results, the following models were chosen for further fine-tuning:

(1) **Mistral family:** Mistral v0.3 Instruct (7B)
(2) **LLaMA 3 family:** LLaMA 3.1 (8B)
(3) **DeepSeek family:** DeepSeek LLaMA (8B)

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

Model	ROUGE-L	BERT F1	BLEU	GLEU	Inference Time
LLaMA 3.1 (8B)	0.5178	0.9378	0.3253	0.3334	12.3s
Mistral v0.3 Instruct (7B)	0.4796	0.9350	0.2745	0.3496	16.5s
DeepSeek LLaMA (8B)	0.4527	0.7903	0.3103	0.2396	19.2s

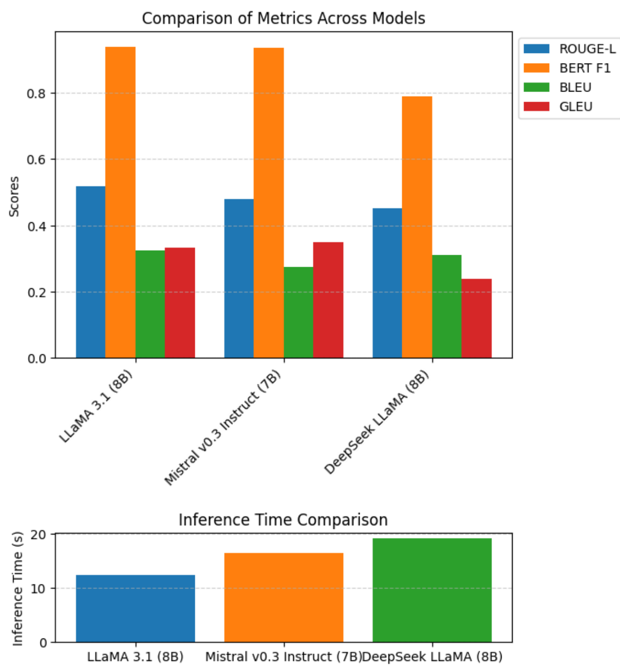

Figure 1: Performance comparison of fine-tuned models on English transcripts.

5.2 Results of Fine-Tuning Process and Comparative Analysis

The fine-tuning was executed as outlined in Section 4.3. Table 3 shows a detailed comparison of the three fine-tuned models, and Figure 1 shows a graphical comparison on English meeting transcripts.

- **LLaMA 3.1 (8B):** This model achieved a ROUGE-L score of 0.5178 and a BERT Score F1 of 0.9378. It also reached a BLEU of 0.3253 and a GLEU of 0.3334. These results reflect an improvement of approximately 7% in ROUGE-L and more than 35% in BERT Score F1 compared to some pre-fine-tuning results.
- **Mistral v0.3 Instruct (7B):** This model recorded a ROUGE-L of 0.4796 and a BERT Score F1 of 0.9350 along with a BLEU of 0.2745 and a GLEU of 0.3496.
- **DeepSeek LLaMA (8B):** While DeepSeek LLaMA (8B) achieved a moderate BLEU score of 0.3103, its overall performance was lower, with a ROUGE-L of 0.4527 and a BERT Score F1 of 0.7903.

5.2.1 Inference Time Trade-offs. Practical uses also depend on inference time, which is as important as correctness. With an average time of 12.3 seconds per transcript, LLaMA 3.1 (8B) not only offered the best accuracy, but also attained the fastest inference. In contrast, DeepSeek LLaMA (8B) was the slowest at 19.2 seconds, while Mistral v0.3 Instruct (7B) needed 16.5 seconds. This trade-off between speed and accuracy is significant; quicker inference allows real-time summarization, which is absolutely vital for interactive systems. LLaMA 3.1 (8B) is the most practical option for deployment given the balance of high performance and low inference time.

DeepSeek LLaMA (8B) encountered challenges during fine-tuning. While it performed reasonably well 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.

The comparison shows that the summary quality improved significantly as a result of fine-tuning. Compared to their pre-fine-tuning outputs presented in Table 2, the models produced more accurate and coherent summaries with faster processing times. The metrics make it evident that accuracy and speed must be traded off; LLaMA 3.1 (8B) provides the fastest inference time and high accuracy (with notable improvements in ROUGE-L and BERT Score F1), achieving the best overall balance. This led to LLaMA 3.1 being chosen for further multilingual fine-tuning and analysis.

5.3 Multilingual Performance Evaluation

After identifying LLaMA 3.1 (8B) as the top model on English transcripts, this model was further fine-tuned on an expanded dataset that includes English, Hindi and Marathi transcripts as explained in Section 4.1. A detailed analysis of the performance of the model in these languages is presented in Table 4 and Figure 2 using the metrics explained in Section 4.5.

The evaluation shows that the model delivers consistent performance in all languages. For example, Hindi transcripts achieved a slightly higher BERT Score (0.7281) than English (0.6533) and Marathi (0.6807), indicating a strong ability to capture semantic meaning. Low Redundancy values confirm that the summaries avoid repetitive content, while high Abstractivity scores demonstrate the model's ability to paraphrase and generate novel expressions while retaining essential information. Furthermore, the elevated N-gram Ratio and solid Conciseness scores attest that the summaries are both varied in vocabulary and succinct.

