Summarization Is Not Dead Yet

Dongqi Liu, Chenxi Whitehouse, Zheng Zhao

Zhuchen Cao, Jian Li, Yabiao Wang

Saarland University, Max Planck Institute for Informatics
University of Cambridge, University of Edinburgh
Zhejiang University, Tencent Research

dongqi.me@gmail.com

EMNLP2026 BUDAPEST

TL;DR

Across domains spanning *ultra-long documents, style transfer, multi-document, multimodal, and multilingual summarization*, humans still lead on informativeness and faithfulness.
The advantage persists with fresh human-written summaries.

Why Revisit Summarization?

- Some prior studies favor LLM summaries
- Overall rankings may mix fluency with information quality
- Judge bias and evidence scope can affect the comparison

What Counts as Quality?

- Content: informativeness and faithfulness
- Form: coherence and conciseness
- Separate these dimensions before judging overall quality

Evaluation Setup

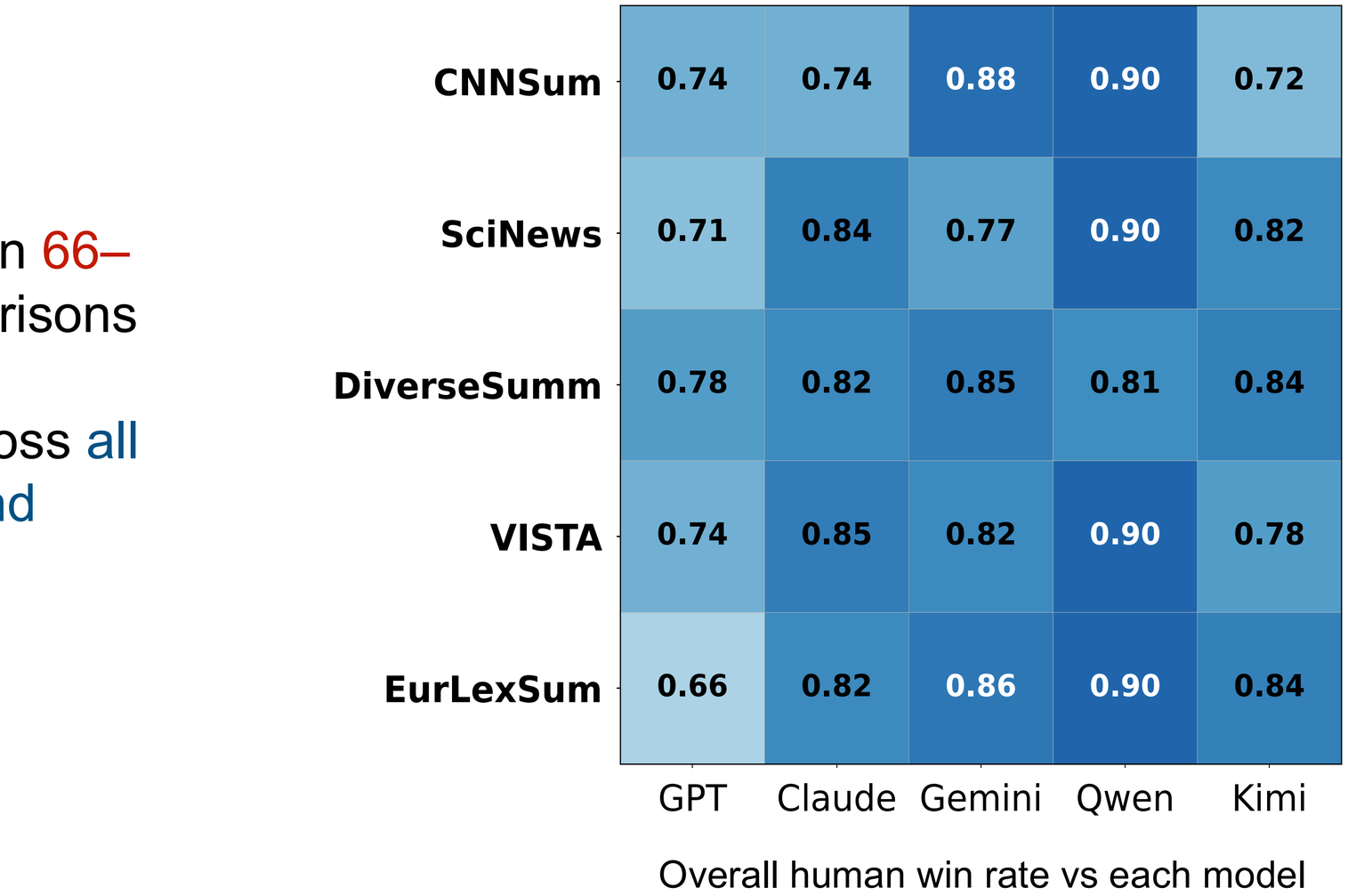
- 5 LLMs: GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro Preview, Qwen3.5-397B-A17B, Kimi-K2.5
- 5 datasets: ultra-long documents, style transfer, multi-document, multimodal and multilingual
- 4 evaluation tracks: human ratings, LLM judges, factuality verification and linguistic analysis

