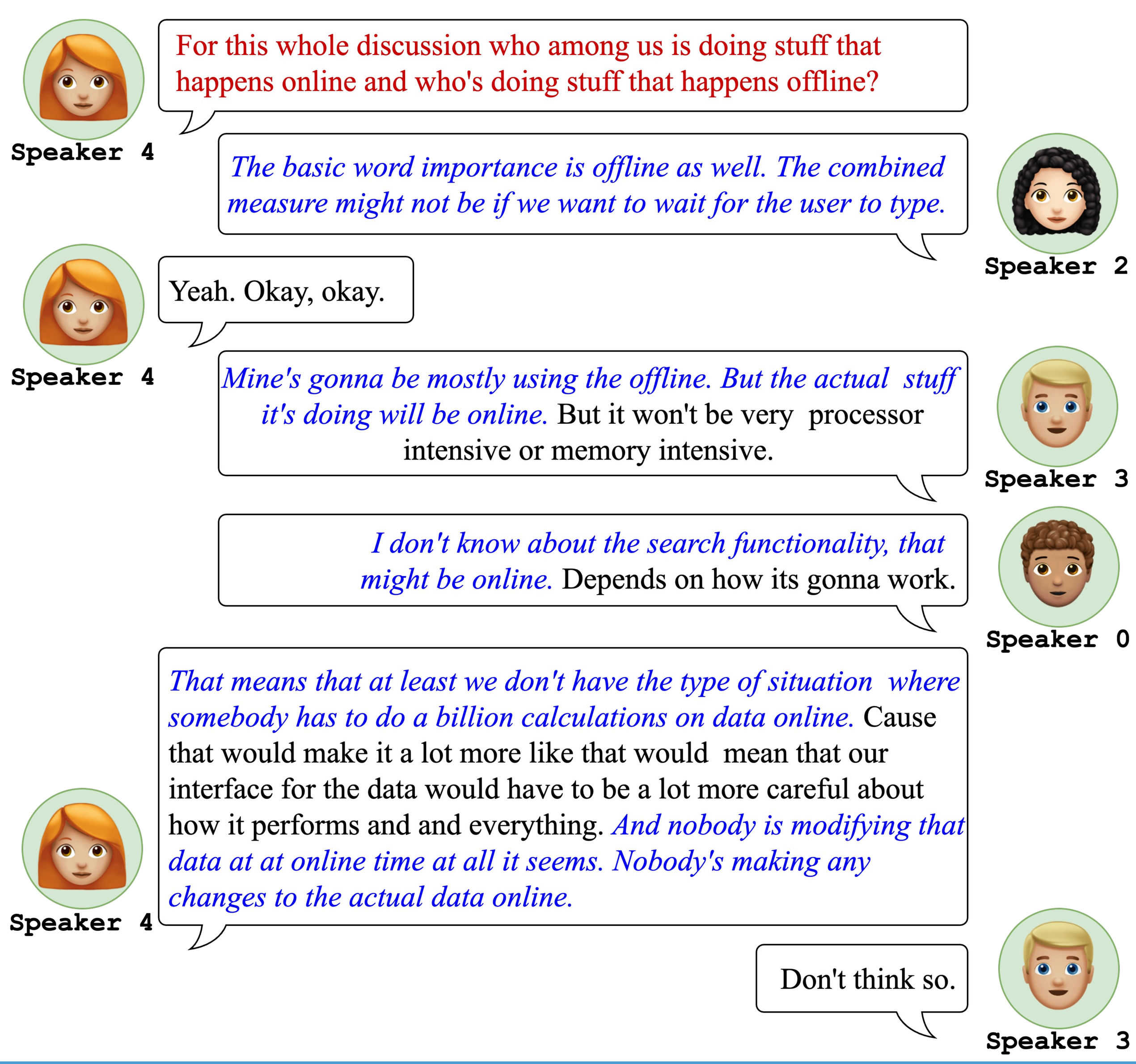
MeetingQA: Extractive Question-Answering on Meeting Transcripts

Archiki Prasad[1], Trung Bui[2], Seunghyun Yoon[2],
Hanieh Deilamsalehy[2], Franck Dernoncourt[2], Mohit Bansal[1]
[1]University of North Carolina at Chapel Hill, [2]Adobe Research

