

Comparative Evaluation of Pre-trained AI Models for Automated Meeting Summarization and Action Item Extraction

Assignment 3 Report

Submitted by Name: Mohammed Omar Abdullah Alqathmi

Lecturer: Dr. Mohammed Anbar

Assignment 3 Report: Comparative Evaluation of Pre-trained AI Models for Automated Meeting Summarization and Action Item Extraction

Course Title: Research Methodology in Computer Science (AI800)

Lecturer: Dr. Mohammed Anbar

Hand-out Date: 15 November 2025

Hand-in Date: 20 December 2025

Hand-in Method: Online Submission via E-Learning

Submitted by Name: Mohammed Omar Abdullah Alqathmi

Email: 253001437@st.midocean.edu.km

Student Number: 253001437

Table of Contents

Abstract	1
Chapter 1: Introduction	1
- Introduction	1
- Background	2
- Problem Statement.....	3
- Research Objectives	5
- Research Scope	6
Chapter 2: Literature Review	7
Chapter 3: Research Methodology	15
- Research Methodology	11
- Expected Outcomes	13
- Gantt Chart.....	15
References	17

Abstract

This research evaluates the performance of pre-trained AI models for automated meeting summarization and action item extraction. Large volumes of recorded content is generated by Virtual meetings, however manual review is time-consuming and error-prone. 3-4 hours per week are spent by employees to review meeting notes, and 20-30% of action items are forgotten or delayed. This study addresses these challenges by comparing open-source transformer models (BART, PEGASUS, T5, LED) against commercial API-based models (GPT-4, Claude, Gemini) using public datasets like AMI and ICSI. The evaluation uses ROUGE, BERTScore, and qualitative metrics to assess summarization quality and action item extraction accuracy. The research also analyzes computational costs, API pricing, deployment complexity, and the impact of different prompting strategies. An application will be developed to demonstrate practical implementation. Outcomes expected include a comparison of the framework, deployment guidelines and insights into cost and performance tradeoffs across different models.

Keywords: Meeting Summarization, Action Item Extraction, Transformer Models, Large Language Models, BART, PEGASUS, T5, GPT-4

Chapter 1: Introduction

Introduction

Meetings are increasingly being held online using various platforms like Zoom, Google Meet, Teams etc. This has become standard practice today across organizations. Thousands of meetings are recorded monthly and they contain important decisions, action items, and assigned responsibilities. Reviewing meetings manually is time consuming and error prone. Studies show that 3-4 hours per week is spent by employees to review meeting notes. Manual extraction leads to missed action items and unclear follow-ups.

The problem becomes worse for people who miss meetings or new employees who need past meeting context. Meeting transcripts include irrelevant content, repeated discussions, and off-topic conversations. Understanding context, speaker turns, and conversation flow is essential for extracting clean summaries and structured action items. These tasks are hard to automate effectively.

Recent advances in transformer-based models and large language models show promise for automated meeting analysis. Models like BART, PEGASUS, T5, and LED have demonstrated strong performance on summarization tasks. Commercial APIs like GPT-4, Claude, and Gemini

offer powerful alternatives with minimal setup. However, a systematic comparison of these approaches on meeting datasets is missing.

This research aims to fill this gap by conducting a practical, comparative evaluation of both open-source and commercial models for meeting summarization and action item extraction. The study will assess quality metrics, computational requirements, costs, and deployment considerations to provide actionable guidance for practitioners.

Background

Evolution of Meeting Summarization

Summarization of meetings are increasingly using automated AI driven analysis evolving from simple note-taking activity. Early approaches relied on extractive methods. This approach selected important sentences from transcripts. These methods worked well for structured text. However they struggled with informal language common in meetings, conversational speech patterns and speaker interruptions.

The introduction of transformer architectures in 2017 changed the landscape. Models like BERT enabled better contextual understanding of language. Specialized summarization models like BART (2019), T5 (2020), PEGASUS (2020), LED (2020) etc were subsequently developed. These newer models trained on large text corpora using encoder-decoder architectures. This enabled abstractive summarization that generates new sentences rather than just extracting existing ones.