Human Evaluation

- Rate four quality dimensions on a 1–5 scale
- Rank six anonymous summaries in random order
- Three independent annotators per instance
- Qualified annotators, with agreement checks

Humans Still Lead

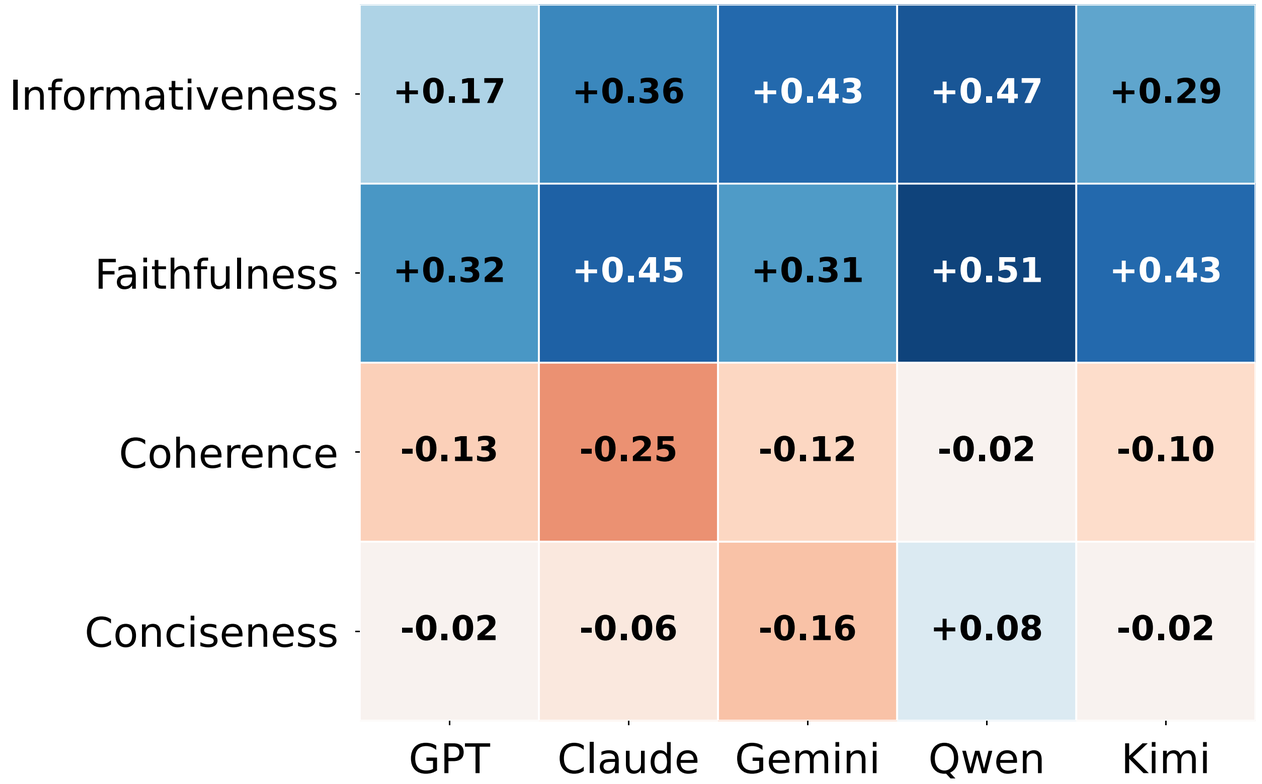
- Human summaries win 66–90% of overall comparisons
- The pattern holds across all evaluated datasets and models