Motivation & Contributions

- Abundance of meeting transcripts – typically domain-specific & information-rich long documents for building NLP systems.
- Prior works focus on summarization and extracting action items, underutilizing significant QA components of meetings.
- We introduce ***MeetingQA***, an extractive QA dataset comprising questions asked by participants during a meeting and corresponding answer spans from relevant discussions.
- Questions asked by participants in MeetingQA are **longer**, **open-ended**, and **discussion-seeking** including interesting scenarios such as **rhetorical questions**, **multi-span answers** and/or answers contributed by **multiple speakers**.
- Despite high human performance (F1=84.6), the best QA models yield F1 of 57.3 making MeetingQA a challenging dataset with substantial room for improvement.



MeetingQA: Details and Analysis

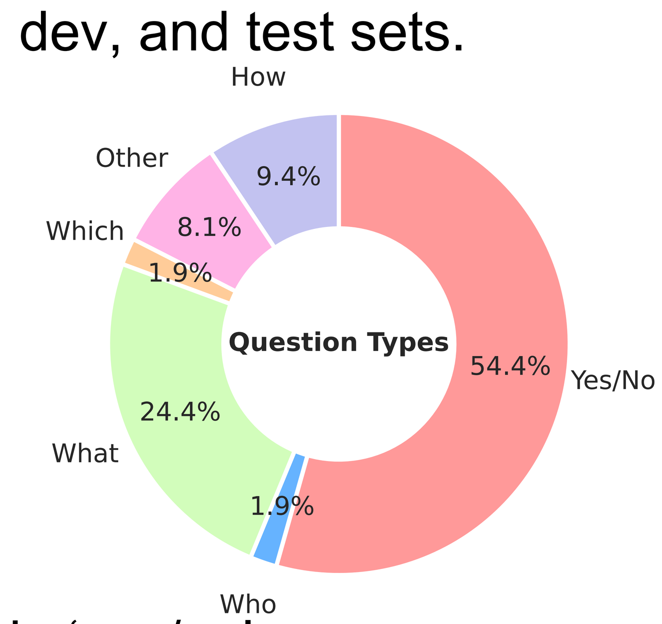
Data Collection and Annotation

- Annotate public meetings from AMI (Augmented Multi-party Interaction) corpus with ~100 hours manually transcribed meetings.
- Question Selection: based on punctuation and number of words.
- Answer Annotation: Recruit annotators to label which sentences from the transcript answer each question along with meta-data.
- High inter-annotator agreement with Krippendorff's α of 0.73, obtaining annotations for 166 meetings at $61 per meeting.

Dataset Information and Analysis

- Total of 7,735 questions split across train, dev, and test sets.

	Train	Dev	Test
Number of Meetings	64	48	54
Number of Questions	3007	2252	2476
w/ No Answer	956	621	764
w/ Multi-Span Answers	787	548	663
w/ Multi-Speaker Answers	1016	737	840
Avg. Questions per Meeting	46.98	46.92	45.85



- **Question Types**: Even questions framed in 'yes/no' manner are information-seeking and elicit detailed responses, ~50% of questions are opinion-seeking and ~20% are framed rhetorically.
- **Answer Types**: 30% of questions are unanswerable, 40% of answers are multi-span (non-consecutive sentences) and 48% involve multiple speakers. Nearly 70% of multi-speaker answers contain some level of disagreement among participants.
- **Length Distribution**: Average length of a transcript, question, and corresponding answer is 5.9K, 12, and 35 words, respectively.
- **Human Performance**: F1=84.6 on 250 questions from the test set.

Methods and Experimental Results

Finetuned Performance

	Model	Overall F1	No Ans. F1	Answerable F1		
				All	M-Span	M-Speaker
SS	RoBERTa-base	**56.5**	41.0	**63.1**	**60.8**	**64.1**
SS	Longformer-base	55.6	**46.1**	59.9	55.3	59.4
MS	RoBERTa-base	54.0	41.1	59.8	58.2	60.9
MS	Longformer-base	53.8	39.4	60.3	58.8	62.0
	Human Performance	**84.6**	**80.7**	**86.3**	**88.1**	**87.7**

- **Short-context models**: Context (that fits in model's input) retrieved from transcript based on location of the question.
- **Single-span models:** Predict single-span from first to last relevant sentence.
- **Multi-span models:** QA as token-classification.