Large Language Models and APIs

The arrival of LLMs like GPT-3, Claude and Gemini introduced a new paradigm. These models were trained on massive datasets. They can perform summarization through prompting without fine-tuning. Commercial APIs make these models accessible without requiring specialized hardware or ML expertise. This democratizes access but introduces concerns about cost, data privacy, and dependency on external services.

Evaluation Challenges

Evaluating meeting summarization quality remains challenging. Traditional metrics like ROUGE measure word overlap with reference summaries but may not capture semantic quality. Recent research questions the reliability of automatic metrics, showing weak correlation with human judgment. Action item extraction adds another layer of complexity, requiring identification of tasks, assignees, and deadlines from conversational text.

Current State

Most research focuses on individual models or specific datasets and as a result, despite progress, several gaps remain. Systematic comparisons across multiple model types using consistent evaluation criteria are rare. Practical considerations like computational cost, deployment complexity, and real-world usability receive limited attention. This research addresses these gaps through comprehensive evaluation.

Problem Statement

Problem Domain

Automated meeting summarization and action item extraction in virtual collaboration environments.

Problem Description

Thousands of meetings are recorded every month by various organizations on platforms such as Zoom, Teams and Google Meet. These meetings contain important decisions, actions to be taken and responsibilities assigned. Reviewing these long recordings manually is slow and prone to human errors.

Studies show that employees spend 3–4 hours per week reviewing meeting notes (Singh et al., 2023), and manual extraction often leads to missed action items and unclear follow-ups (Golia & Kalita, 2023). The problem becomes worse for people who miss meetings or new employees who rely on past meeting context.

Meeting transcripts often include irrelevant content and repeated discussions, off topic conversations. Understanding context, speaker turns and flow of conversation is essential for extracting clean summaries and structured action items. These tasks are generally hard to automate.

Significance of the Problem

Poor summarization / documentation of meetings leads to 20-30% of action items being forgotten or delayed resulting in repeated discussions and leads to productivity loss and missed deadlines.

Saving at least 1 hour per employee in a week through automation results in savings of over 5,200 hours per year in an organization of about 100 employees.

Automated summaries also improve knowledge sharing by helping new or remote team members quickly understand past decisions without watching long recordings.

Existing Solutions

- **Transformer-based Summarization (Golia & Kalita, 2023):** BART with recursive summarization improved meeting action-item extraction, achieving a BERTScore of 64.98 on AMI.
- **Multimodal Meeting Minutes System (Singh et al., 2023):** Whisper AI + NLP achieved 92% key point and 87% action item accuracy, producing minutes 5× faster than manual writing.
- **Evaluation Metrics Analysis (Kirstein et al., 2024):** Showed that many summarization metrics are unreliable, with weak correlation to human judgment ($\alpha \approx 0.79$).

Research Gaps

Despite progress, key issues remain:

- **High computational requirements:** Models like BART and PEGASUS need powerful GPUs, limiting accessibility.
- **Weak generalization:** Models trained on AMI or clean datasets struggle with noisy, real-world meetings (Singh et al., 2023).
- **Lack of fair comparison across models:** No consistent benchmarking of common pre-trained models (BART, T5, PEGASUS, LED) using the same dataset.
- **Action item extraction is still difficult:** Many models summarize well but fail to extract structured tasks (owner, deadline, action).
- **Accessibility gap:** Most research focuses on model design, not on easy, practical deployment for students or small teams.

Because of these gaps, a clear, practical comparison of readily available summarization models is needed.

Research Objectives

Primary Goal

Evaluation and comparison of the performance of existing pre-trained AI models for meeting summarization and action item extraction using publicly available datasets.

