Concise Answers to Complex Questions: Summarization of Long-form Answers

Abhilash Potluri* acpotluri@utexas.edu
Fangyuan Xu* fangyuan@utexas.edu
Eunsol Choi eunsol@utexas.edu

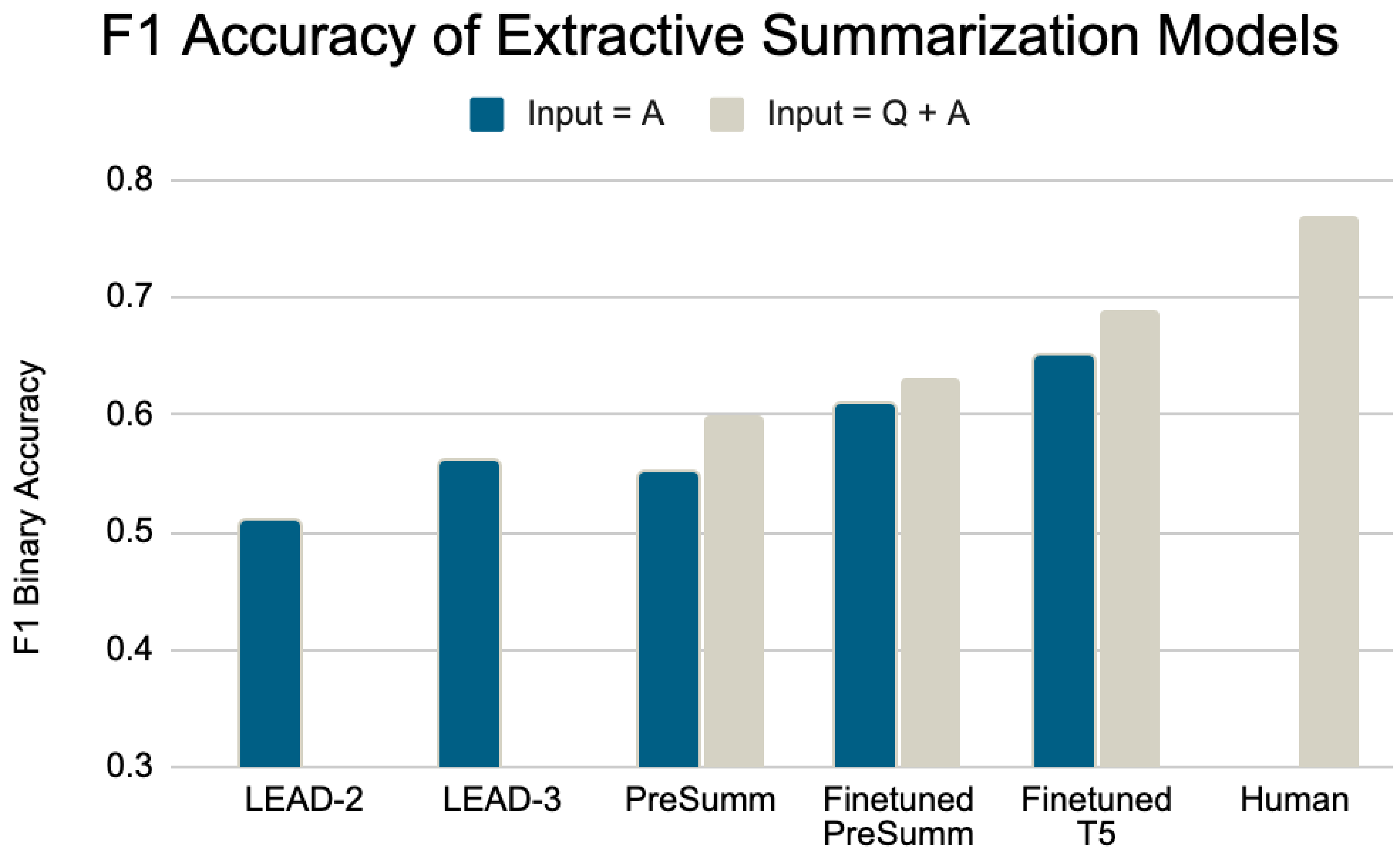
Motivation for Concise Answers

- Long-form question answering (LFQA) opens door to generate comprehensive answers to complex questions. But users prefer concise answers. Can we summarize long-form answers while still addressing questions?
- Recent study on discourse structure of long-form answers [1] found that nearly 40% of answer sentences contain non-essential information (e.g., providing examples, providing background information).