Results:

(i) Short-context models slightly outperform long-context models by 1-2 F1 points.
(ii) Multi-span models have comparable or less performance than single-span models.
(iii) ≥ 25 F1 points gap with human performance.

Zero-shot Performance

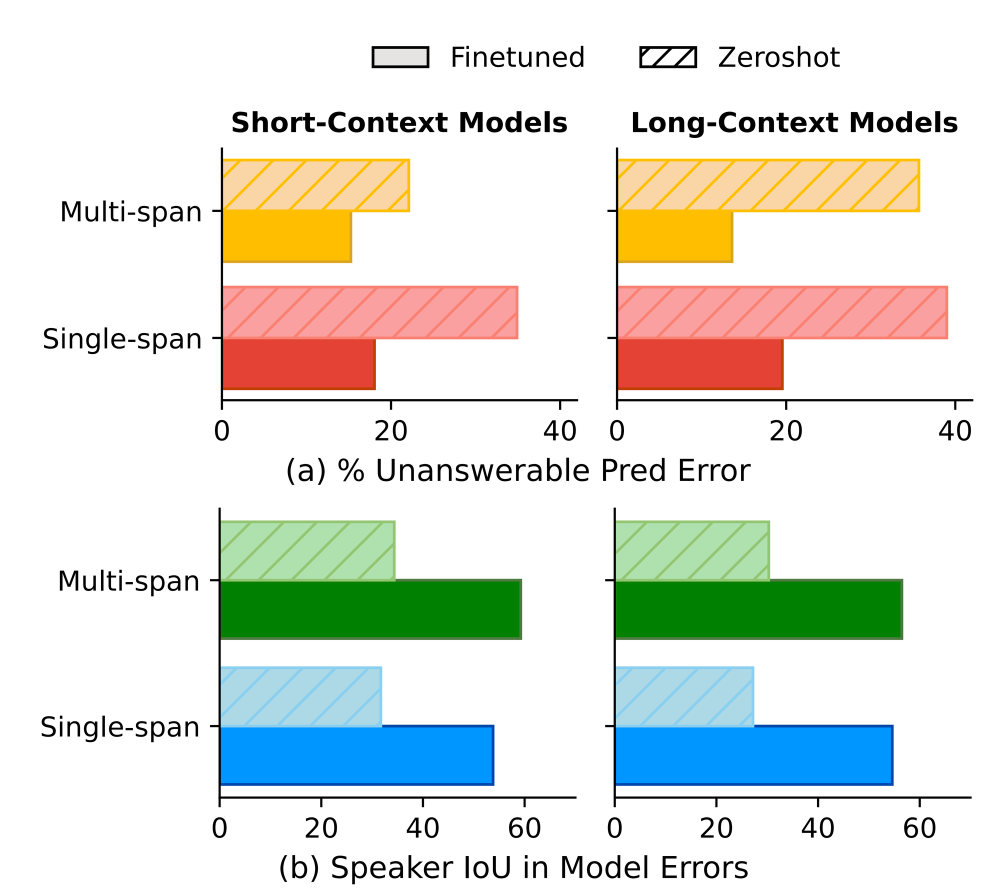
Model	Inter. Data	Overall F1
RoBERTa-base	SQuADv2	27.9
	+ silver	**34.6**
Longformer-base	SQuADv2	15.1
	+ silver	32.5
FLAN-T5 XL	–	33.8
FLAN-T5 XL (self ans)	–	34.0
Human performance	–	**84.6**

- **Silver Data Augmentation**: Augment training data with automatically annotated answer spans for interviews from MediaSum dataset.

Results:

(i) All models exhibit poor zero-shot performance (~50 F1 point gap).
(ii) Augmenting with silver data improves zero-shot performance.
(iii) Larger instruction-tuned LMs (Flan-T5) yield comparable performance.

Error Analysis



- Models struggle at identifying rhetorical questions, especially in zero-shot.
- Single-span predictions contain a greater fraction of irrelevant sentences.
- Models struggle to identify which speakers answer a question, especially in zero-shot setting.

Project Page →
Contact: archiki@cs.unc.edu