Overall human win rate vs each model

The Gap Is in Content

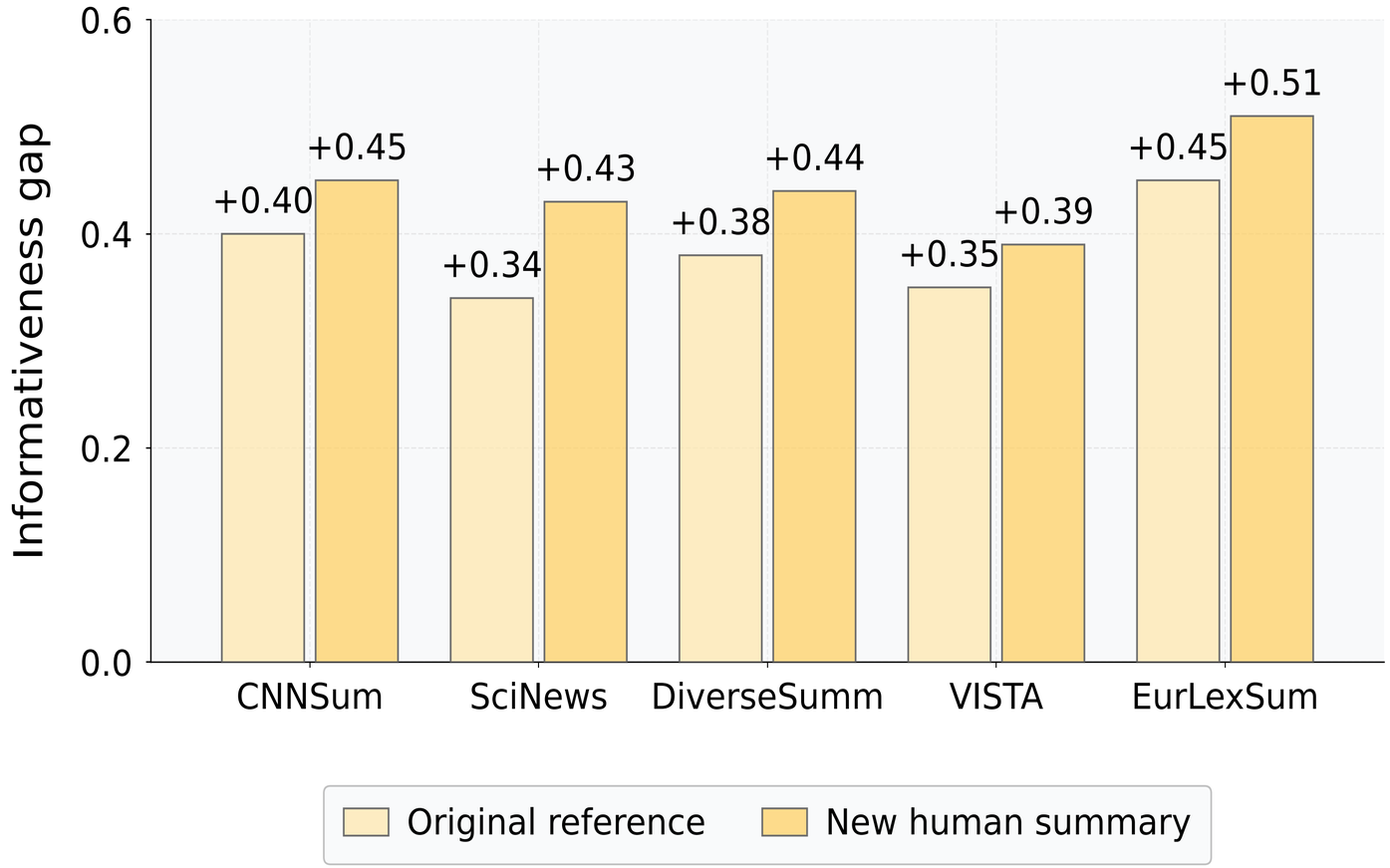
- Humans lead on informativeness and faithfulness
- GPT and Claude score higher on coherence