Question: What is the weather usually like in Australia?

Long Answer: Australia's climate is governed largely by its size and by the hot, sinking air of the subtropical high pressure belt. This moves north and south with Antarctica. But it is variable, with frequent droughts lasting several seasons - thought to be caused in part by the El Nino - Southern Oscillation. ***The climate varies widely due to its large geographical size, but by far the largest part of Australia is desert or semi-arid. Only the south - east and south - west corners have a temperate climate and moderately fertile soil. The north part of the country has a topical climate, varied between tropical rainforests, grasslands, and part desert., grasslands, and part desert.***

Example question, long-form answer, and ***extractive summary of the answer***.



New Extractive Summarization Dataset for LFQA

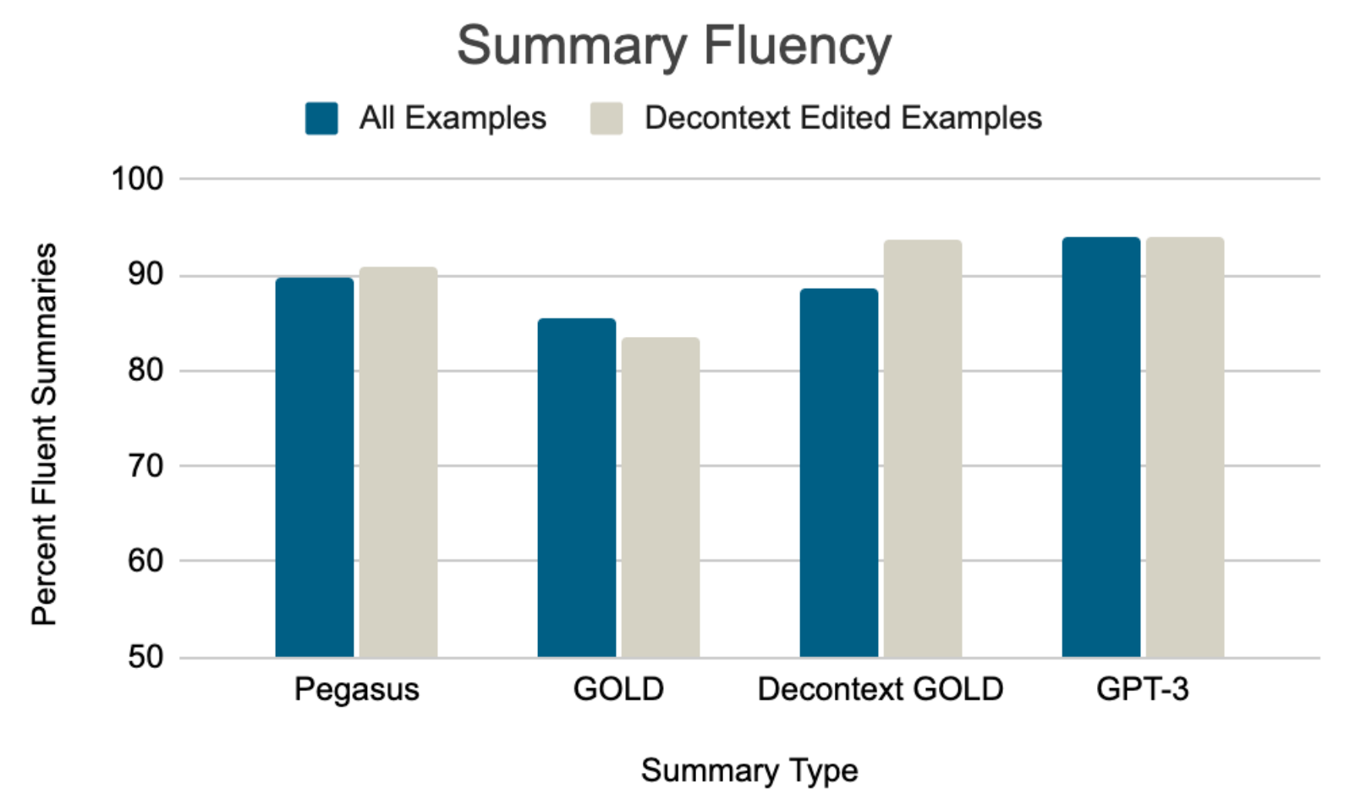
- Annotators select a subset of answer sentences as "summary" for long-form answers.
- Question/answer pairs sourced from three datasets: ELI5 [3], NQ [4], WebGPT [5]
  - Total of 1134 examples (three-way annotated), with average long answer length of 118 tokens and average summary length of 35 tokens.
- We evaluate extractive and abstractive summarization models (Presumm, T5, Pegasus, GPT-3) on this new benchmark. Having question as an input always helps!

