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

TEXAS

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

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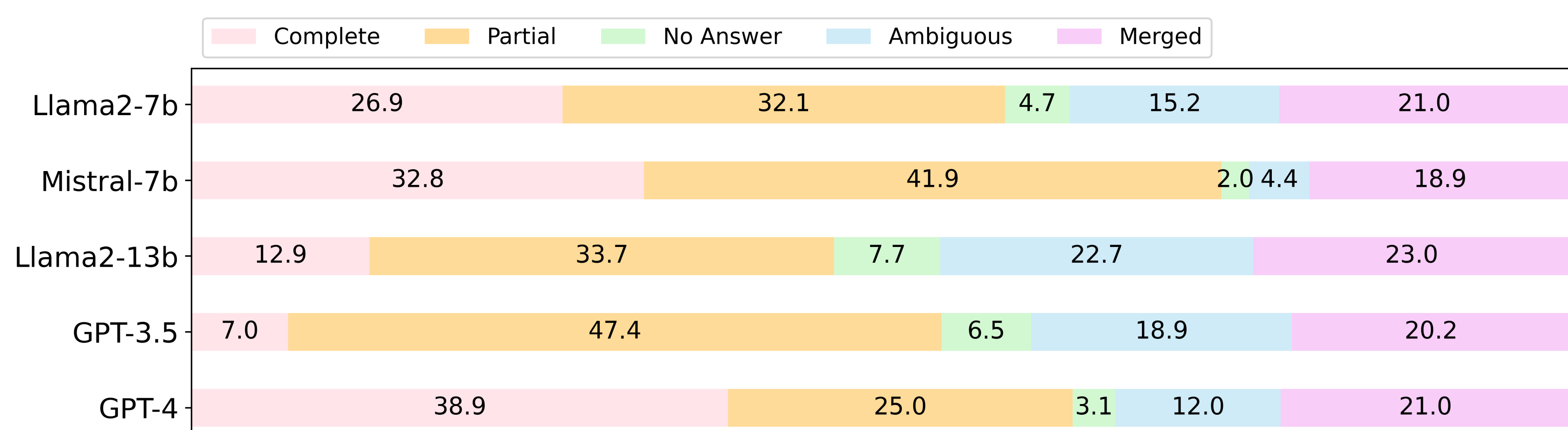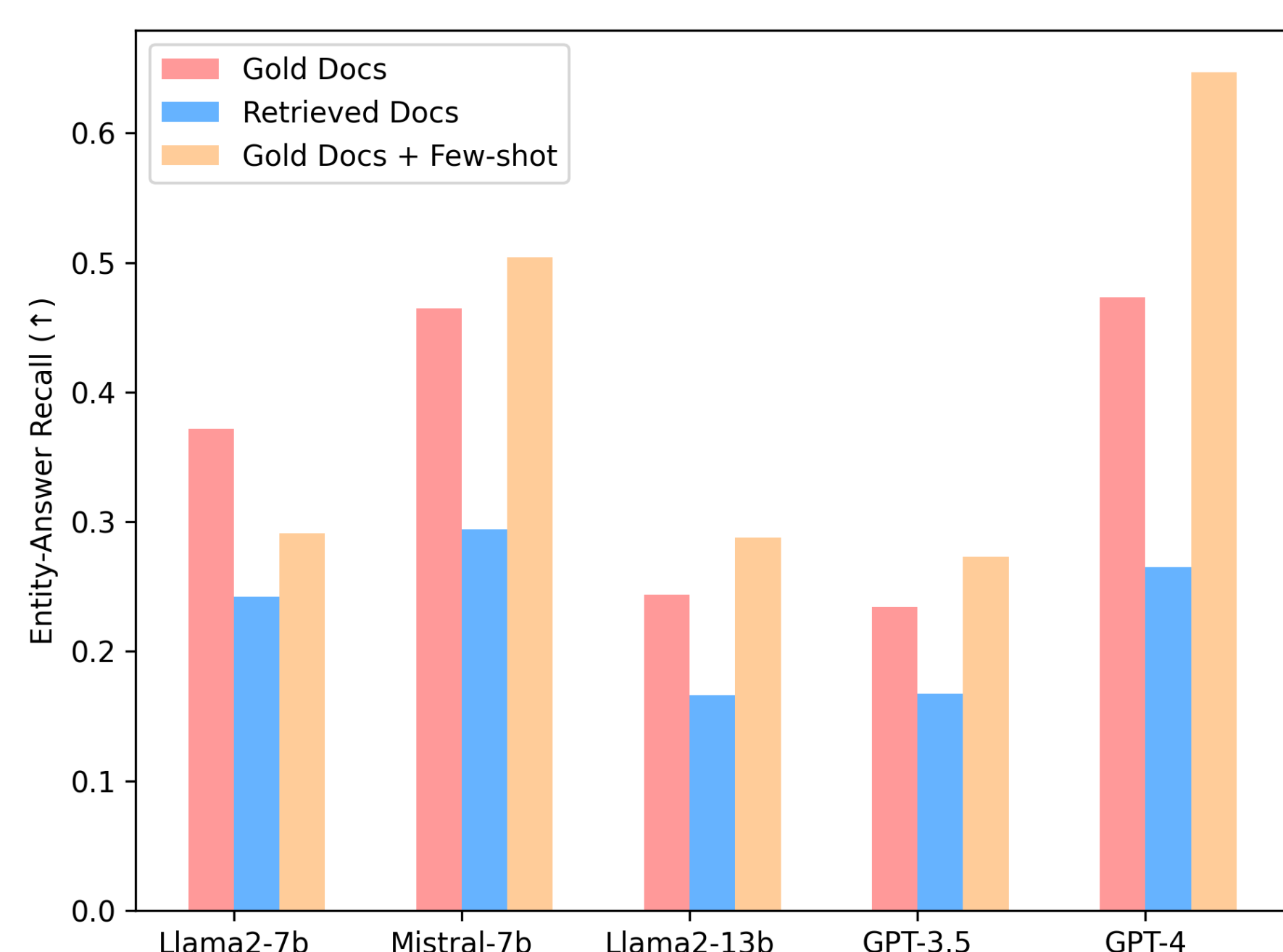
- Few-shot learning encourages to generate complete answers
- Mistral and GPT-4 performs the best among current LMs
- Different LMs show different failure modes