Mean Likert gaps across datasets
Human − model

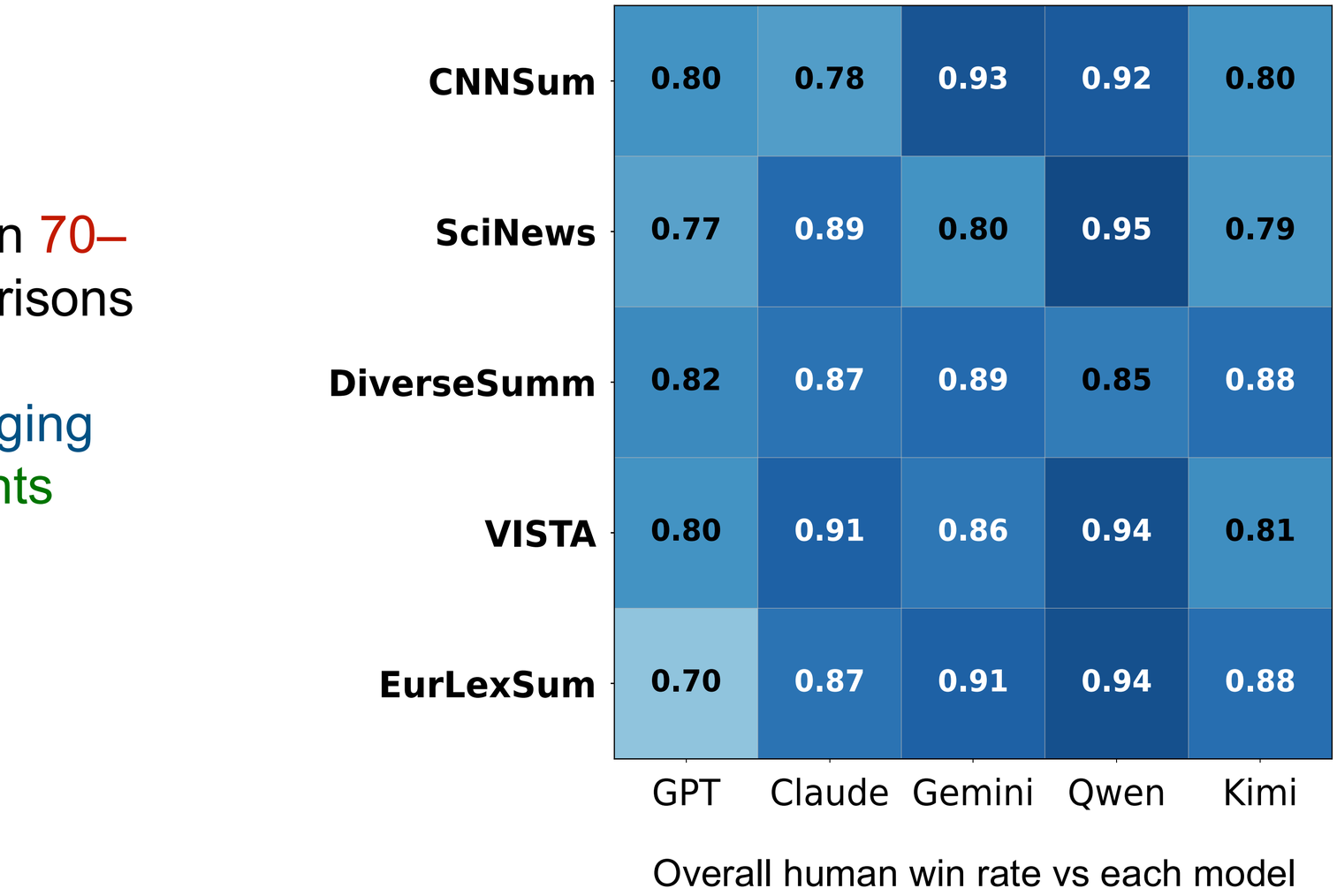
Fresh Summaries, Larger Gaps

- 120 new human summaries
- The gap widens in every dataset