Specific Objectives

1. To identify and prepare suitable public meeting transcript datasets (AMI, ICSI) with proper preprocessing such as speaker normalization and text cleaning.
2. To implement summarization using two approaches:
 - Open-source models including BART-base, PEGASUS-base, T5-small, LED-base using HuggingFace.
 - Commercial APIs including GPT-4, Claude, Gemini with structured prompts.
3. To evaluate summarization quality using:
 - ROUGE and BERTScore (where references exist)
 - qualitative review of clarity and correctness
 - action item extraction accuracy
4. To implement action item extraction using:
 - prompt-based or fine-tuned sequence labeling (open-source)
 - structured JSON prompting (API models)
5. To compare practical aspects of each approach including:
 - computational requirements for open-source models
 - cost, response time, and context limits for API models
 - ease of setup and deployment
6. To analyze the impact of different prompting strategies including but not limited to zero-shot, few-shot, chain-of-thought on summarization and action item extraction.
7. To implement a simple demo application (Streamlit/Gradio) that:
 - accepts meeting transcripts
 - allows model/method selection
 - generates summaries and action items
 - displays quality metrics and cost/time info
 - provides downloadable outputs

Research Scope

This research focuses on the following areas:

Dataset Selection: The study will use publicly available meeting transcript datasets, specifically AMI and ICSI corpora. These datasets provide annotated meeting transcripts with reference summaries and action items, enabling quantitative evaluation. The research will not collect new meeting data or create new datasets.

Model Coverage: Evaluation will include open-source transformer models (BART-base, PEGASUS-base, T5-small, LED-base) and commercial API-based models (GPT-5, Claude, Gemini). These models are commonly used and accessible rather than experimental or proprietary architectures.

Evaluation Metrics: Assessment will combine automatic metrics (ROUGE, BERTScore) with qualitative analysis of summary quality and action item extraction accuracy. The research will not conduct large-scale human evaluation studies but will include manual review of sample outputs.

Practical Focus: The study emphasizes practical deployment considerations including computational requirements, API costs, setup complexity, and response times. This addresses real-world usability for small teams and students rather than purely theoretical performance optimization.

Prompting Strategies: Analysis will cover zero-shot, few-shot, and chain-of-thought prompting approaches for API-based models. The research will document the impact of different prompt designs on output quality and extraction accuracy.

Demo Application: A functional demo application will demonstrate practical implementation. The application will showcase model integration, user interaction, and output generation but will not include production-ready features like user authentication, persistent storage, or enterprise security.

Limitations: The research will not address real-time summarization, multi-language support, or audio processing. The focus remains on English text transcripts. The study will not develop new models or architectures but will evaluate existing ones.

Chapter 2: Literature Review

This chapter reviews existing research on meeting summarization and engagement analysis. The first section covers six papers from initial literature review, followed by four additional recent studies that provide deeper insights into model comparisons and LLM-based approaches.

Paper 1: Action-Item-Driven Summarization of Long Meeting Transcripts (Golia & Kalita, 2023)

This paper focuses on extracting actionable content from condensed speech. The paper highlights that meetings are not dialogues, they are structured discussions with goals, decisions and responsibilities.

They propose a recursive summarization approach combined with an action-item extraction algorithm using BART transformers. By segmenting meeting transcripts into topic-based sections, the model reduces memory loss common in long-context models. The system was evaluated on the AMI corpus and achieved a BERTScore of 64.98, improving by nearly 5% over existing baselines.

While effective, the model's reliance on synthetic meeting data limits generalization to real corporate scenarios. Moreover, processing multiple transcript sections increases computational cost, making large-scale deployment challenging.

🔗 <https://arxiv.org/abs/2312.17581>

Paper 2: Investigating Automatic Metrics on Meeting Summarization (Kirstein et al., 2024)

This research evaluates the reliability of metrics used to measure meeting summarization quality. The team applied PRISMA review methodology, developing an eight-category taxonomy of possible summarization errors—ranging from redundancy to hallucination. Using 35 annotated QMSum samples, they assessed five leading models (LED, DialogLED, PEGASUS-X, GPT-3.5 turbo, and Zephyr-7B- α) across nine metrics.

