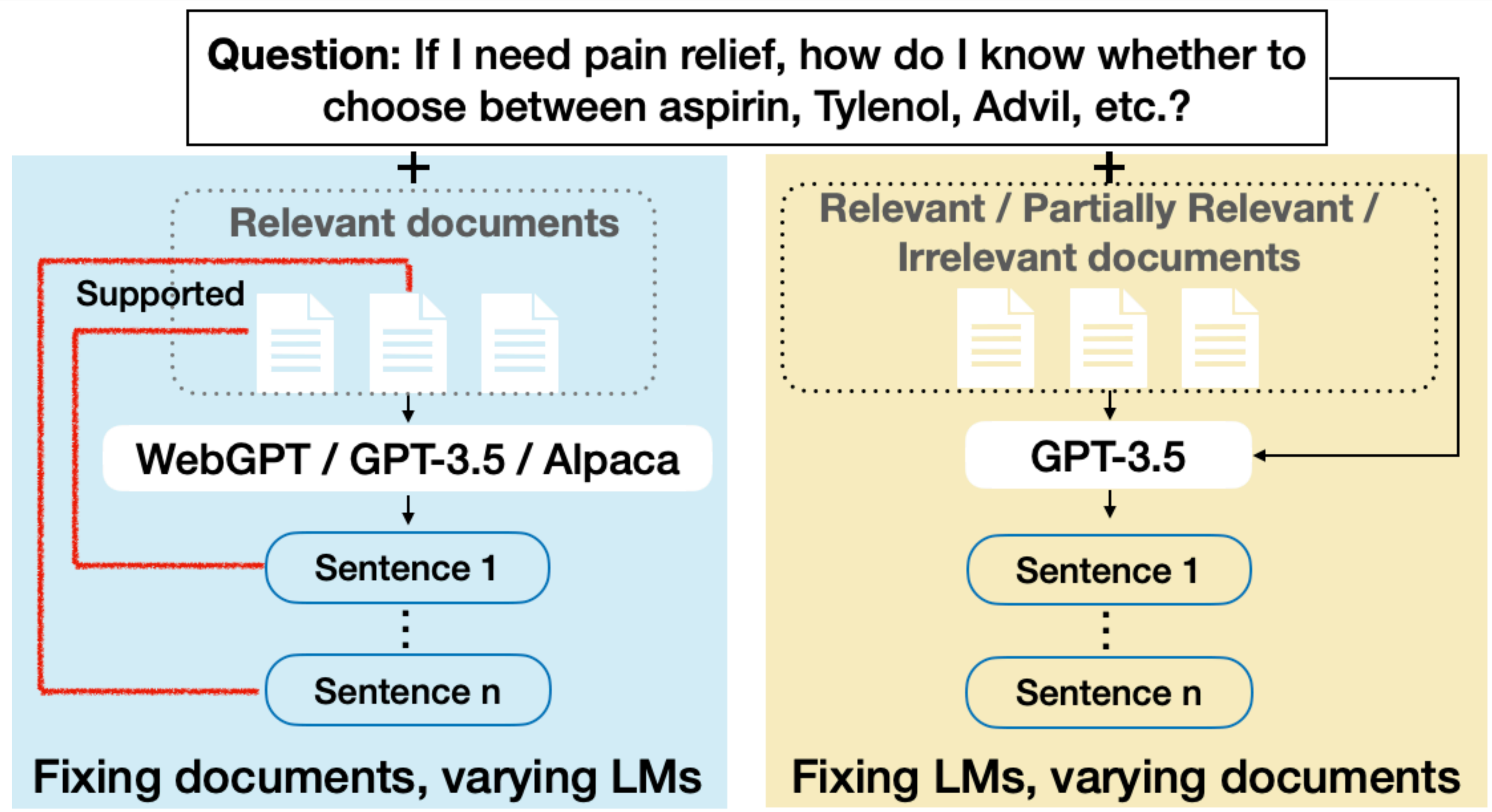
Understanding Retrieval Augmentation for Long-Form Question Answering

Hung-Ting Chen, Fangyuan Xu*, Shane Arora*, Eunsol Choi
University of Texas at Austin

Code & Data

How does the outputs from retrieval-augmented LMs change when varying the in-context documents or the base LM?