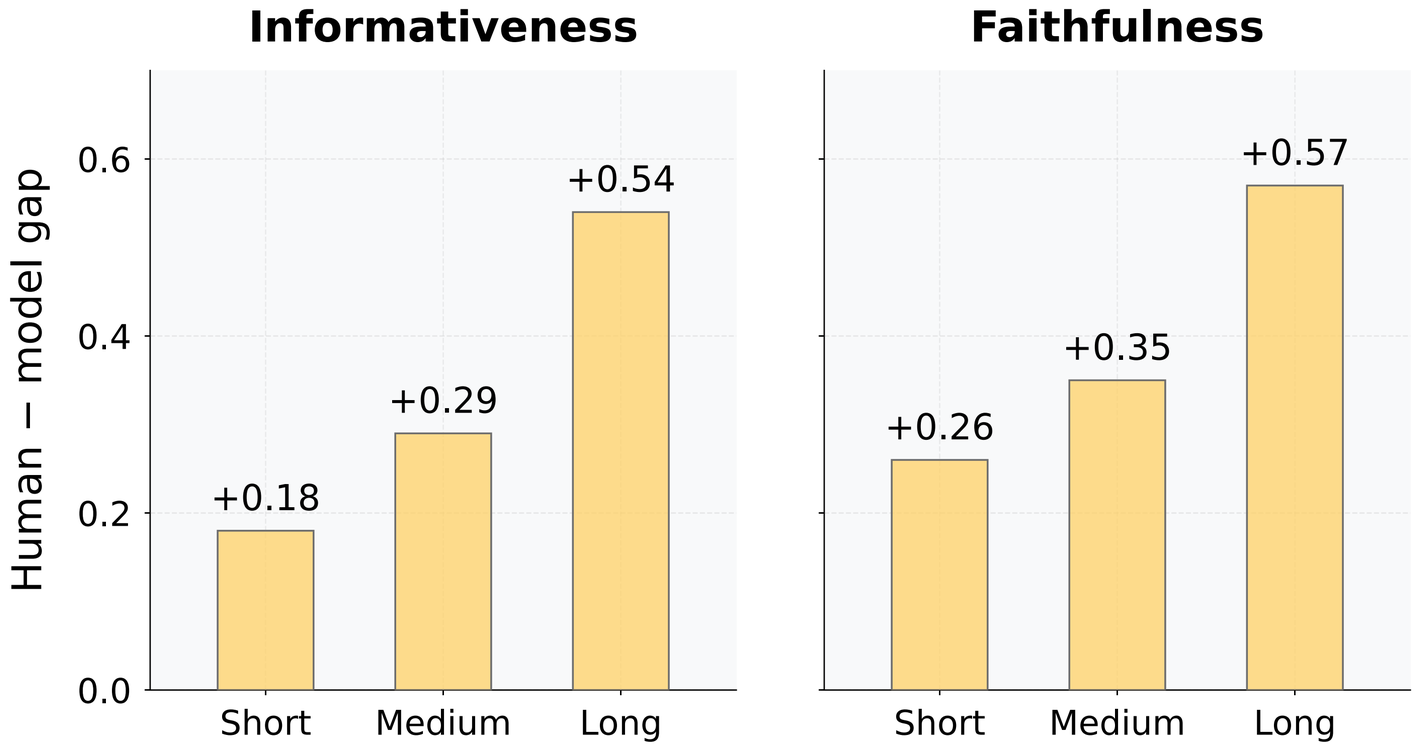
LLM Judges Also Favor Humans

- Human summaries win 70–95% of overall comparisons
- Blind, randomized judging excludes self-judgments