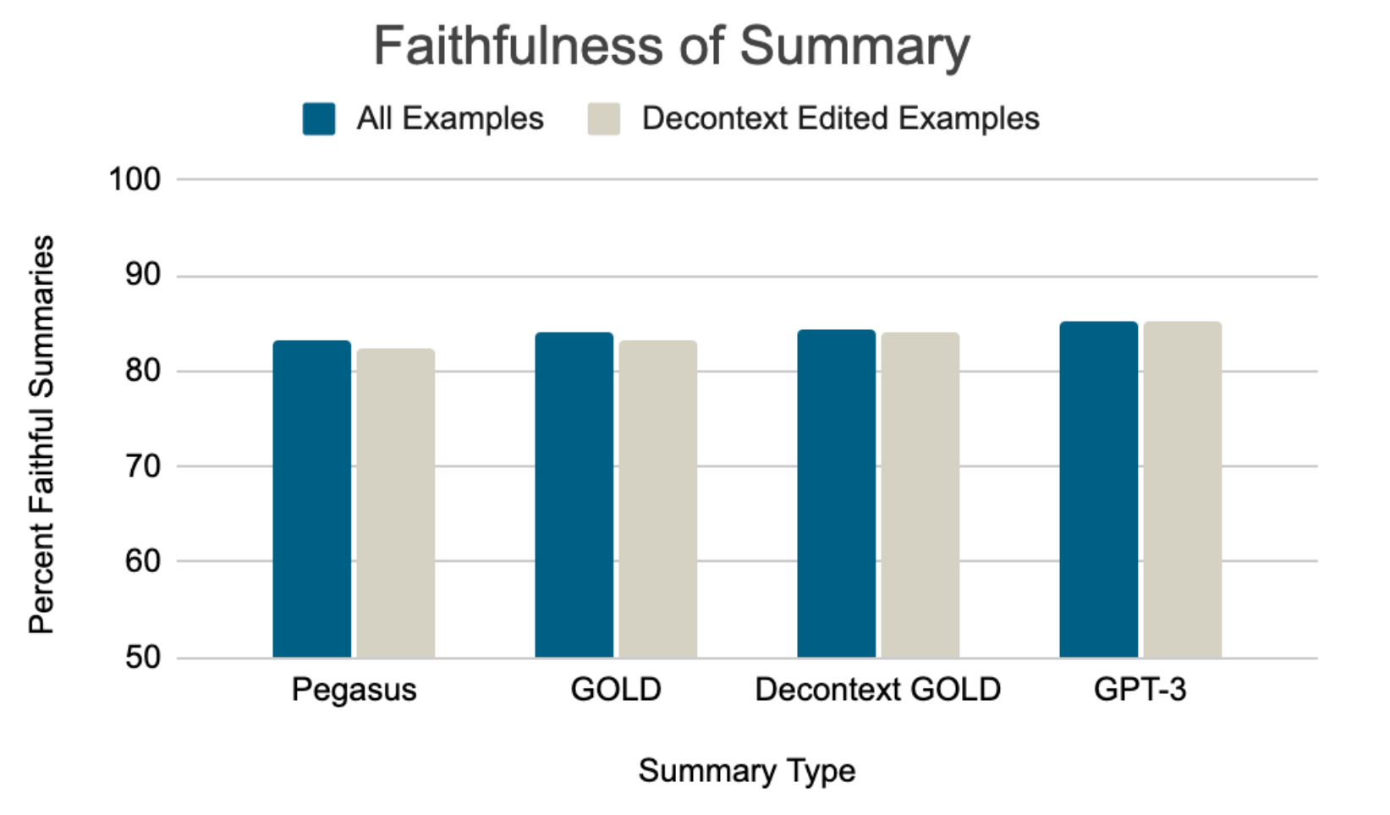
Question	Long Answer (Abridged)	Decontextualized Extractive Summary
What is the social construction of reality in sociology?	Berger and Luckmann introduced the term "social construction" into the social sciences and were strongly influenced by the work of Alfred Schutz. ***Their central concept is that people and groups interacting in a social system create, over time, concepts or mental representations of each other's actions ...***	**-Their** **+Berger and Luckmann's** central concept is that people and groups interacting in a social system create, over time, concepts or mental representations of each other's actions ...
How did Switzerland stay out of WWII?	They were literally the bankers of the war. ***The Nazis and the allies both kept their assets there.*** This is how they stayed neutral, because if either side invaded, that side's assets would either be seized by the other side, or seized by the Swiss.	The Nazis and the allies both kept their assets **-there** **+in Switzerland** .