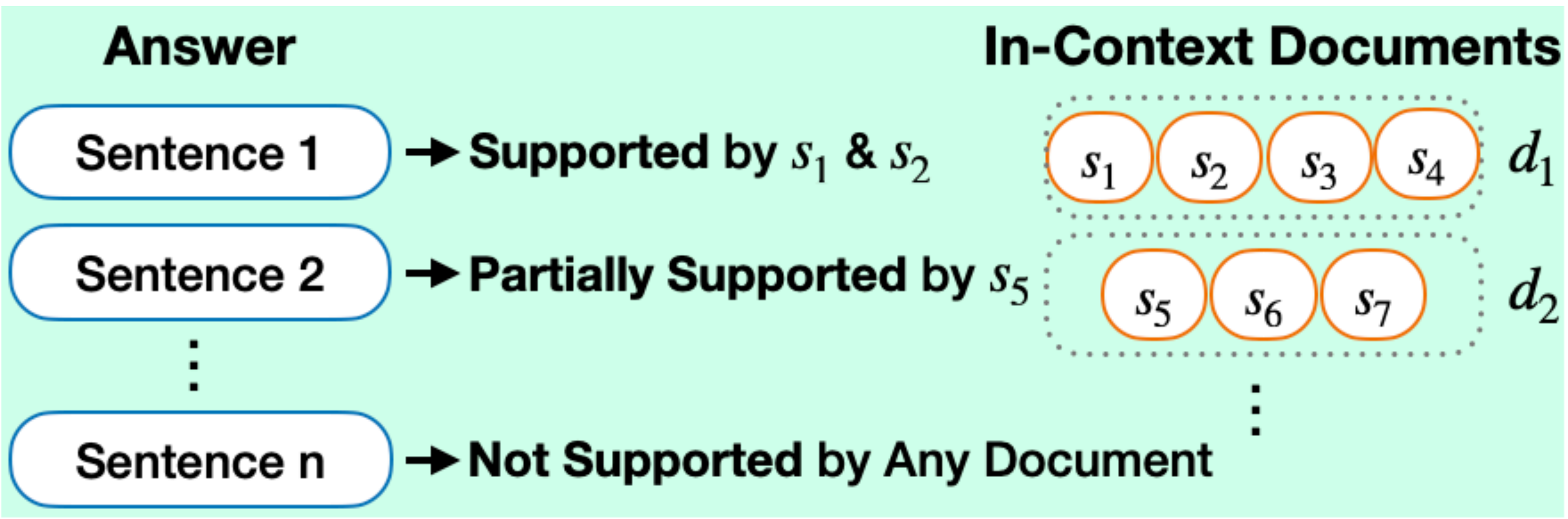
- Can LMs properly use relevant in-context documents and ignore irrelevant ones?
- Does retrieval augmentation in training matter? (Takeaway 1)
- Are there patterns of attribution that guide the designing of RAG systems? (Takeaway 2)
- Can NLI models be used to identify unsupported sentences? (Takeaway 3 & 4)



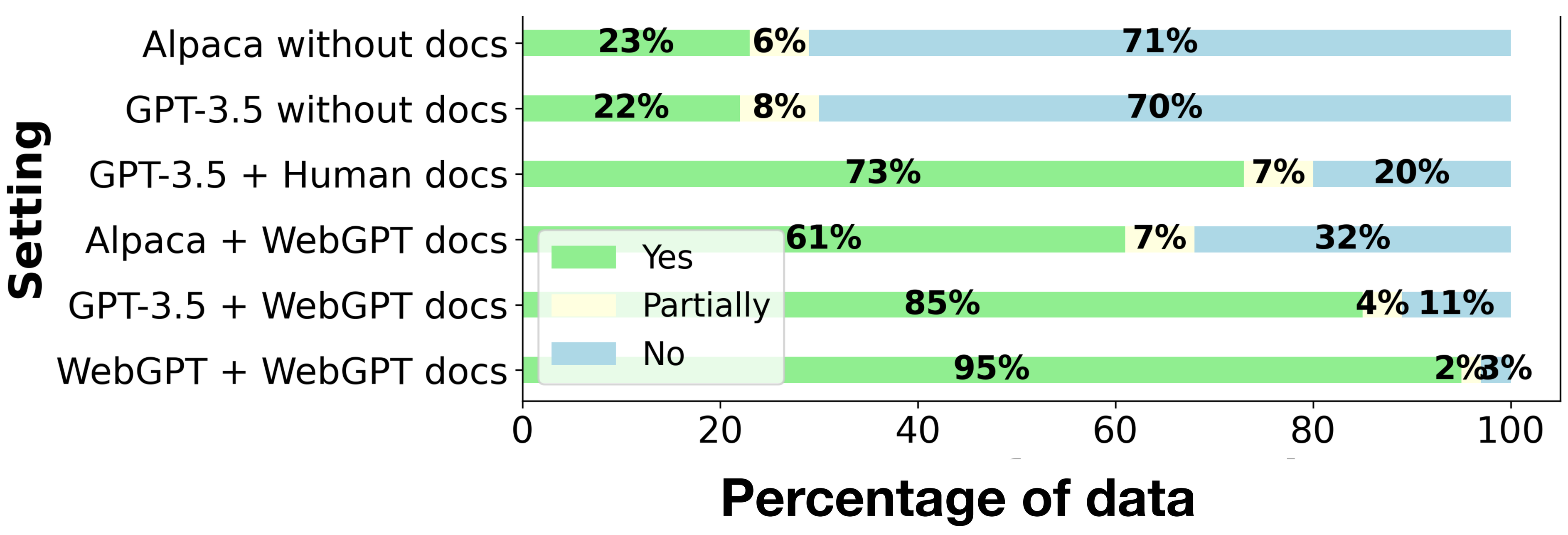
Sentence-level attribution to reference documents (whether each sentence is **supported** by some document)