Longer Inputs, Wider Gaps

- On SciNews, longer inputs show larger content gaps
- Both informativeness and faithfulness gaps widen



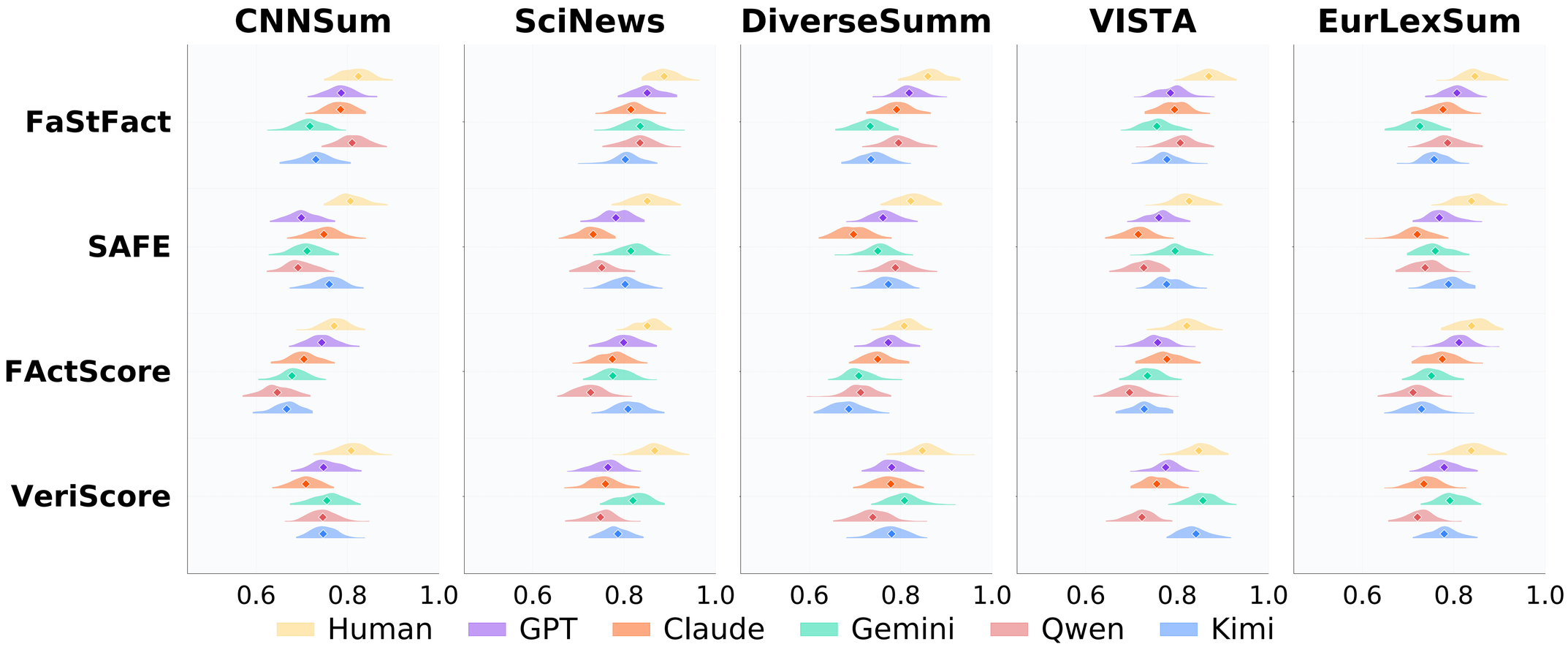
SciNews, LLM-judge gap: human − model mean

