# Scaling Multi-Document Event Summarization
## Evaluating Compression vs. Full-Text Approaches

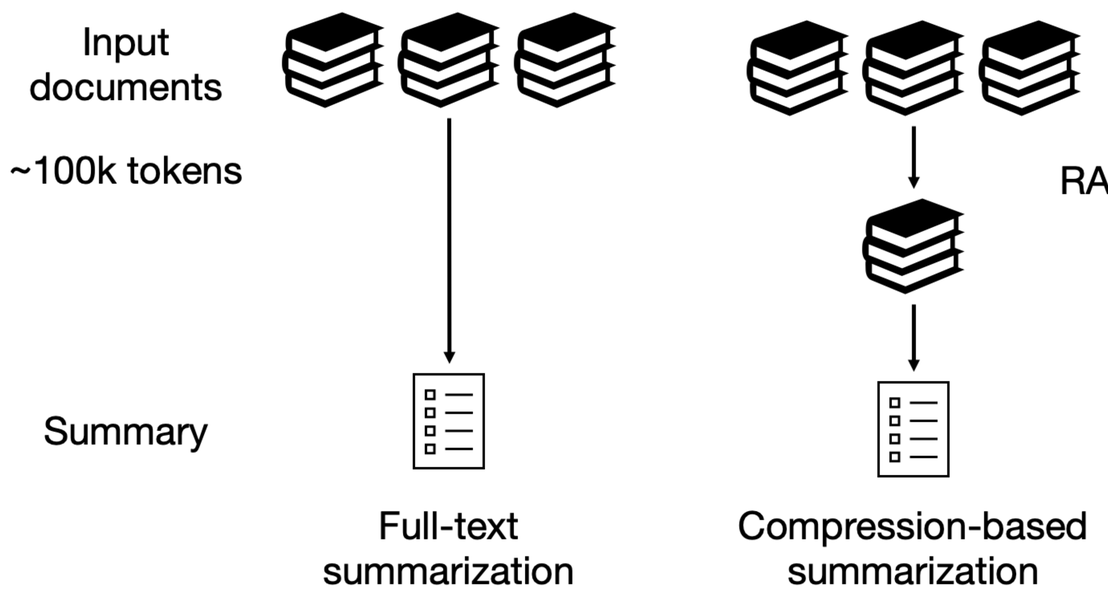
Carnegie Mellon University

Adithya Pratapa* Teruko Mitamura

*On the industry job market!

{vpratapa, teruko}@cs.cmu.edu

— Full input (w. long context window) or compressed input (w. short context window)? —




## Compression

- Efficient
- Scalable
- Retrieval, Hierarchical, Incremental

## Full-text

- Effective
- Utilizes long-context windows of recent LLMs

## Base models

- Llama-3.1 (8B, 70B)
- Command-R (32B)
- Jamba 1.5 Mini (52B MoE, 12B active)

## Datasets

- SummHay (Laban et al., 2024)
- Background (Pratapa et al., 2023)
- WCEP (Gholipour Ghalandari et al., 2020)

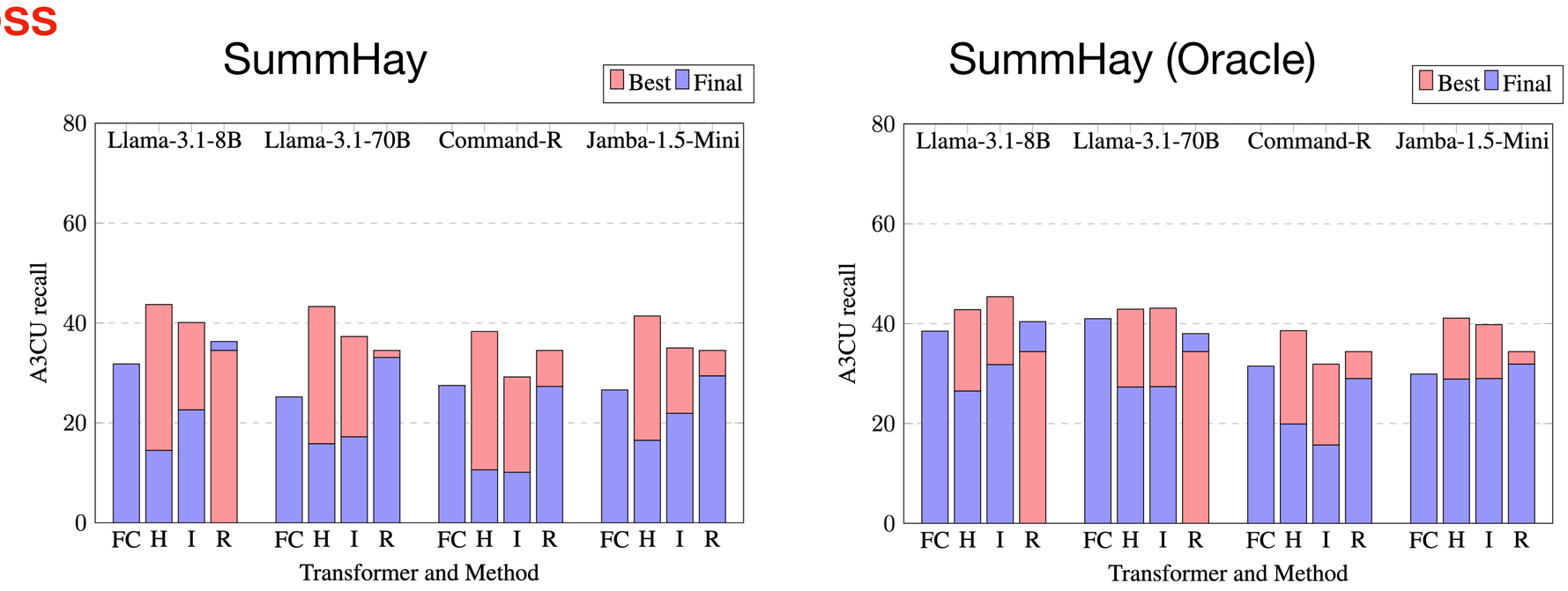
## Results

| | Llama-3.1-8B | Llama-3.1-70B | Command-R | Jamba-1.5-Mini |
|---|---|---|---|---|
| SummHay | -53% -32% +4% 33.9 | -44% -37% +17% 31.1 | -63% -63% 0% 30.4 | -54% -32% +6% 32.9 |
| SummHay (oracle) | -27% -17% +1% 37.1 | -35% -29% -6% 41.8 | -33% -47% -3% 32.6 | -16% -18% +3% 35.1 |
| Background | -16% -36% +8% 15.6 | -9% -31% -3% 16.1 | -10% -6% +15% 10.3 | -17% -15% -11% 12.3 |
| WCEP | -15% -22% -3% 30.7 | -13% -22% -2% 31.1 | -11% -13% -1% 28.9 | -2% -13% -1% 29.3 |

Relative performances of hierarchical, incremental and retrieval compared to full-text baseline (AutoACU F1)

## Tracking information loss

We measure A3CU recall scores of the final system summary and the *best* intermediate output.




## Proposals for future work

- Hybrid approaches that combine compression with long-context windows could be more efficient and effective.
- We need automatic metrics that do not rely on human-written references.