Results revealed that one-third of metric–error combinations either ignored errors or rewarded them. Metrics like Perplexity often favored grammatically fluent but semantically incorrect summaries. (Krippendorff's $\alpha \approx 0.79$), meaning Human correlation remained weak suggests evaluation benchmarks may not align with actual human judgment.

In spite of limited dataset and English-only scope, the study underscores a critical issue: the field may be optimizing for the wrong targets, mistaking surface-level coherence for comprehension.

🔗 <https://arxiv.org/pdf/2404.11124.pdf>

Paper 3: The Application of Real-time Emotion Recognition in Video Conferencing (Lin, 2024)

Lin's paper explores how AI can improve the emotional quality of virtual meetings. The proposed framework is integrated with CNN-based facial feature extraction based on LSTM and TCN architectures to analyze emotion sequences over a period of time.

Edge computing handles almost all processing locally leading to significantly reduced latency and assured real-time responsiveness. Multi-modal fusion (combining facial, vocal, and physiological cues) enables enhanced emotion detection accuracy while Security is strengthened through homomorphic encryption and federated learning.

However, continuous emotion tracking raises significant privacy concerns. Differences in environmental conditions such as lighting, camera angle and even cultural expression differences reduce reliability. Despite these challenges, the study shows that emotion-aware conferencing can significantly improve engagement and communication quality.

❓ <https://www.suaspress.org/ojs/index.php/JCTAM/article/view/v1n4a10>

Paper 4: Automated Minutes of Meeting Using a Multimodal Approach (Singh et al., 2023)

This work targets a practical business problem—automating meeting minutes generation. Using Whisper AI for speech-to-text transcription, the pipeline integrates NLP techniques for text normalization, topic extraction, and speaker diarization.

Results were impressive: 92% of key points and 87% of action items were identified correctly, with a Word Error Rate under 10% for clear audio. The system produced minutes five times faster than humans.

This solution does not work well in noisy environments or when transcription contains domain-specific jargon and overlapping conversations. It however works well in structured, well-recorded meetings, limiting its flexibility in diverse conditions.

❓ <https://www.ijraset.com/research-paper/automated-minutes-of-meeting-using-a-multimodal-approach>

Paper 5: Facial Expression Recognition using Visible and IR by Early Fusion (Multiple Authors, 2025)

This paper explores dual-modal facial expression recognition using both visible and infrared (IR) imaging. The idea is that visible light captures surface-level expressions, while infrared captures subtle temperature shifts linked to emotional states.

The authors enhance ResNet-18 with an attention mechanism and apply early fusion to combine both modalities. Tested on VIRI and NVIE datasets, the model achieved 84.44% and 85.20% accuracy, outperforming prior methods by over 10%.

However, small datasets and hardware dependency (IR cameras) limit scalability. Performance drops result in occlusions (masks, glasses). poor lighting remains unresolved. In spite of these issues, the study points toward richer, more reliable emotion detection in hybrid environments.

🔗 <https://peerj.com/articles/cs-2676/>

Paper 6: Automatic Student Engagement Recognition in Online Classrooms (Multiple Authors, 2025)

Focusing on virtual learning, this paper presents a model for detecting student engagement using multi-scale CNNs and attention fusion modules. Spatial interaction modeling captures relationships between the student, environment, and device.

The model achieved 88% accuracy for highly engaged students and maintained 20–25 FPS in real time. The study demonstrates that engagement recognition can work effectively with quality video data.

Obstacles include a requirement for higher resolution videos ($\geq 480p$), limited scalability with larger classes, and concerns related to privacy because of continuous monitoring. Despite these drawbacks, this research shows how computer vision can bridge the observation gap in remote education.

