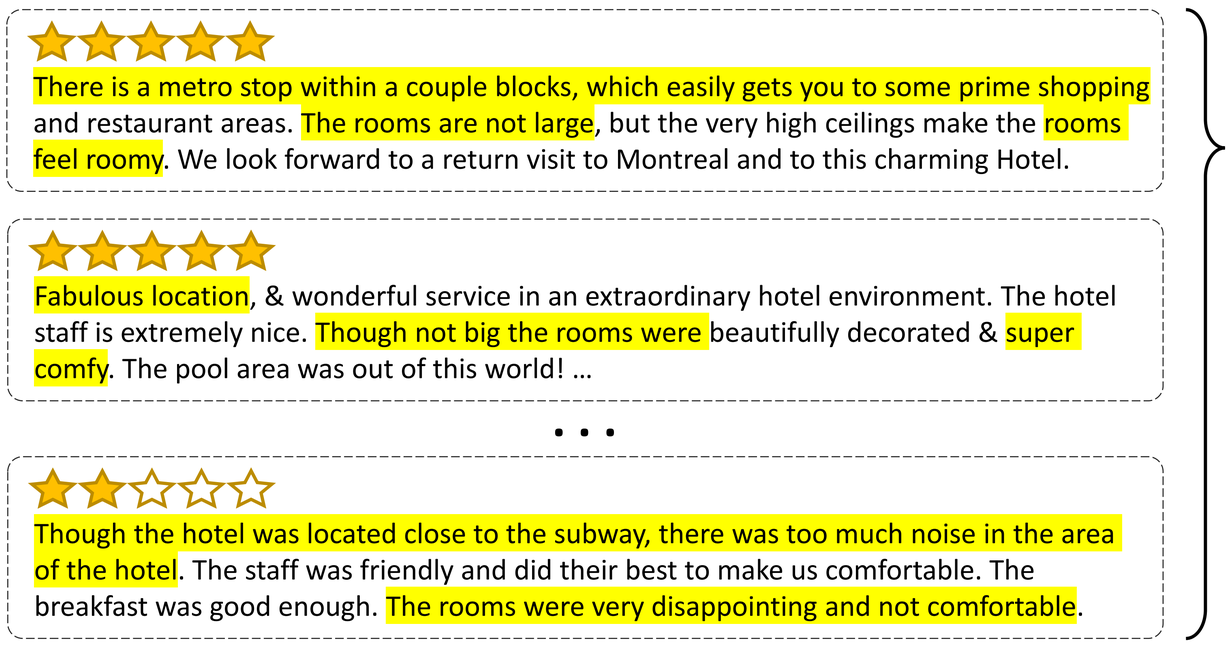
Multi-Review Fusion-in-Context

Aviv Slobodkin, Ori Shapira, Ran Levy, Ido Dagan

Bar-Ilan University
אוניברסיטת בר-אילן

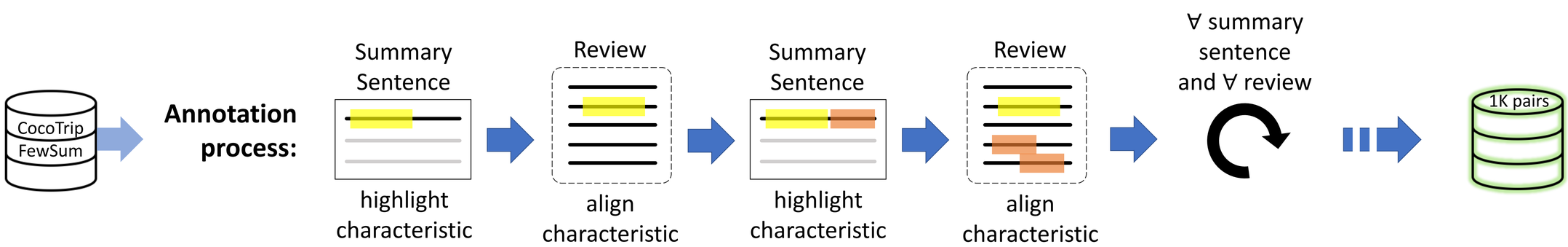
amazon | science

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



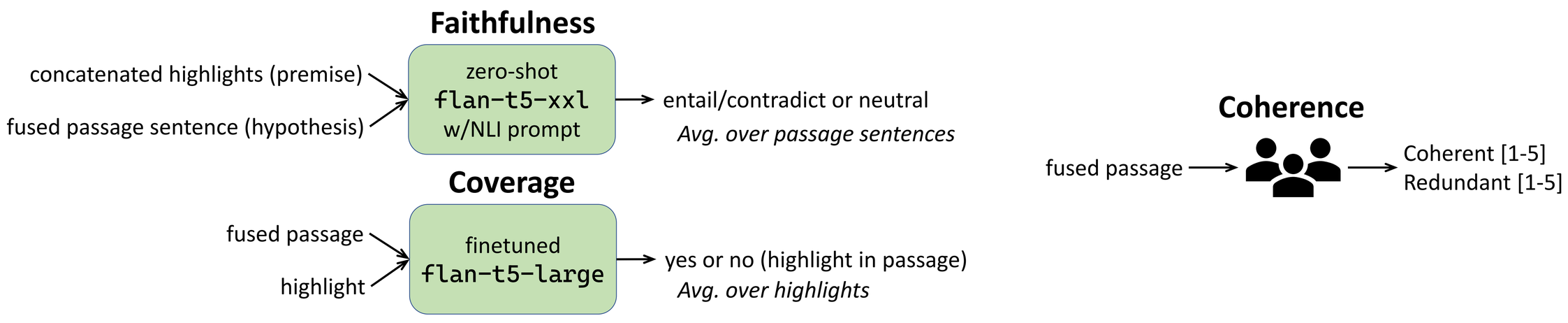
- **Generalizability:**
  "great sushi" → *"great food"*
- **Aggregation/Abstraction:**
  "beds clean" + *"bathrooms spotless"* → *"rooms are clean"*
- **Conflict merging:**
  "good service" + *"unprofessional staff"* → "service mostly good"

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	F-1	Coherence	Redundancy
Reviews w/highlights	Flan-T5$_{H}$	72.8	86.4	79.0	4.3	4.1
Only highlights	Flan-T5$_{only\text{-}H}$	**84.6**	**87.8**	**86.2**	3.6	3.8
Only reviews	Flan-T5$_{no\text{-}H}$	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 → highest faithfulness and coverage
- Adding context → improves coherence
- Without highlights → desired information ignored
- In-context GPT-4 good, but training smaller model is effective
- More work to be done!

`https://fusereviews.github.io`