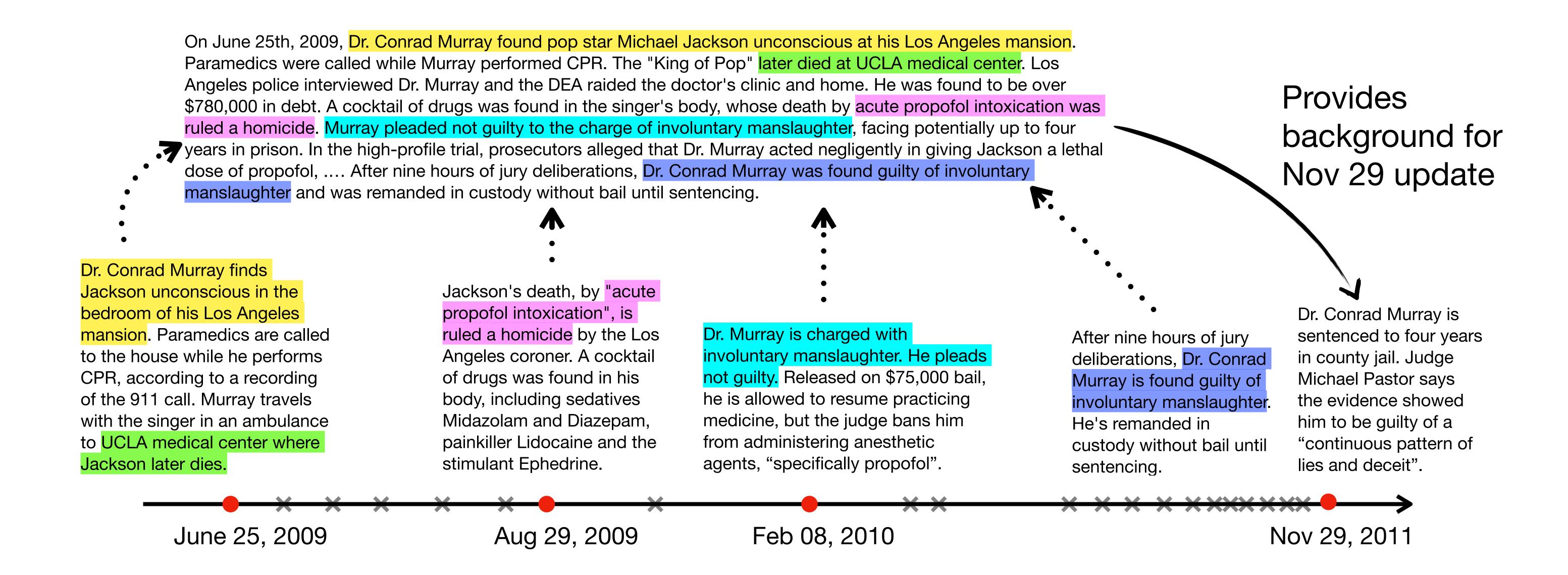
Background Summarization of Event Timelines



Adithya Pratapa, Kevin Small, Markus Dreyer

vpratapa@cs.cmu.edu, {smakevin,mddreyer}@amazon.com





New Summarization Benchmark

- Generate backgrounds for complex long-running events
- 14 major events from 2005—2014, over 1000 backgrounds
- Timelines from: Timeline17, Crisis, SocialTimeline
- 3 expert-annotated backgrounds per update