🔗 <https://www.iijeta.org/journals/ts/paper/10.18280/ts.420428>

Paper 7: Summaries, Highlights, and Action Items: Design, Implementation and Evaluation of an LLM-powered Meeting Recap System (Asthana et al., 2024)

This paper presents a production-level meeting recap system deployed at Microsoft. The authors designed two recap formats: Highlights (concise bullet points of outcomes) and Hierarchical Minutes (structured meeting notes with nested sections). The system uses BART

for extractive summarization and deBERTa for identifying important segments. Seven Microsoft users evaluated the system on their real meetings. Results showed that Highlights format was preferred for quick reviews while Hierarchical format worked better for detailed reference. Users reported that action item extraction was accurate but occasionally missed context about who was responsible. The study provides valuable insights into user preferences and practical deployment challenges. However, evaluation was limited to internal Microsoft users and meetings, which may not generalize to other organizations. The system requires access to meeting transcripts and does not handle audio processing directly.

🔗 <https://arxiv.org/abs/2307.15793>

Paper 8: Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction (Dharrao et al., 2024)

This study directly compares three popular transformer models (BART, T5, PEGASUS) on business news summarization using BBC News articles. ROUGE and METEOR metrics were used to evaluate model performance across different text lengths and domains. The highest score was achieved by T5, followed by BART and PEGASUS. T5-small showed better efficiency with comparable performance to larger variants. PEGASUS struggled with domain-specific terminology in business contexts. The study found that all models performed worse on longer articles (>1000 words), suggesting context length limitations. While this research focuses on news rather than meetings, the model comparison methodology and findings are relevant for meeting summarization tasks. The limitation is that business news has different structure and language patterns compared to conversational meeting transcripts.

🔗 <https://www.iieta.org/journals/ria/paper/10.18280/ria.380311>

Paper 9: Re-FRAME the Meeting Summarization SCOPE: Fact-Based Summarization and Personalization via Questions (Kirstein et al., 2025)

This recent work introduces FRAME, a pipeline that improves meeting summarization through fact-based approaches. The system extracts atomic facts from transcripts, organizes them into structured notes, and enriches content using GPT-4o and Gemini 1.5 Pro. The authors also present SCOPE, a protocol that generates personalized summaries by asking users about their information needs. Evaluated on QMSum and FAME datasets, FRAME reduced hallucination errors from 3 to 1, omission errors from 2 to 1, and irrelevance errors from 2 to 1 compared to baseline approaches. The study introduces P-MESA, a framework for evaluating personalization quality. Results show that user-specific summaries improve satisfaction and information retrieval. The main limitation is computational cost, as fact extraction and enrichment require multiple LLM calls. The approach also assumes users can articulate their information needs clearly.

🔗 <https://arxiv.org/abs/2509.15901>

Paper 10: Meeting Delegate: Benchmarking LLMs on Attending Meetings on Our Behalf (Hu et al., 2025)

This paper evaluates whether LLMs can effectively "attend" meetings and answer questions on behalf of absent participants. The authors tested GPT-4, GPT-4o, Gemini 1.5 Pro, Gemini 1.5 Flash, and Llama3 (8B and 70B) on 846 test cases from ELITR Minuting Corpus. The benchmark includes four scenarios: understanding discussions, identifying action items, answering specific questions, and generating summaries. Results showed that about 60% of responses successfully captured key points, with GPT-4o and Gemini 1.5 Pro performing best. Smaller models like Llama3-8B struggled with complex multi-turn discussions. The study found that chain-of-thought prompting improved performance by 12-15% across all models. Action item extraction remained challenging, with accuracy around 70% even for top models. This research demonstrates that LLMs show promise for meeting delegation tasks but are not yet reliable enough for critical business scenarios.

🔗 <https://arxiv.org/abs/2502.04376>

Chapter 3: Research Methodology

Research Methodology

This research uses a practical, experimental approach to compare pre-trained AI models for meeting summarization and action item extraction. The methodology consists of five phases executed sequentially.

Phase 1: Dataset Preparation

The first phase involves selecting and preparing meeting transcript datasets. The research will use two publicly available corpora: AMI (Augmented Multi-party Interaction) and ICSI (International Computer Science Institute) meeting datasets. These datasets contain annotated meeting transcripts with reference summaries and action items.