Surface features: Diversity (Self-BLEU), Perplexity (Analysis in the paper)

SALAD: Sentence-level Attribution of Long-form Answers to evidence Documents



- 3-way human annotation on whether each sentence is supported, and by which sentence in documents
- Question source: ELI-5 (Fan et al., 2019)
- Answers are generated in the settings on the right. WebGPT docs are used in "without docs" settings
- The dataset is open-sourced and can be used for developing automatic attribution evaluation models

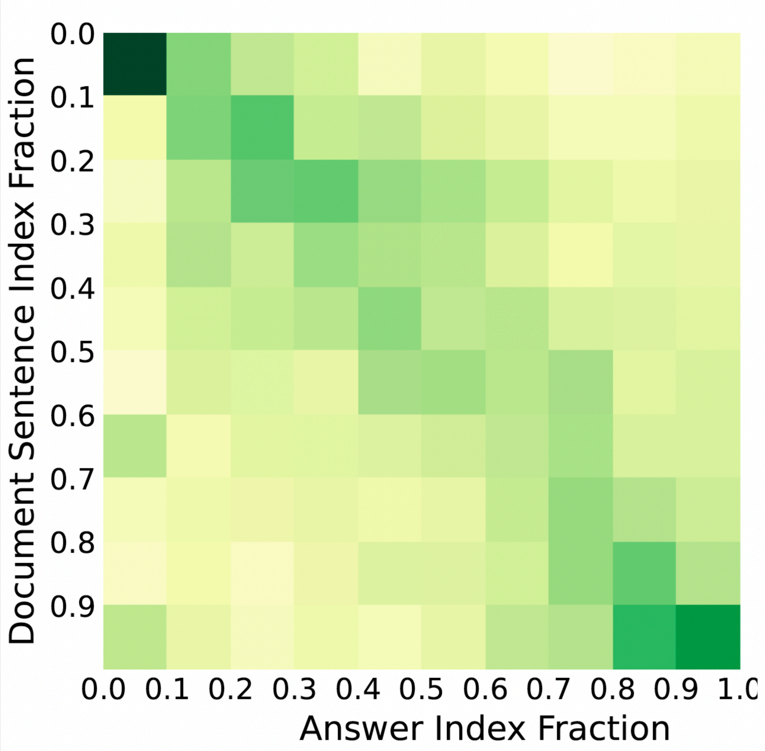


- 100 Answers in each subset = 400 ~ 800 instances
- LMs: WebGPT > GPT-3.5 > Alpaca
- Documents: WebGPT > Human > No Document

Takeaway 1:

An LM trained with retrieval(**WebGPT**) generates sentences that are **most attributed** to in-context evidence documents

➜ Retrieval / attribution during training



Takeaway 2:

