## **Multi-Review Fusion-in-Context**



amazon

science

Aviv Slobodkin, Ori Shapira, Ran Levy, Ido Dagan

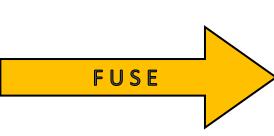
## **Task:** Generate a coherent passage that fuses all and only the highlighted spans within a multi-document source.

#### $\bigstar \bigstar \bigstar \bigstar \bigstar$

There is a metro stop within a couple blocks, which easily gets you to some prime shopping and restaurant areas. The rooms are not large, but the very high ceilings make the rooms feel roomy. We look forward to a return visit to Montreal and to this charming Hotel.

#### $\bigstar \bigstar \bigstar \bigstar \bigstar \bigstar$

Fabulous location, & wonderful service in an extraordinary hotel environment. The hotel staff is extremely nice. Though not big the rooms were beautifully decorated & super comfy. The pool area was out of this world! ...



The location of the hotel was highly appreciated, especially its proximity to the subway, though one reviewer said that there was too much noise in the area. Also, most reviewers appreciate the rooms, stating they were conformable, albeit small.

#### • • •

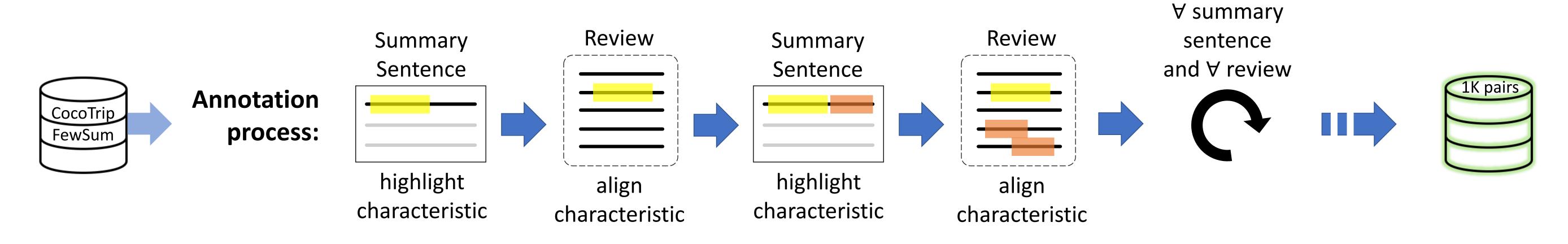
#### 

Though the hotel was located close to the subway, there was too much noise in the area of the hotel. The staff was friendly and did their best to make us comfortable. The breakfast was good enough. The rooms were very disappointing and not comfortable.

- Generalizability:
  *"great sushi"* → *"great food"*
- Aggregation/Abstraction:
  "beds clean" + "bathrooms spotless" → "rooms are clean"
- Conflict merging:

"good service" + "unprofessional staff"  $\rightarrow$  "service mostly good"

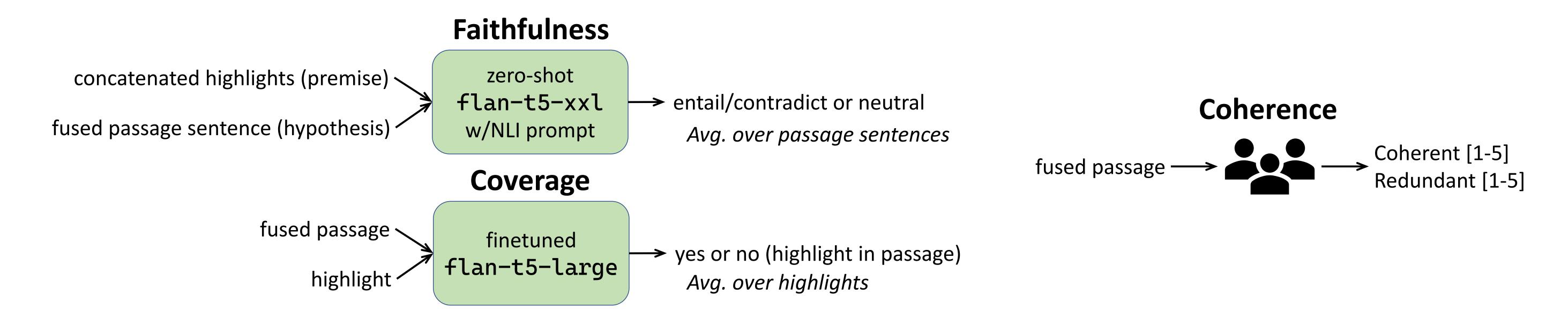
### **Data:** Backwards engineer a multi-document summarization dataset.



80% of summary sentences align to multiple reviews

50% of summary sentences align with non-consecutive spans 🛛 🔴 25% of review tokens are highlighted

# **Evaluation:** Passage should (1) be faithful to highlights, (2) cover all highlights, (3) be coherent.



## **Experiments:** Finetuned flan-t5-large vs GPT-4 with in-context exemplar

	Model	Faithfulness	Coverage	<b>F-1</b>	Coherence	Redundancy
Reviews w/highlights	Flan-T5 <sub>H</sub>	72.8	86.4	79.0	4.3	4.1
Only highlights	Flan-T5 <sub>only-H</sub>	84.6	87.8	86.2	3.6	3.8
Only reviews	Flan-T5 <sub>no-H</sub>	53.7	76.9	63.2	4.1	3.9
Reviews w/highlights	GPT-4	81.6	85.6	83.6	4.7	4.5

- Highlights only  $\rightarrow$  highest faithfulness and coverage
- Adding context  $\rightarrow$  improves coherence
- Without highlights  $\rightarrow$  desired information ignored
- In-context GPT-4 good, but training smaller model is effective
- More work to be done!



### https://fusereviews.github.io