Factuality Needs Context

- SciNews lay summaries may add (valid) background knowledge
- Verify claims against the source and external evidence
- Added context should follow the task instructions

Factuality Results

- With source + web evidence, humans generally score higher

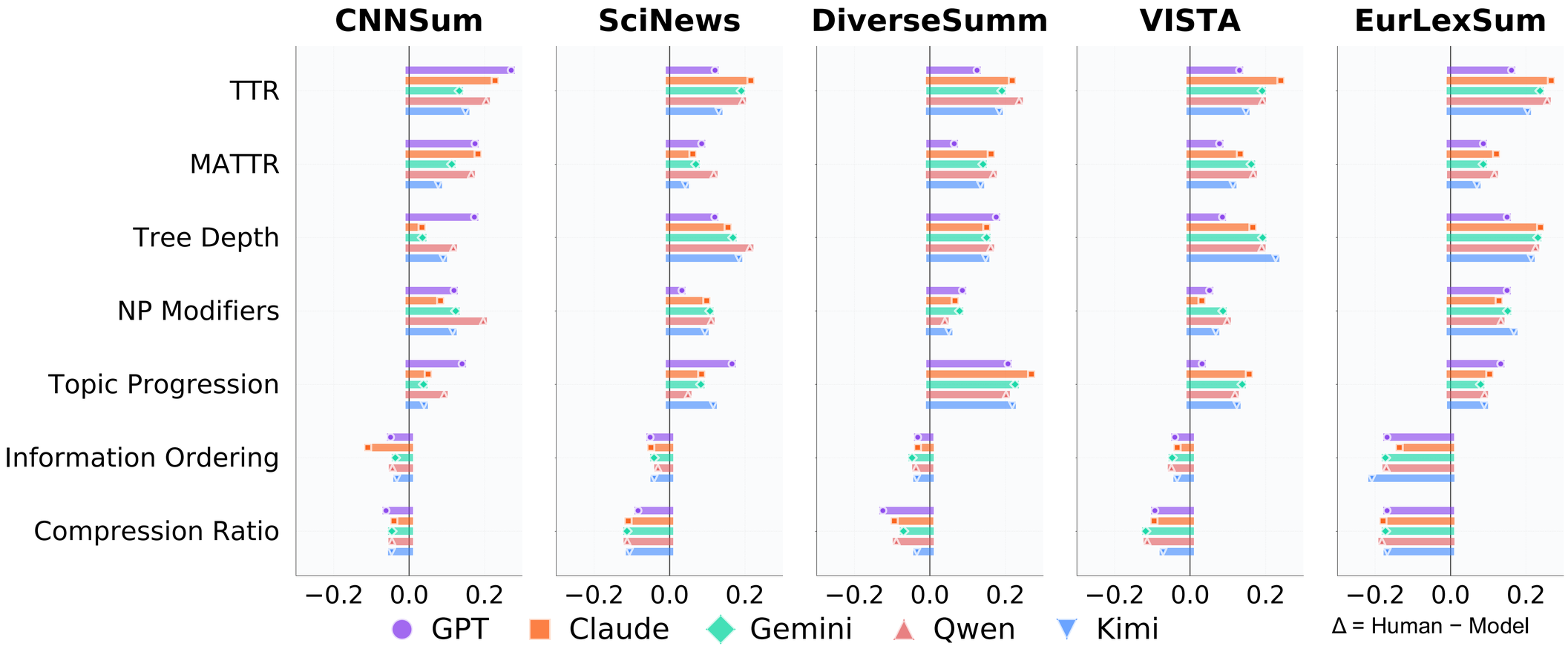


Linguistic Measures

- Words: vocabulary variety (TTR and MATTR)
- Sentences: dependency depth and noun-phrase modifiers
- Discourse: topic progression, ordering and compression

Different Writing Patterns

- Humans use more varied vocabulary and deeper dependency trees
- Humans follow source order less closely and compress more



Conclusion

- Humans still lead on informativeness and faithfulness
- Fluency, factual support and task compliance need separate checks
- Summarization remains an open research problem

Thanks for listening