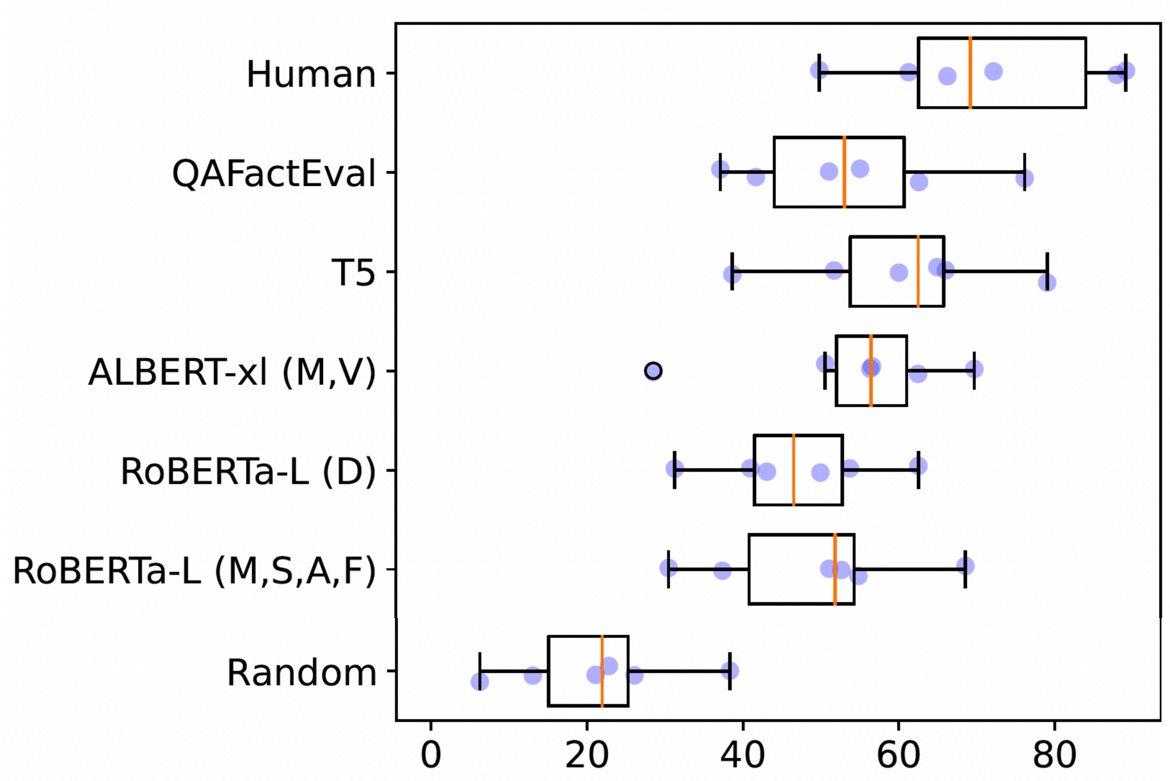
We should carefully add evidence documents to LMs

1. The order of information in documents largely affects the order of outputs.
2. Irrelevant document meaningfully change surface features

Model (+ evidence)	# Sentences	RankGen (↑)	Self-BLEU (↓)	Perplexity (↓)
GPT-3.5	$9.3_{1.5/2.6}$	$12.77_{0.67/1.87}$	$0.71_{0.04/0.06}$	$6.13_{0.02/1.37}$
+Human docs	$6.6_{0.9/1.8}$	$11.89_{0.60/1.86}$	$0.62_{0.04/0.07}$	$10.94_{0.05/3.94}$
+WebGPT docs	$6.8_{0.9/1.8}$	$11.97_{0.60/1.79}$	$0.62_{0.04/0.07}$	$11.63_{0.13/4.16}$
+Bing docs	$6.9_{1.0/1.9}$	$12.13_{0.68/1.91}$	$0.64_{0.04/0.07}$	$9.03_{0.12/3.24}$
+Random docs	$7.6_{1.1/2.1}$	$12.40_{0.67/2.13}$	$0.68_{0.04/0.07}$	$6.76_{0.05/1.86}$

Takeaway 3:

NLI models are promising for identifying generated unsupported sentences



Takeaway 4: LM answers are more supported when documents are more relevant

% Supported Sentences w.r.t.

Model (+ evidence)	Human	WebGPT	Bing	Rand.
GPT-3.5	27.59	34.04	24.79	4.49
+Human docs	65.13	37.99	20.19	3.67
+WebGPT docs	31.37	73.53	20.24	3.90
+Bing docs	24.12	30.17	48.53	4.09
+Random docs	26.13	33.52	26.13	5.19