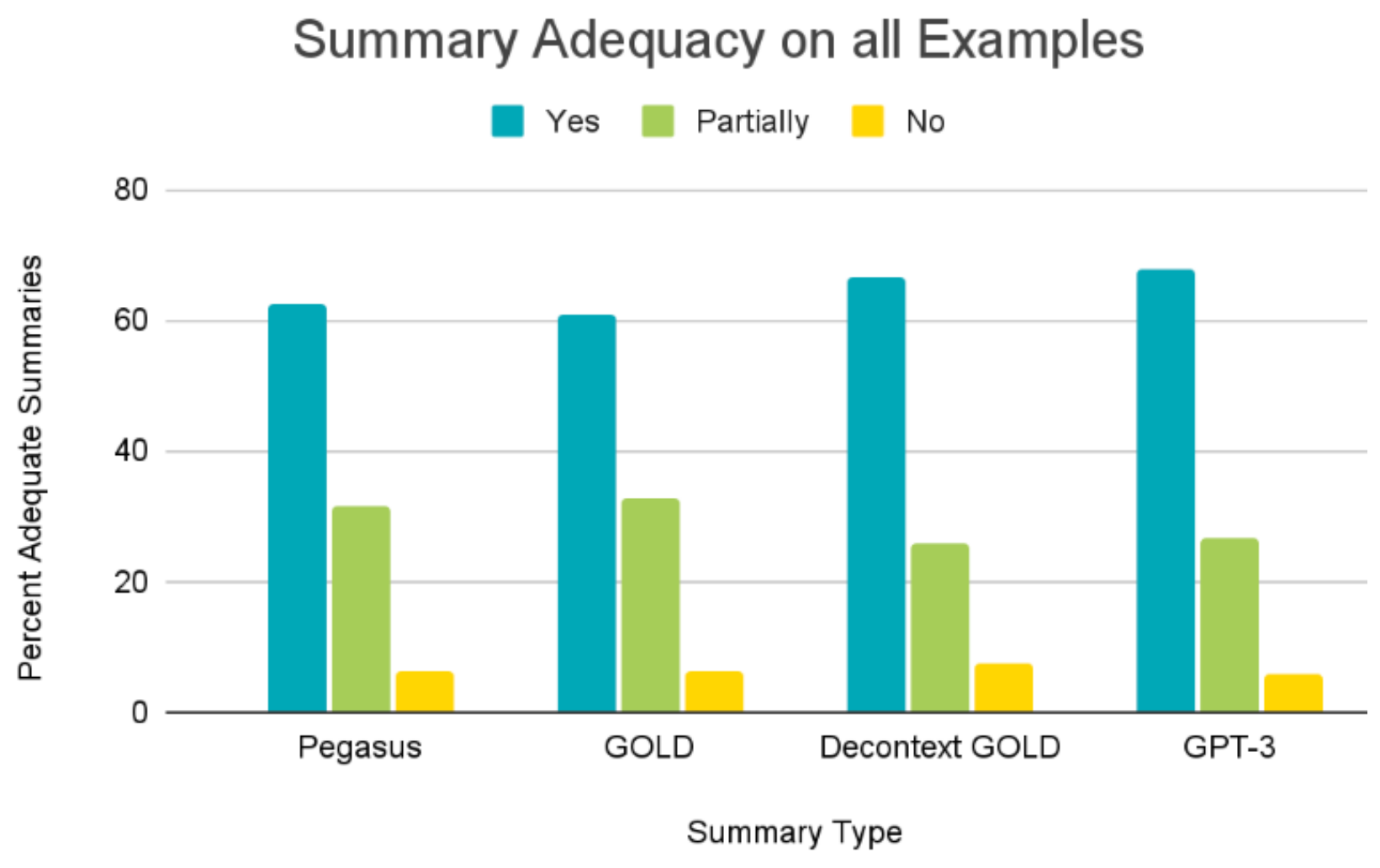
Example question, long-form answer, ***extractive summary***, and decontextualized summary