Data preprocessing includes speaker normalization (standardizing speaker labels), removing filler words and repetitions, and segmenting long meetings into manageable chunks. Each dataset will be split into training (if fine-tuning is required) and test sets following standard protocols. Basic statistics about meeting length, speaker count, and vocabulary will be documented.

Phase 2: Model Implementation

The second phase implements two categories of models: open-source transformers and commercial API-based models.

Open-source models include BART-base, PEGASUS-base, T5-small, and LED-base from HuggingFace. These models will be used in two configurations: zero-shot (pretrained only) and fine-tuned on meeting data. Implementation will use Python with PyTorch and HuggingFace Transformers library. GPU resources will be documented including memory requirements and processing time.

Commercial API models include GPT-4, Claude, and Gemini. These will be accessed via official APIs using structured prompts. Three prompting strategies will be tested: zero-shot (simple instruction), few-shot (with examples), and chain-of-thought (step-by-step reasoning). All prompts will be documented for reproducibility.

Phase 3: Evaluation

The third phase evaluates model outputs using multiple metrics. Automatic metrics include ROUGE-1, ROUGE-2, ROUGE-L for summary quality and BERTScore for semantic similarity. These metrics will be computed where reference summaries are available.

Qualitative evaluation involves manual review of sample outputs. A small set of generated summaries (20-30 per model) will be assessed for clarity, completeness, and correctness. Action item extraction will be evaluated by checking whether extracted items match reference annotations, measuring precision, recall, and F1 scores.

Phase 4: Comparative Analysis

The fourth phase compares models across practical dimensions. For open-source models, computational requirements (GPU memory, processing time per transcript) will be measured. Setup complexity (installation steps, dependencies) will be documented.

For API models, cost per summary will be calculated based on token usage and current pricing. Response time and context length limits will be recorded. The analysis will compare ease of use, including API documentation quality and error handling.

Phase 5: Demo Application Development

The final phase implements a web-based demo application using Streamlit or Gradio. The application will allow users to upload meeting transcripts, select a model (open-source or API), configure parameters, and generate summaries with action items. The interface will display quality metrics and cost/time information. Users can download outputs in JSON or text format.

Tools and Technologies

- **Programming:** Python 3.9+
- **Libraries:** HuggingFace Transformers, PyTorch, ROUGE, BERTScore
- **APIs:** OpenAI (GPT-4), Anthropic (Claude), Google (Gemini)
- **Development:** Streamlit/Gradio for demo, Jupyter for experimentation
- **Infrastructure:** GPU access (Google Colab or local), API credits

Ethical Considerations

All datasets used are publicly available with appropriate licenses. No new human data will be collected. API usage will follow provider terms of service. Generated summaries will be evaluated for potential bias or harmful content.

Limitations

The research is limited to English transcripts and does not address multi-language scenarios. Real-time processing is not evaluated. The study does not develop new models but evaluates existing ones. Computational resources may limit the extent of fine-tuning experiments.

Expected Outcomes

Comprehensive Model Comparison: A detailed comparison of seven models (four open-source, three commercial) on meeting summarization tasks. The comparison will include quantitative metrics (ROUGE, BERTScore, F1 for action items) and qualitative assessment of summary quality. This outcome will help practitioners choose appropriate models based on their requirements.

Performance Analysis Report: Documentation of model performance across different meeting types, lengths, and complexity levels. The report will identify which models handle long context better, which extract action items more accurately, and how performance varies with meeting characteristics.

Cost-Benefit Analysis: Clear documentation of computational costs for open-source models (GPU hours, memory requirements) and monetary costs for API models (price per summary, token usage). This analysis will enable informed decisions about deployment options based on budget and scale.

Prompting Strategy Guidelines: Analysis of how different prompting approaches (zero-shot, few-shot, chain-of-thought) affect output quality for API-based models. This outcome will provide practical guidance for prompt engineering in meeting summarization contexts.

Deployment Guidelines: Practical documentation covering setup procedures, dependency management, API integration, and troubleshooting for each model type. These guidelines will reduce barriers to implementation for students and small teams.

