UNIVERSITY of WASHINGTON

Retrieval Augmentation of Large Language Models for Lay Language Generation

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, Trevor Cohen

RALL Paper

CV

Problem:

- Automated lay summary generation (LLG) can improve the accessibility of health information, but is challenging because of the need to provide background information absent in source documents.

What is already known:

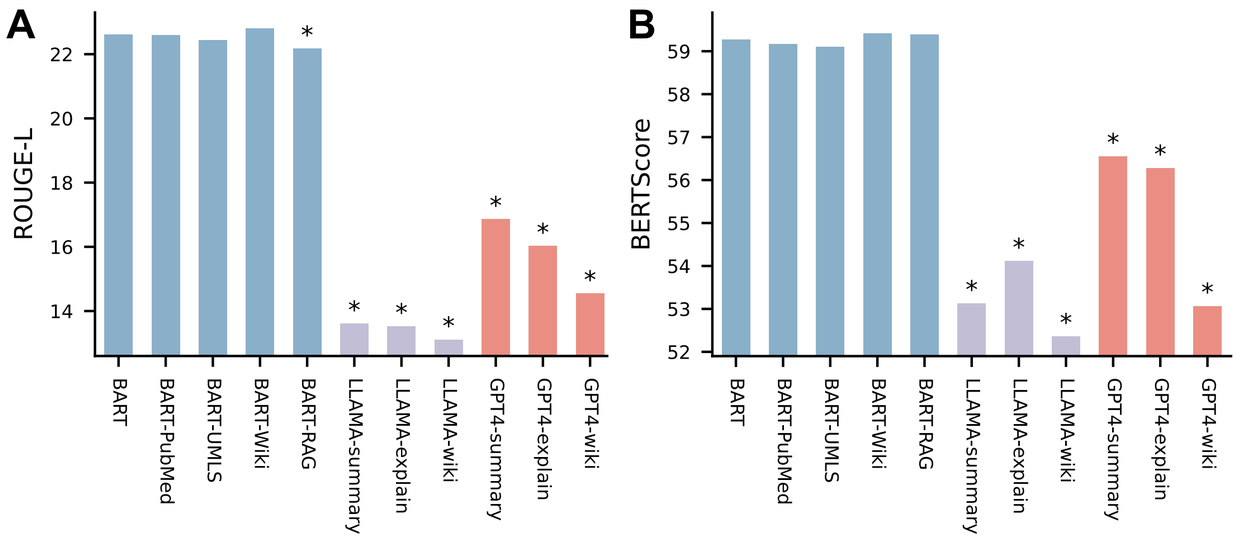
- Current models face constraints due to corpus size, topic diversity, and untested utility of external information retrieval.

What this paper adds:

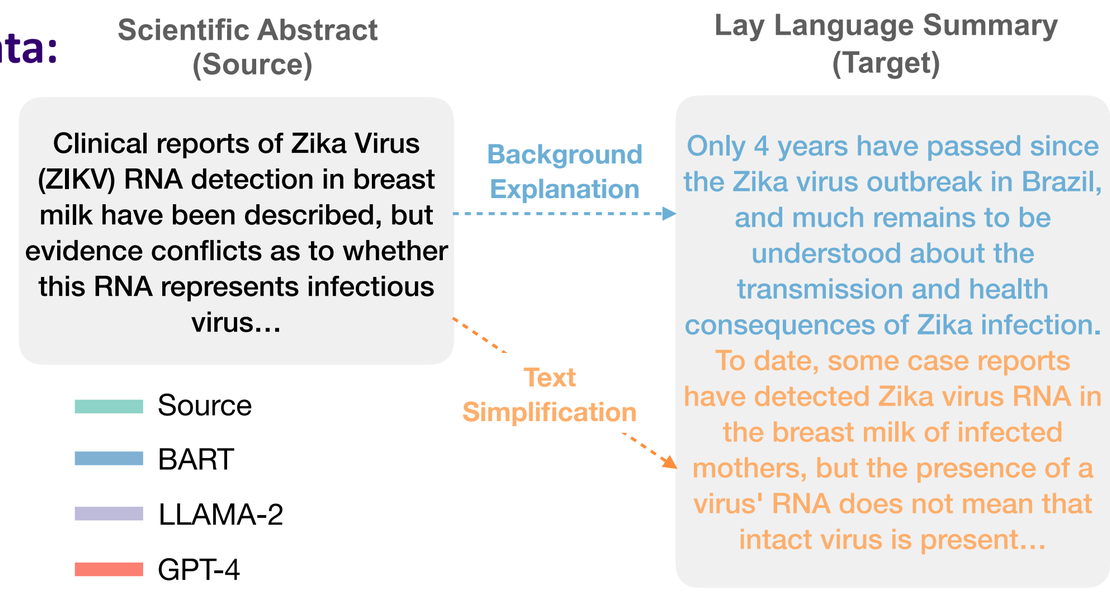
- We approach LLG by simplifying content and also generating background explanations, achieved through innovative Retrieval-Augmented Lay Language (RALL) methods.
- We introduce CELLS, the largest (63k pairs) and most diverse (12 journals) parallel corpus for lay language generation, with a specialized subset to advance background explanation capabilities.