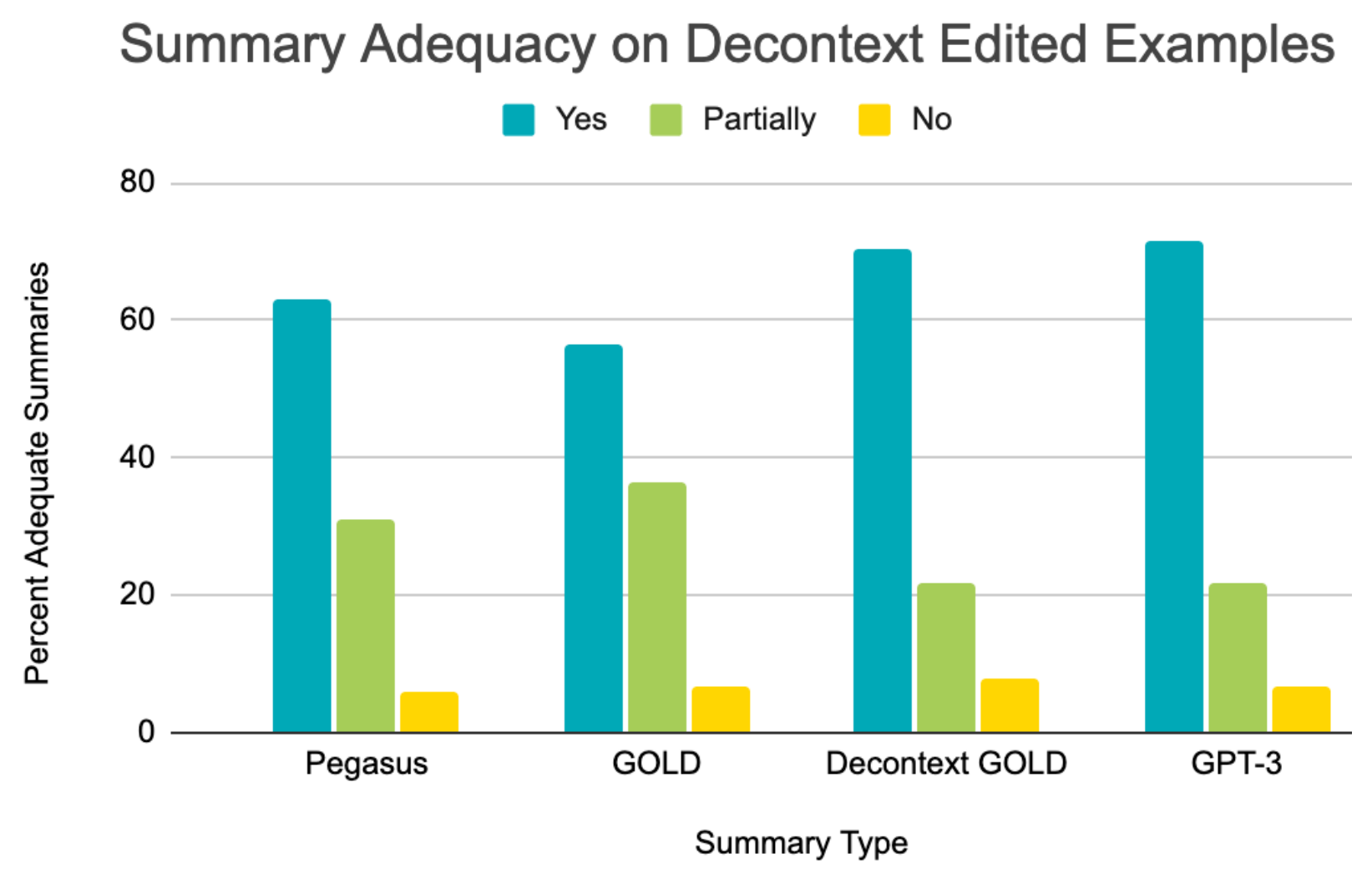
New summarization method: Extract-and-decontextualize