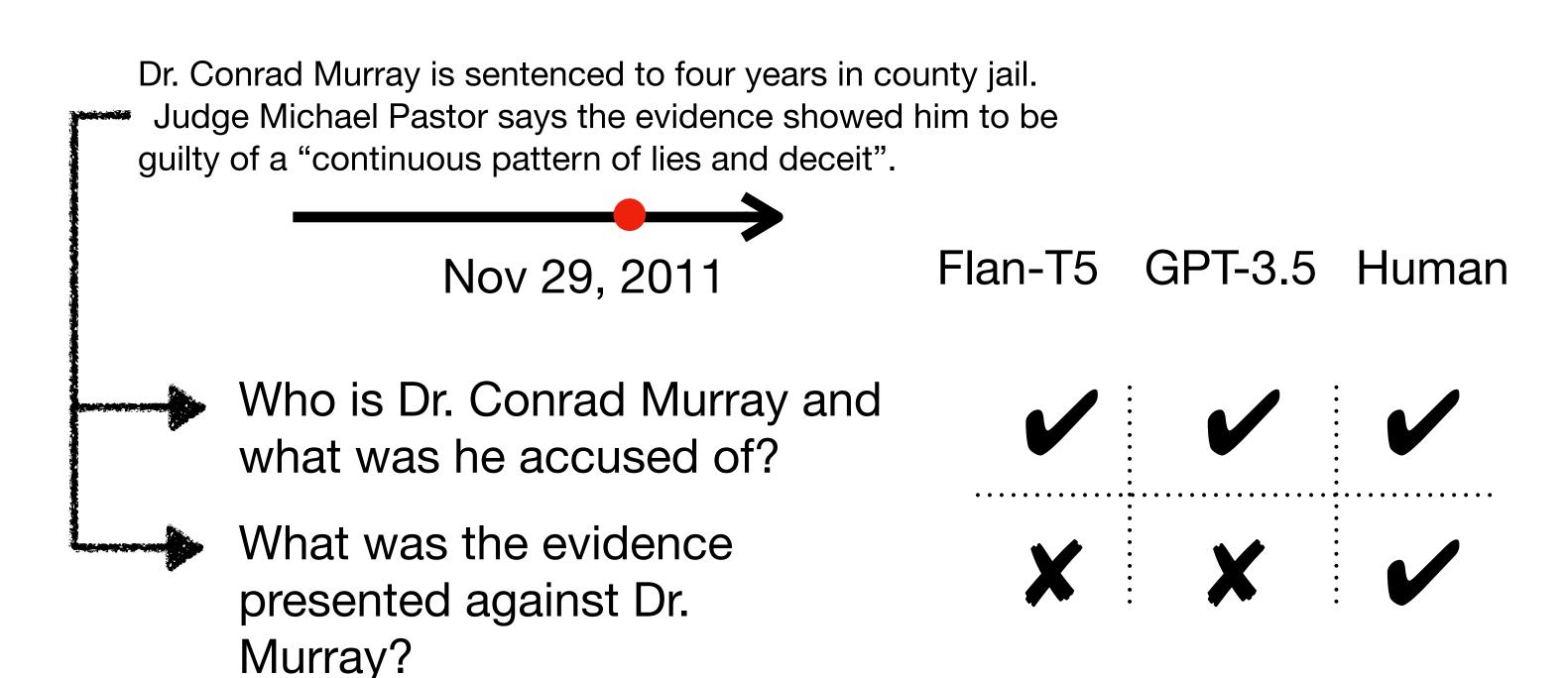
New QA-based Evaluation Metric

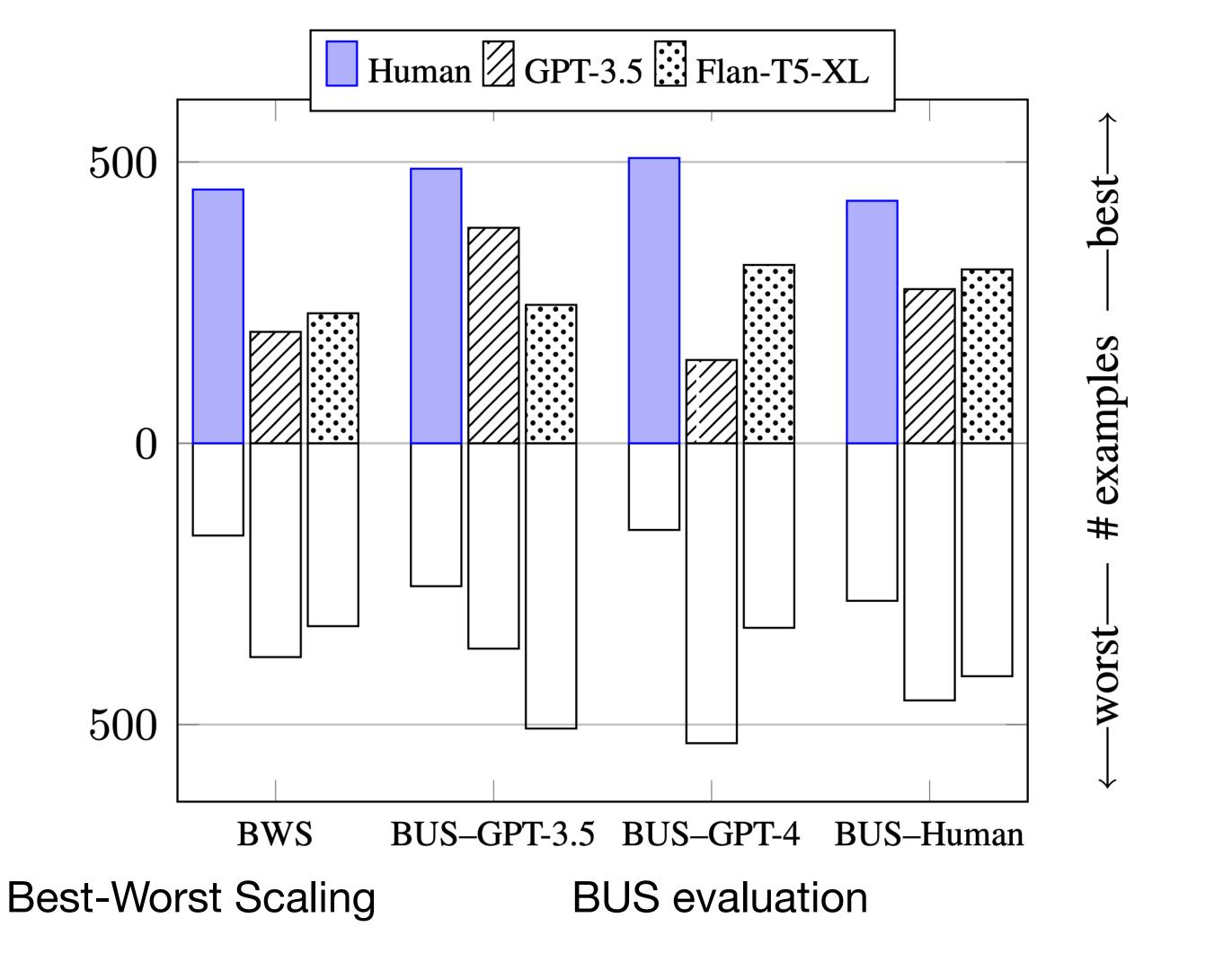
- Utility of a background to news reader
- Crowdsourced and GPT-based

Experiments

	ROUGE-1	ROUGE-2	ROUGE-L	QuestEval	BERTScore Precision
Generic summarization					
Flan-T5-XL	41.4	17.4	37.6	25.0	85.6
GPT-3.5 (zero-shot)	37.7	11.7	33.0	30.5	87.2
Query (current update) focused					
Flan-T5-XL	41.3	17.4	37.6	24.9	85.6
GPT-3.5 (zero-shot)	40.5	12.9	35.9	31.7	87.5

Background Utility Score (BUS)





Key Takeaways

- A new summarization benchmark
- Proposed utility based measure correlates well with human evaluation
- Zero-shot GPT-3.5 backgrounds are competitive to fine-tuned systems

Future Directions

- Personalized backgrounds for news readers
- Utilize news articles and short-form content as event sources



