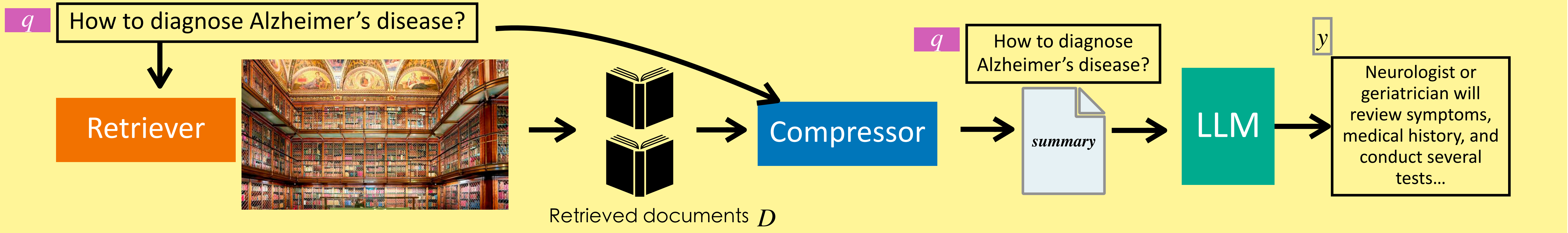
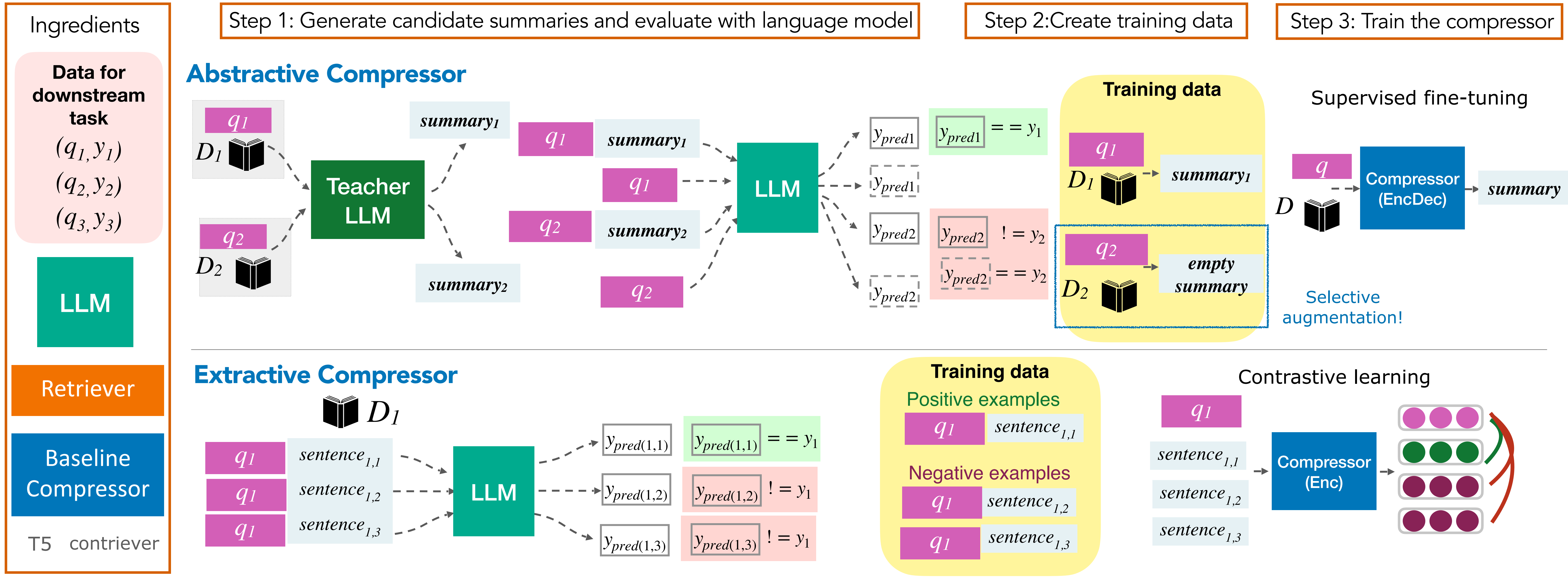
Fangyuan Xu[1], Weijia Shi[2], Eunsol Choi[1]
[1]The University of Texas at Austin,
[2] University of Washington

RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation

Key Idea: Learn a compressor that maps noisy retrieval outputs into a summary and then prepend the summary as an input to LM