- Extractive summary often suffers from fluency issues
- Decontextualization task rewrites a sentence in a document such that the sentence is intelligible when presented alone.
- We use an off the shelf decontextualization model [2] to rewrite the first sentence in extractive summaries (GOLD) to improve extractive summary (Decontext GOLD)









Human Evaluation: Setting

- Evaluated 175 question/long-form answers with Amazon Mechanical Turk.
- Compared four summarization systems: Pegasus, human-annotated extractive summary (GOLD), decontextualized GOLD summary (Decontext GOLD), GPT-3 (text-davinci-002).
- Out of 175 examples, decontextualization edited 63 sentences .
- Crowdworker rated:
  a. **Summary Fluency**: is the summary grammatically correct with no incomplete references? Yes/No
  b. **Summary Faithfulness:** does the summary maintain the same main idea as the original long answer? Yes/No
  c. **Summary Adequacy:** does the summary provide a sufficient answer to the question? Yes / Partially / No

Human Evaluation: Results

- Decontextualization improves fluency of the gold extractive summaries (by almost 10% on sentences where decontextualization changed the sentence).
- GPT3 shows the strongest performance, followed by Decontext GOLD.
- Over 90% of questions had a functional (fluent + adequate + faithful) summary from at least one summarization system.
- Some long form answers are challenging to summarize (one example on the right, more in paper)

Question: Why do most restaurants sell Pepsi instead of Coke, and yet Coke is seen to be a bigger competitor?

Answer: Coke sells way more soda by volume than Pepsi. As a response, Pepsi offers its products to restaurants at a reduced cost, which is why many restaurants carry it. But only up to midscale places -- no nice restaurant serves Pepsi, because Coke has more cachét, and also you need it for mixed drinks. Note also that McDonald's, the single biggest restaurant chain in the world, serves Coke.

Example question and long-form answer that is hard to summarize.

Check out our code and data at:
github.com/acpotluri/lfqa_summary

References

1. Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics.*
2. Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *CoRR*, abs/2102.05169
3. Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long-form question answering. *CoRR*, abs/1907.09190.
4. Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics,* 7:452–466
5. Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint* arXiv:2112.09332.