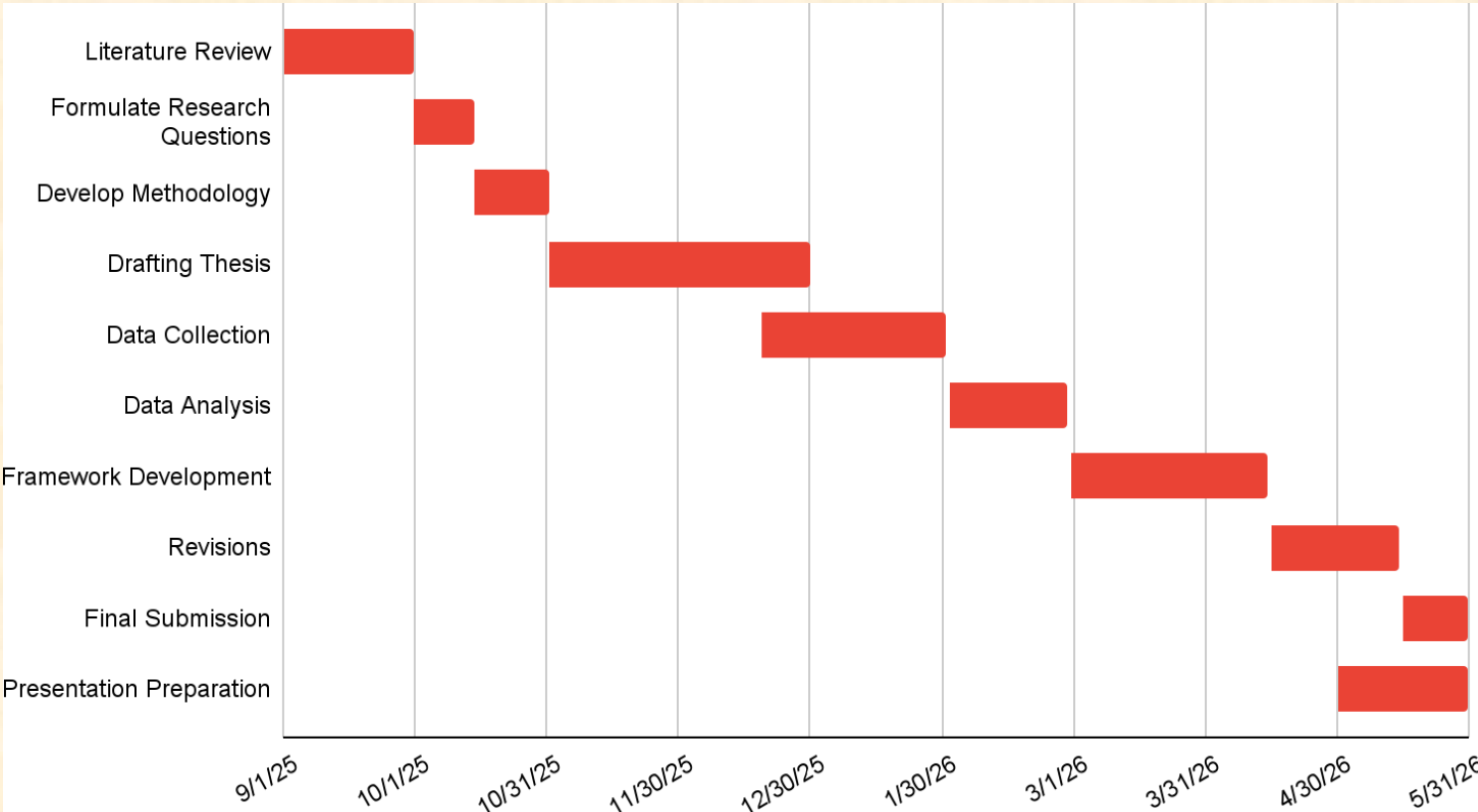
Demo Application: A functional web application demonstrating practical implementation of multiple models. The application will serve as a reference implementation and teaching tool for meeting summarization systems.

