Tell Me What I Need to Know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization

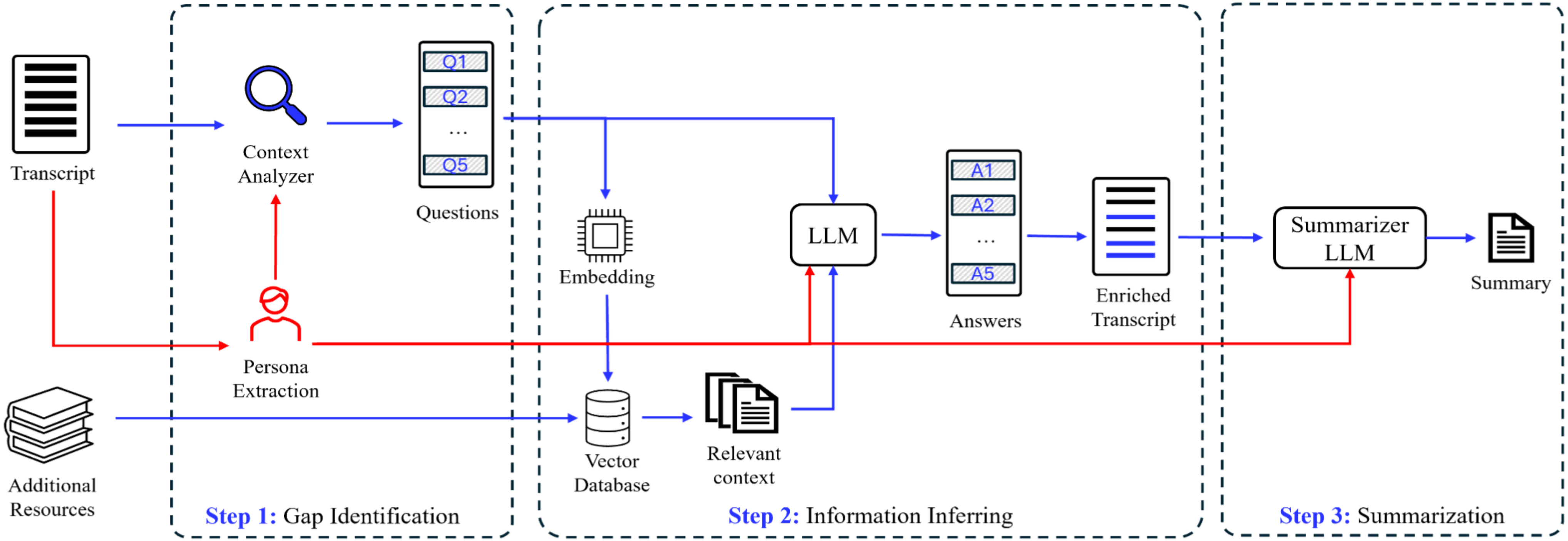
Frederic Kirstein, Terry Ruas, Robert Kratel & Bela Gipp
University of Göttingen, Germany

EMNLP 2024 Industry Track

Current summarization systems rely only on the transcript to produce general summaries

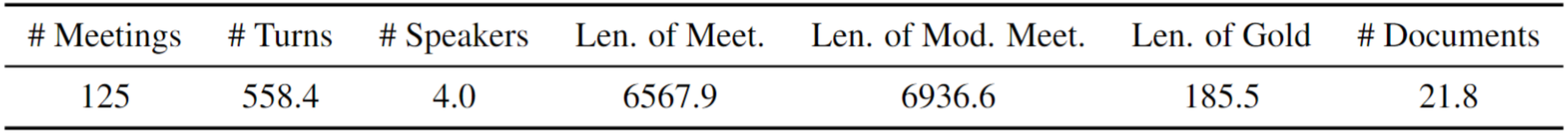
- Current summarization systems neglect that meetings are often accompanied by additional resources
  - Additional content-related sources can be considered to improve the model's context understanding
- Typically, generated summaries are general and do not fit what the target reader looks for
  - Not aligned with the reader's needs, reducing efficiency
  - Multi-source allows to include different sources/materials/etc. that are tailored to the reader’s knowledge level and help them to better understand the content

We propose a RAG-based architecture for multi-source summarization and a persona extraction module for personalization