Table 4: Multilingual Evaluation of the Fine-tuned Model

Language	BERT Score	ICE	Redundancy	Abstractivity	N-gram Ratio	Conciseness
English	0.6533	0.4043	0.0347	0.8007	0.9690	0.6788
Hindi	0.7281	0.7702	0.1064	0.5361	0.8113	0.6203
Marathi	0.6807	0.5688	0.0677	0.8207	0.9663	0.6649

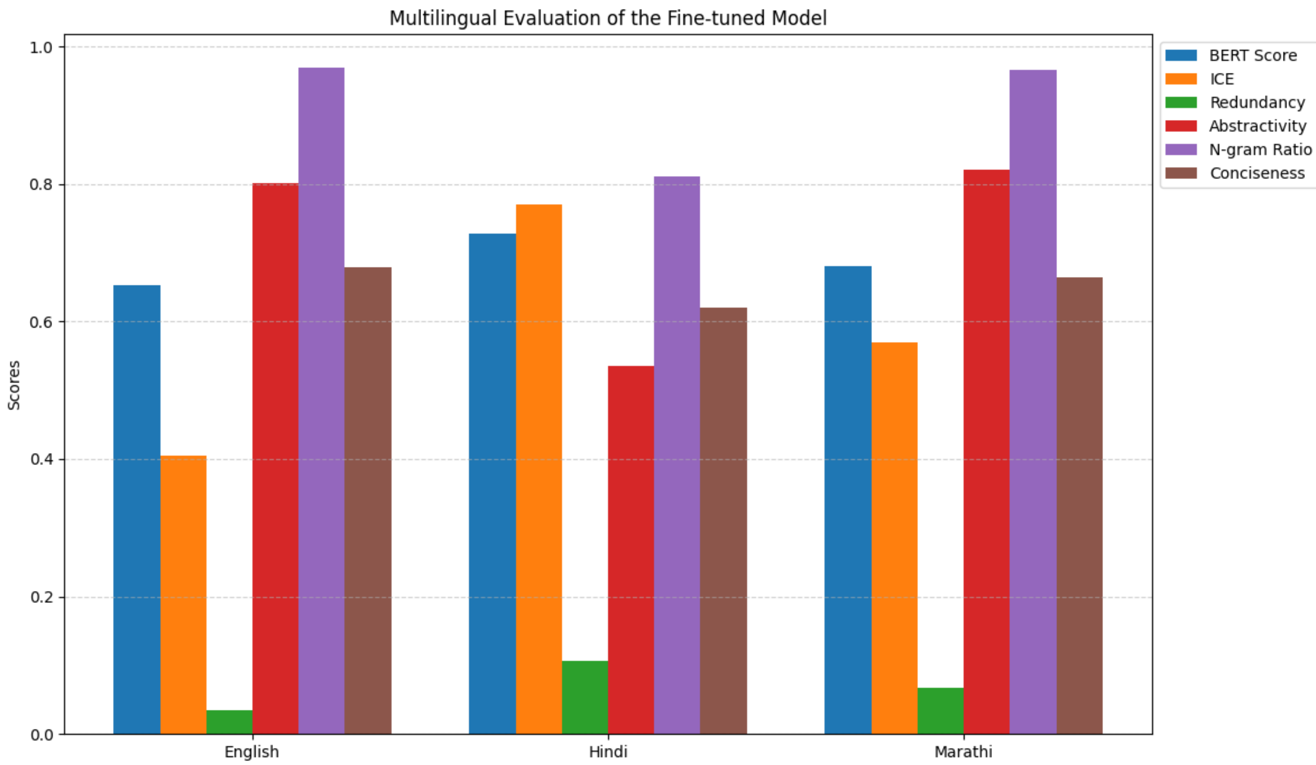

Figure 2: Multilingual Evaluation Metrics for LLaMA 3.1 (8B) across English, Hindi, and Marathi transcripts

6 CONCLUSION

The research in this paper presents a systematic evaluation and optimization procedure for meeting summarization models in a multilingual environment. Three model types were thoroughly tested with zero-shot, one-shot, and three-shot prompting techniques on a specially created dataset of career guidance meeting transcripts. ROUGE-L, BERT Score F1, BLEU, and GLEU were used as metrics to evaluate and determine which model worked best for each type. Fine-tuning experiments with Unsloth's 4-bit quantization architecture and Parameter-Efficient Fine-Tuning (PEFT) with LoRA validated that the LLaMA 3.1 (8B) model not only produced correct summaries, but also performed with exceptional efficiency and pace.

Furthermore, applying the best-performing model to handle multiple languages demonstrated its ability to extract important information in English, Hindi, and Marathi transcripts. The results of the experiment emphasize the requirement of domain-specific fine-tuning for domain-related tasks and indicate the potential of open-source LLMs to develop working and resource-effective AI tools for career guidance.

Future research could investigate comparative tests between supervised fine-tuning and reinforcement learning or reward-based optimization techniques to optimize structured, customized summarization. Future research can also examine trade-offs between instruction tuning and fine-tuning, optimization for live summarization, and more pervasive applications in other professional domains. In general, the study reflects a clear methodology from model choice to effective multilingual summarization, and with it, establishes a solid foundation for next-generation AI-powered 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] Geetanjali Singh, Namita Mittal, and Satyendra Singh Chouhan. Hindisumm: A hindi abstractive summarization benchmark dataset. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(12), November 2024. ISSN 2375-4699. doi: 10.1145/3696207. URL https://doi.org/10.1145/3696207.
[20] Daisy Monika Lal, Paul Rayson, Krishna Pratap Singh, and Uma Shanker Tiwary. Abstractive Hindi text summarization: A challenge in a low-resource setting. In Jyoti D. Pawar and Sobha Lalitha Devi, editors, *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 603–612, Goa University, Goa, India, December 2023. NLP Association of India (NLPAI). URL https://aclanthology.org/2023.icon-1.58/.
[21] Tong Xiao and Jingbo Zhu. Foundations of large language models, 2025. URL https://arxiv.org/abs/2501.09223.
[22] 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/.
[23] 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/.
[24] 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/.
[25] 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.