Research Publication: Findings will be compiled into a research paper suitable for submission to conferences or journals in NLP or AI applications domains. The paper will contribute to academic discourse on practical AI deployment.

Open-Source Contribution: Code, prompts, and evaluation scripts will be shared in a public repository to support reproducibility and enable other researchers to build on this work.

Knowledge Transfer: Documentation and tutorials will help bridge the gap between research prototypes and production-ready systems, particularly benefiting students and small organizations with limited ML expertise.

Gantt Chart



Task	Start Date	Duration
Literature Review	01-Sep-2025	1 Month
Formulate Research Questions	01-Oct-2025	2 Weeks
Develop Methodology	15-Oct-2025	2 Weeks
Drafting Thesis	01-Nov-2025	2 Months
Data Collection	20-Dec-2025	6 Weeks
Data Analysis	01-Feb-2026	1 Month
Framework Development	01-Mar-2026	1.5 Months
Revisions	16-Apr-2026	1 Month
Final Submission	16-May-2026	2 Weeks
Presentation Preparation	01-May-2026	1 Month

References

- [1] L. Golia and J. Kalita, "Action-Item-Driven Summarization of Long Meeting Transcripts," *arXiv preprint arXiv:2312.17581*, 2023. [Online]. Available: <https://arxiv.org/abs/2312.17581>
- [2] F. Kirstein, J. P. Wahle, T. Ruas, and B. Gipp, "Investigating Automatic Metrics on Meeting Summarization," *arXiv preprint arXiv:2404.11124*, 2024. [Online]. Available: <https://arxiv.org/pdf/2404.11124.pdf>
- [3] W. Lin, "The Application of Real-time Emotion Recognition in Video Conferencing," *Journal of Computer Technology and Applied Mathematics*, vol. 1, no. 4, pp. 79-88, 2024. [Online]. Available: <https://www.suaspress.org/ojs/index.php/JCTAM/article/view/v1n4a10>
- [4] A. Singh et al., "Automated Minutes of Meeting Using a Multimodal Approach," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 12, 2023. [Online]. Available: <https://www.ijraset.com/research-paper/automated-minutes-of-meeting-using-a-multimodal-approach>
- [5] Multiple authors, "Facial Expression Recognition using Visible and IR by Early Fusion of Deep Learning with Attention Mechanism," *PeerJ Computer Science*, e-2676, 2025. [Online]. Available: <https://peerj.com/articles/cs-2676/>
- [6] Multiple authors, "Automatic Student Engagement Recognition in Online Classrooms Based on Deep Image Segmentation and Attention Mechanisms," *Traitement du Signal*, vol. 42, no. 4, 2025. [Online]. Available: <https://www.iieta.org/journals/ts/paper/10.18280/ts.420428>
- [7] S. Asthana et al., "Summaries, Highlights, and Action Items: Design, Implementation and Evaluation of an LLM-powered Meeting Recap System," *Proc. ACM Hum.-Comput. Interact.*, vol. 9, no. CSCW1, 2025. [Online]. Available: <https://arxiv.org/abs/2307.15793>
- [8] S. Dharrao, S. Mishra, and S. Gupta, "Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction," *Revue d'Intelligence Artificielle*, vol. 38, no. 3, pp. 943-950, 2024. [Online]. Available: <https://www.iieta.org/journals/ria/paper/10.18280/ria.380311>
- [9] F. Kirstein, T. Ruas, J. P. Wahle, and B. Gipp, "Re-FRAME the Meeting Summarization SCOPE: Fact-Based Summarization and Personalization via Questions," *arXiv preprint arXiv:2509.15901*, 2025. [Online]. Available: <https://arxiv.org/abs/2509.15901>
- [10] W. Hu, K. Yuan, K. Zhang, Y. Wang, and Y. Zhang, "Meeting Delegate: Benchmarking LLMs on Attending Meetings on Our Behalf," in *Proc. ACL Workshop on Human-Centered NLP*, 2025. [Online]. Available: <https://arxiv.org/abs/2502.04376>