We set up a new version of AMI to allow for multi-source summarization

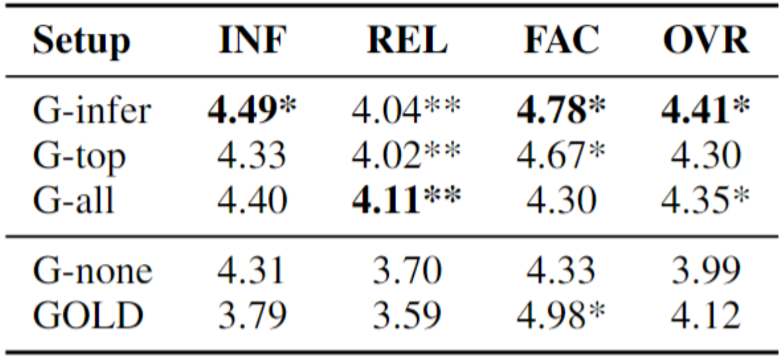
- MS-AMI: A Multi-Source Dataset for Meeting SummarizationDataset:
  - Adapted from AMI
  - 125 Business Meetings with supplementary content (whiteboard drawings, slides, notes)
- Processing Tools:
  - GPT-4 for OCR,
  - Aspose for text extraction
- Each meeting's data is compiled into a JSON file, preserving original structures
- We remove 12 samples from the initial 137 meetings due to processing errors



# Meetings	# Turns	# Speakers	Len. of Meet.	Len. of Mod. Meet.	Len. of Gold	# Documents
125	558.4	4.0	6567.9	6936.6	185.5	21.8

We employ a RAG-based pipeline to improve context understanding through multi-source summarization

- The structured inclusion of inferred details significantly enhances the summary quality
  - ○ Multi-source summarization improves contextual understanding and relevance
  - ○ Inferred details reduce hallucinations
  - ○ Selective integration outperforms simple concatenation



Setup	INF	REL	FAC	OVR
G-infer	**4.49***	4.04**	**4.78***	**4.41***
G-top	4.33	4.02**	4.67*	4.30
G-all	4.40	**4.11****	4.30	4.35*
G-none	4.31	3.70	4.33	3.99
GOLD	3.79	3.59	4.98*	4.12

▶ *LLM-based 5-point Likert scoring of the general multi-source meeting summarization pipeline.*

G-none

"In the kick-off meeting for a new project, the team, led by project manager Laura, gathered to design an original, trendy, and user-friendly remote control. The team members, including Craig (User Interface), Andrew (Marketing Expert), and David (Industrial Designer), shared their favorite animals to highlight characteristics they value, such as affection, personality, and health, which could inspire the remote's design…"

G-infer

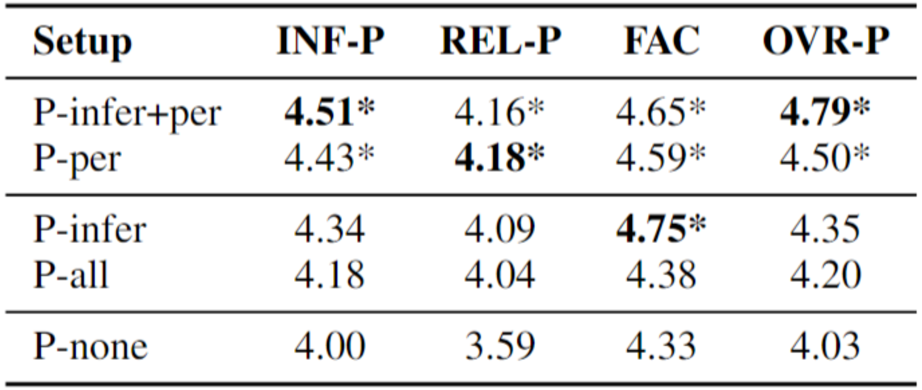
"In a kick-off meeting for a new project, a team discusses the design of an original, trendy, and user-friendly remote control. The design process will involve three stages, with individual work and collaborative meetings. The team considers various features, such as combining multiple device controls into one, adding …"

We extract details about a target reader to personalize summaries

- Detailed personas improve personalization but introduce challenges in handling and linking content to specific readers
  - Improved Personalization with Persona Extraction
  - P-infer+per provides the most tailored summaries
  - Linking content to specific personas remains complex