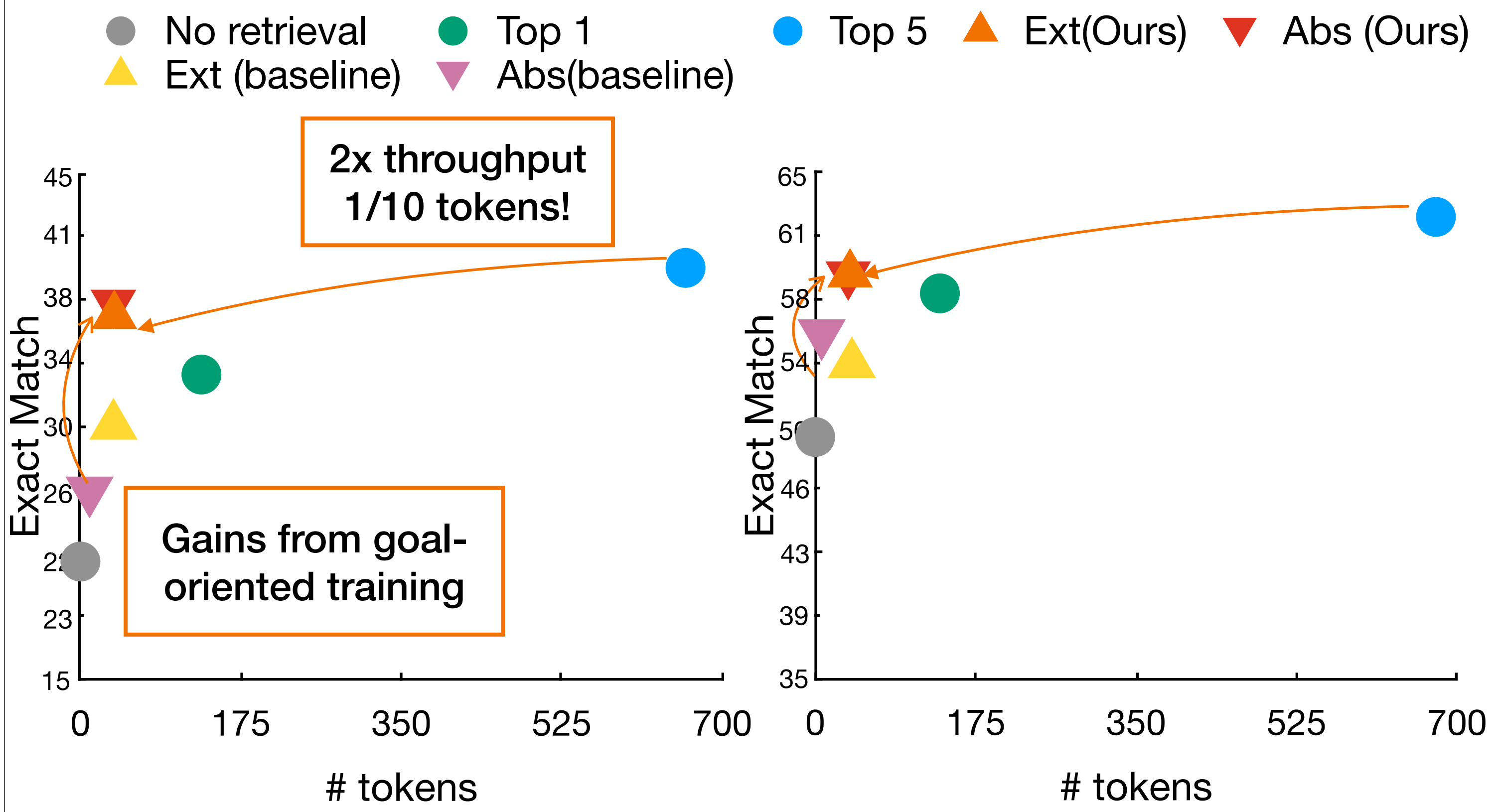


Learning a compressor



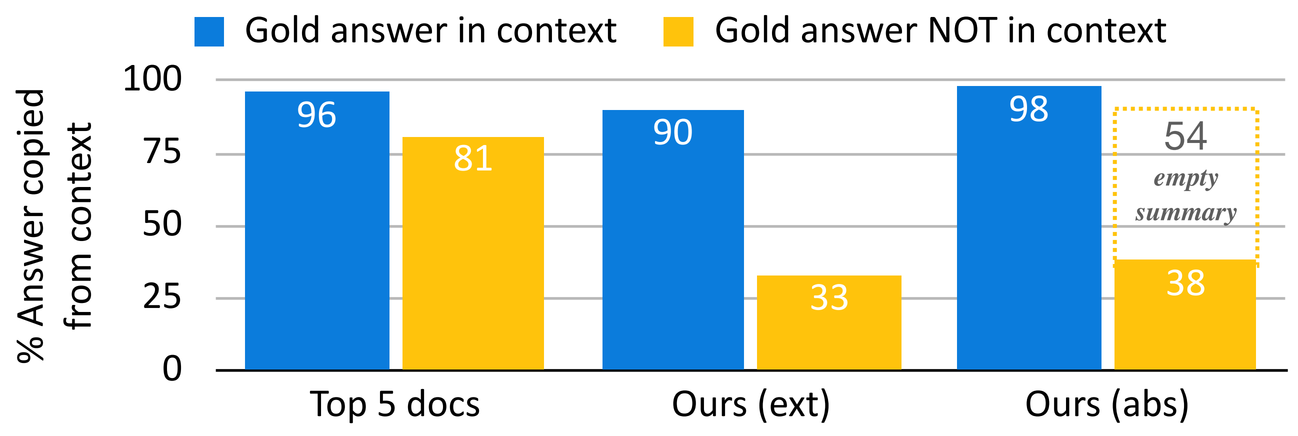
Experiments & Results

✦Our method enables efficient inference with minimal performance drop.



FLAN-UL2 (20B) results on Natural Questions (left) and TriviaQA (right)

✦Our method decreases incorrectly copying from prepended documents.



FLAN-UL2 (20B) results on Natural Questions

More analysis in the paper!

- Does compressor trained with one LM transferred to other LMs?
- Does compressor trained with one retriever transferred to others?
- How often does selective augmentation happen?
- How good is the teacher model for abstractive compressor?