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

TEXAS

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

We create AmbigDocs from Wikipedia disambiguation pages

All Code Available & Easy Evaluation!

- 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

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

● Answer with entity disambiguation ; c_O is number of such answers

▲ Answer without entity disambiguation ; $c_\triangle$ is number of such answers

✕ No answer

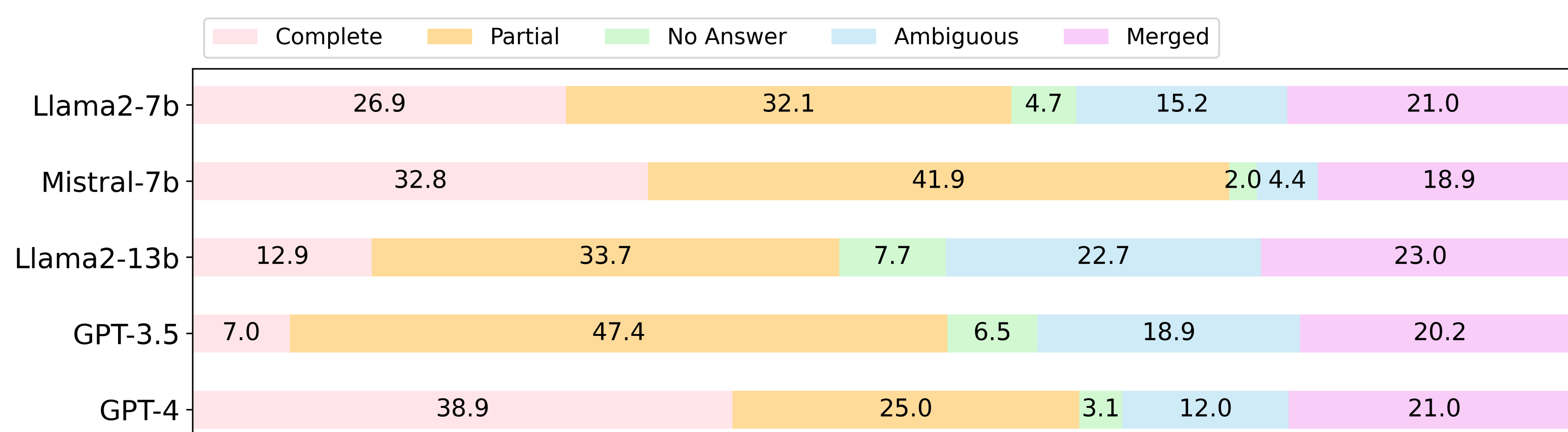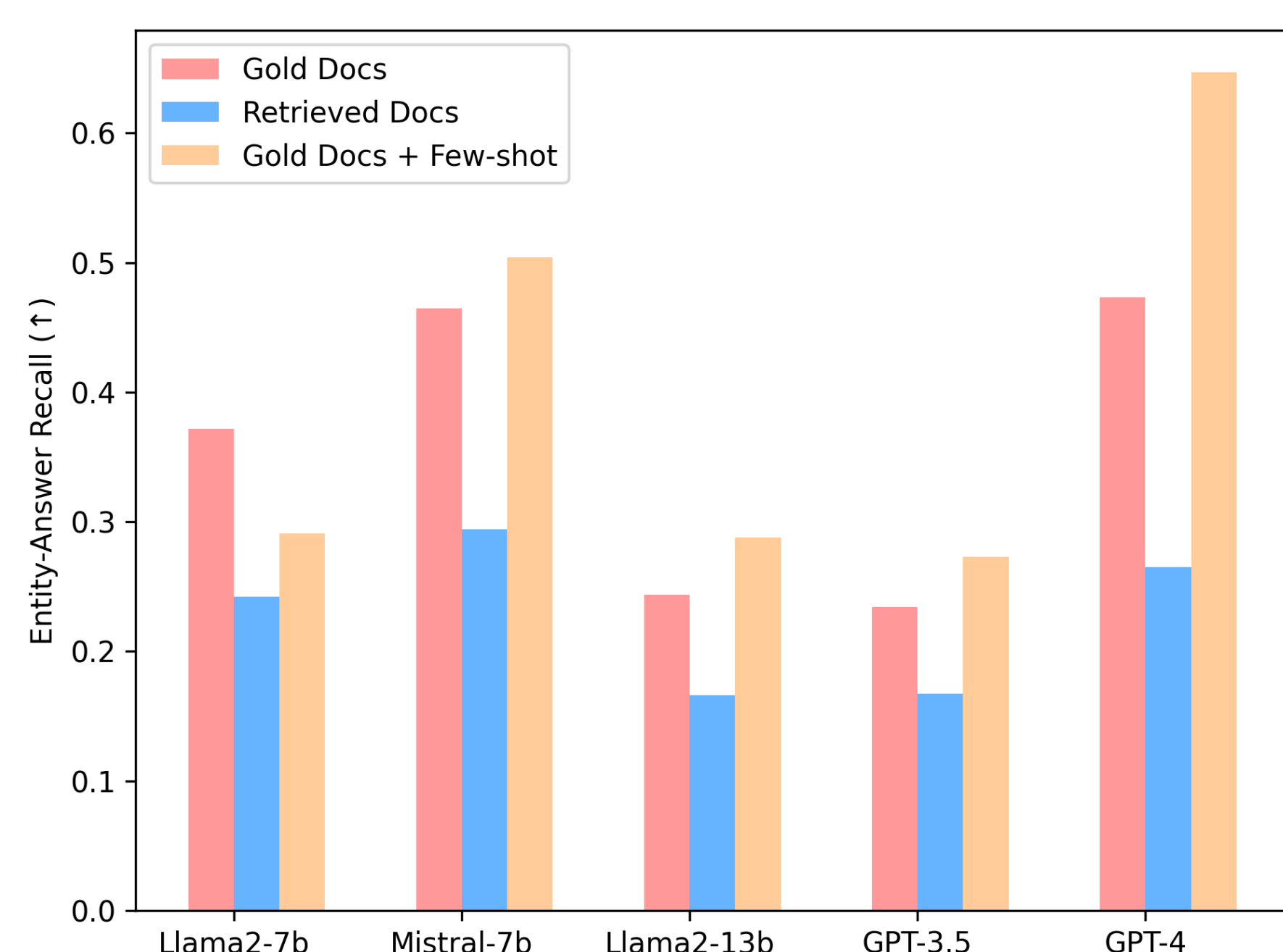✓ Automatic categorization achieves human-human agreement of 0.85 & human-automatic metric agreement of 0.83

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

	Complete	Partial	No Answer	Ambiguous	Merged
Llama2-7b	26.9	32.1	4.7	15.2	21.0
Mistral-7b	32.8	41.9	2.0	4.4	18.9
Llama2-13b	12.9	33.7	7.7	22.7	23.0
GPT-3.5	7.0	47.4	6.5	18.9	20.2
GPT-4	38.9	25.0	3.1	12.0	21.0

- Few-shot learning encourages to generate complete answers

- Mistral and GPT-4 performs the best among current LMs

- Different LMs show different failure modes