AmbigDocs: Reasoning across Documents on Different Entities under the Same Name

Yoonsang Lee[1,2], Xi Ye[1], Eunsol Choi[1]

[1]University of Texas at Austin, [2]Seoul National University

Can LMs generate a complete long-form answer correctly distinguishing confusing entities?

All Code Available & Easy Evaluation!

We create AmbigDocs from Wikipedia disambiguation pages

- Existing datasets lack confusing document sets annotation.
- We collect a new multi-document reasoning dataset using an automatic dataset generation pipeline.
- Total of 36K examples and 102K unique entities, where each question has 2.92 answers on average.

Data Example

Question: What is the state where Judge Day was born?

Doc 1 (Charles Bernard Day)
Charles Bernard Day is a ... Born in Dothan, Alabama ...

Doc 2 (Edward William Day)
Edward William Day was a ... Born in Cranston, Rhode Island ...

Doc 3 (William Louis Day)
William Louis Day was a ... Born in Canton, Ohio ...

5-way Answer Categorization

● Correct answer **with** entity disambiguation ; c_O is count of such answers ▲ Correct answer **without** entity disambiguation ; $c_\triangle$ is count of such answers ✕ No answer

Long-form Answer (y)	Per-Doc Answer Recall	Automatic Categorization (m: # of gold answers)	Answer Category
Charles Bernard Day was born in Alabama. Edward William Day was born in Rhode Island. William Louis Day was born in Ohio.	● ● ●	$c_O = m$	**Complete**
William Louis Day was born in Ohio, and Charles Bernard Day was born in Alabama.	● ✕ ●	$1 \leq c_O < m, c_\triangle = 0$	**Partial**
The context does not provide enough information.	✕ ✕ ✕	$c_O = c_\triangle = 0$	**No Answer**
Judge Day was born in Ohio.	✕ ✕ ▲	$c_O = 0, c_\triangle = 1$	**Ambigous**
Judge Day was born in Alabama, Rhode Island, and Ohio.	▲ ▲ ▲	otherwise	**Merged**

✓ Automatic categorization achieves human-human agreement of 0.85 & human-automatic metric agreement of 0.83

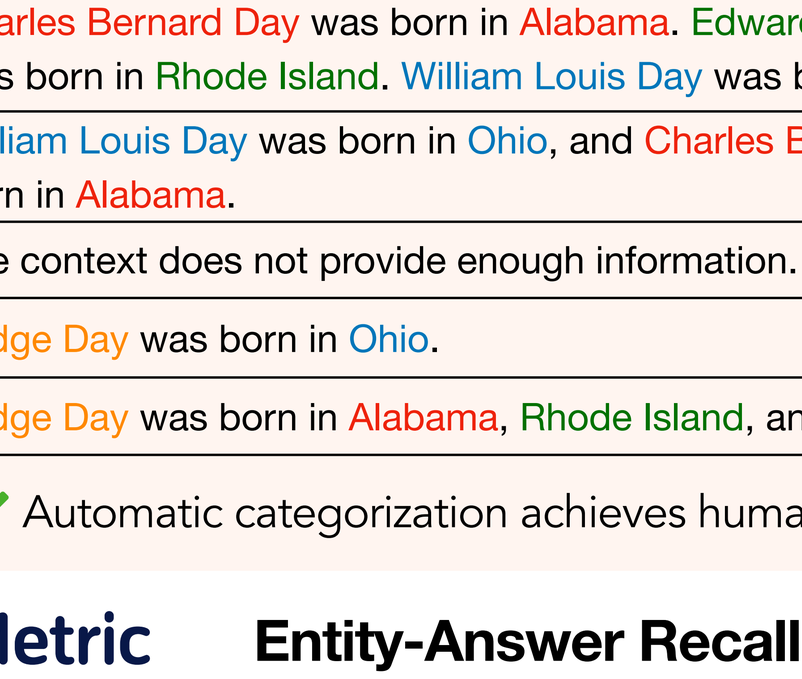
Metric

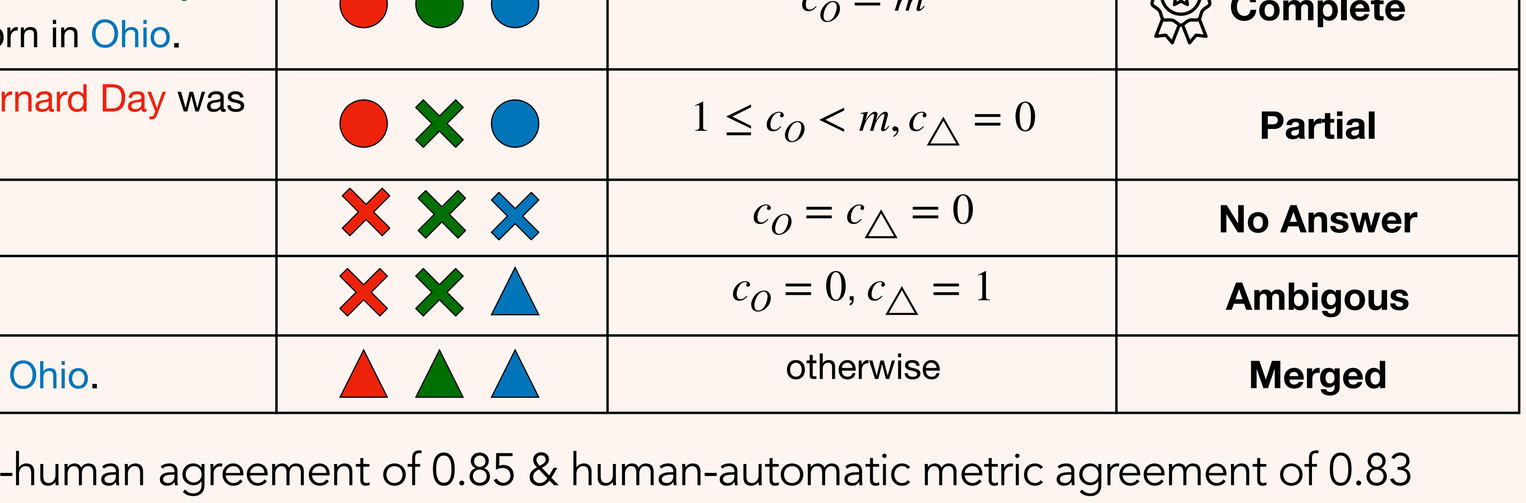
Entity-Answer Recall: $\frac{1}{m}\sum R(DE_i^*, y) \cdot R(a_i^*, y)$ a_i^*: Gold answer / DE_i^*: Disambiguated entity

→ Measure how many answers are paired with its disambiguated entity

Results

Current LMs struggle on multi-document reasoning





- Few-shot learning encourages to generate complete answers
- Mistral and GPT-4 performs the best among current LMs
- Different LMs show different failure modes