P-infer+per

"Laura, you as the Project Manager, led the kick-off meeting for the design of a new remote control, aiming for it to be original, trendy, and user-friendly. The team discussed the project's financial goals, targeting a selling price of 25 Euros and a production cost not exceeding 12.50 Euros. The team explored the idea of integrating multiple device controls into one unit, reflecting on the inconvenience of managing … "



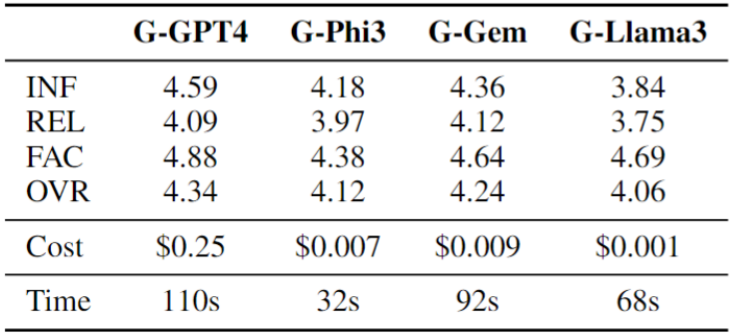
Setup	INF-P	REL-P	FAC	OVR-P
P-infer+per	**4.51***	4.16*	4.65*	**4.79***
P-per	4.43*	**4.18***	4.59*	4.50*
P-infer	4.34	4.09	**4.75***	4.35
P-all	4.18	4.04	4.38	4.20
P-none	4.00	3.59	4.33	4.03

▶ *LLM-based 5-point Likert scoring of the personalized multi-source meeting summarization pipeline.*

More efficient LLMs like Phi3 and Gemini provide viable alternatives to GPT4 for multi-source and personalized summarization

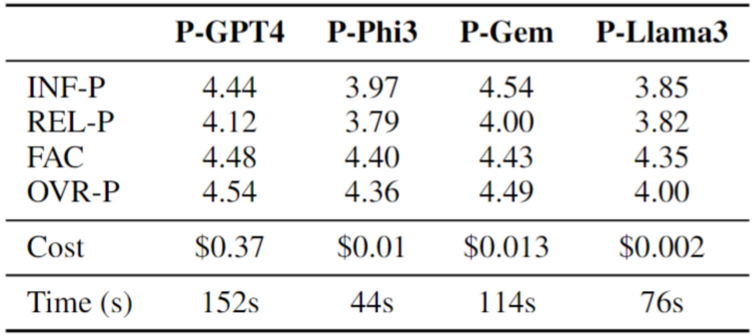
- Phi-3 provides efficient, on-device summarization with a balance of speed and performance
  - Despite being smaller than Llama3, Phi3 often outperforms it in both general and personalized summarization
  - Struggles with context gaps and requires additional quality assurance for personalization

- Gemini produces high-level, shallow summaries
  - Summaries tend to be high-level and sometimes omit important details
  - Suitable for overviews but less ideal for detailed, context-rich summaries



	G-GPT4	G-Phi3	G-Gem	G-Llama3
INF	4.59	4.18	4.36	3.84
REL	4.09	3.97	4.12	3.75
FAC	4.88	4.38	4.64	4.69
OVR	4.34	4.12	4.24	4.06
Cost	$0.25	$0.007	$0.009	$0.001
Time	110s	32s	92s	68s

Table 3: LLM-based 5-point Likert scores of the general summarization pipeline, comparing different model families. Costs are per sample.



	P-GPT4	P-Phi3	P-Gem	P-Llama3
INF-P	4.44	3.97	4.54	3.85
REL-P	4.12	3.79	4.00	3.82
FAC	4.48	4.40	4.43	4.35
OVR-P	4.54	4.36	4.49	4.00
Cost	$0.37	$0.01	$0.013	$0.002
Time (s)	152s	44s	114s	76s

Table 4: LLM-based 5-point Likert scores of the personalized summarization pipeline, comparing different model families. Costs are per sample.

Multi-source summarization and persona-based personalization enhance the quality and relevance of summaries

- Multi-source summarization improves quality
  - Incorporating multiple supplementary sources into the summarization process improves summary quality by at least 0.31 over single-source models

- Persona-based personalization increases relevance
  - Using dynamically generated participant personas for personalization enhances summary relevance by up to 0.44

- Smaller LLMs perform well
  - Phi-3 mini demonstrates that even significantly smaller models can produce high-quality summaries, performing competitively with larger models like GPT-4 turbo