* All results here are for background explanation subset. Results for abstract-level LLG and sentence-level simplification can be found in the QR code.

RQ1. Can RALL improve text generation quality?

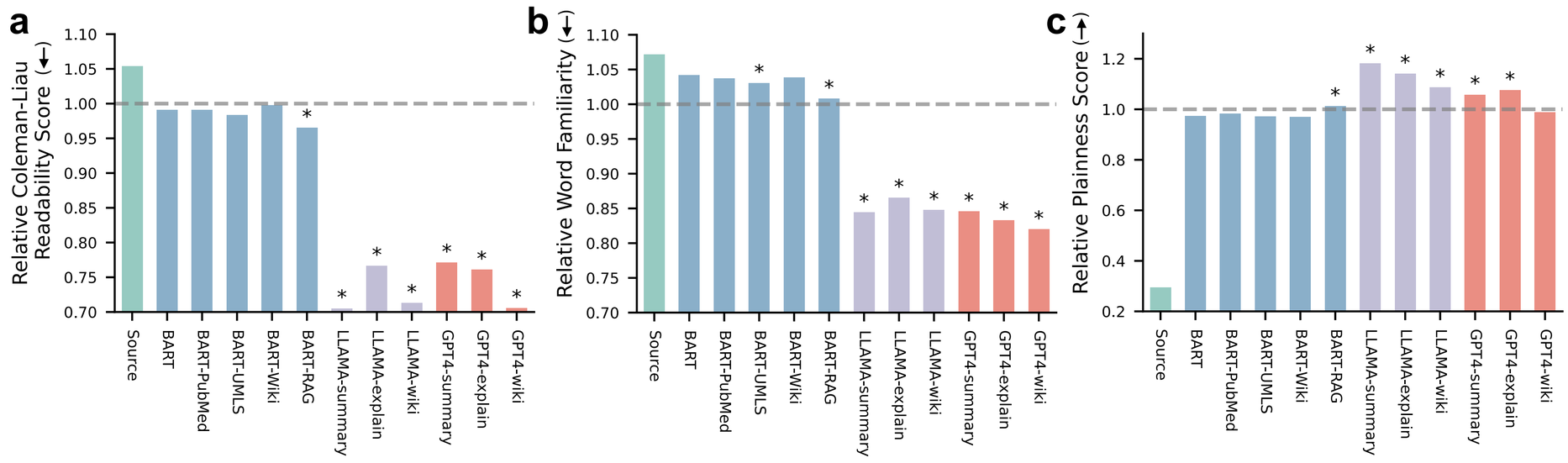


Data:



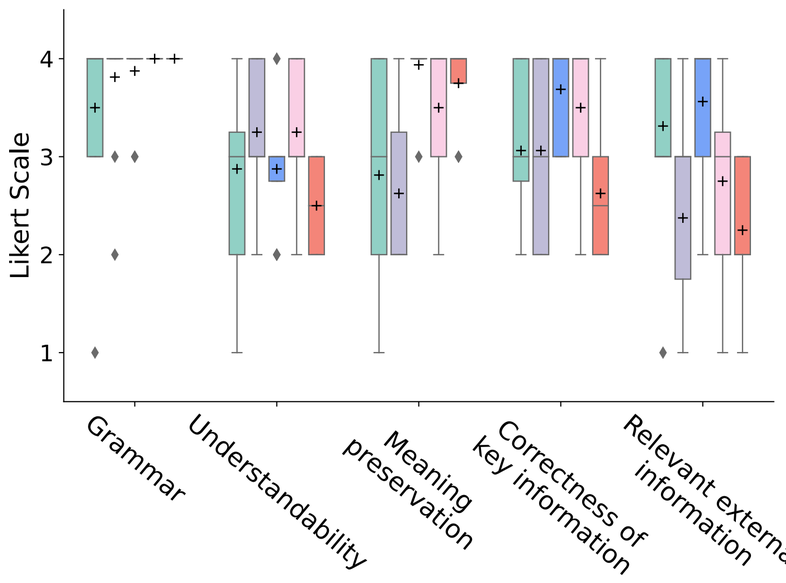
Example abstract/summary pair from CELLS. The lay language summary is written by the abstract's authors. There are two challenges in lay language summary generation: background explanation and text simplification.

RQ2. Is RALL effective in enhancing the simplicity of text?



- A lower readability score and word familiarity indicate that the text is easier to read (values below the dashed line are lower than those from professionally-authored plain language).
- A higher Plainness Score indicates that the text is more representative of an LLS.
- P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART).
- * indicates statistical significance with Bonferroni-Holm correction for multiple hypothesis testing

RQ3. Can RALL preserve grammatical correctness, understandability, key information, factuality, and relevance of information? (Assessed through human evaluation)



Each text was assigned to four raters. For the Likert scale, 1 is very poor, and 4 is very good. "+" indicates the mean.

Takeaways:

> RALL variants help in LLG, and Wikipedia is a good source for external knowledge

> Background explanation generation is challenging and requires considering both knowledge sources relatedness and simplicity of the text

> LLMs show promise in text simplification, but need improvement in extracting related information

Acknowledgment: The study was supported by NLM [R21LM013934]