UNIVERSITY OF ALBERTA

Accurate and Nuanced Open-QA Evaluation Through Textual Entailment

Peiran Yao, Denilson Barbosa

Department of Computing Science, University of Alberta, Canada

What is Open-QA and Why?

Open-domain question answering is ...

- to answer questions **freely** using natural language, instead of selecting from candidate answers
- **evaluated by matching** system answers with reference answers
- **indicative** of LLM's general, factuality, alignment, uncertainty calibration ... abilities [Anil et al., 2023; Touvron et al., 2023; Yang et al., 2023; Tian et al., 2023]

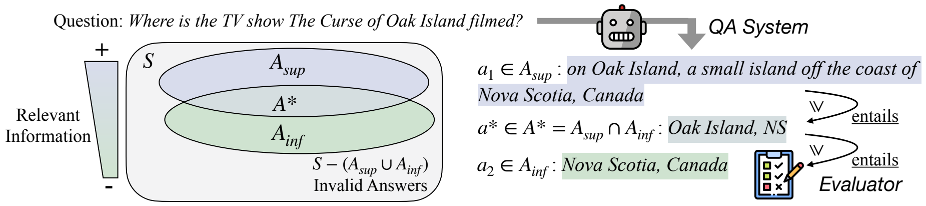
What was wrong with evaluation?

Current evaluations are basic, because

- questions are **under-specified** and reference answers are **non-exhaustive** [Boyd-Graber & Börschinger, 2020]

  exact match and token F_1 are **not considering semantics**, but are still **widely used** [Kamalloo et al., 2023; Wang et al., 2023]
- even semantic similarity models and LLMs are **significantly different** from human judgment [Kamalloo et al., 2023; Wang et al., 2023]

Valid answers are in a hierarchy defined by entailment relations.



Question: *Where is the TV show The Curse of Oak Island filmed?*

QA System

$a_1 \in A_{sup}$: *on Oak Island, a small island off the coast of Nova Scotia, Canada*

$a^* \in A^* = A_{sup} \cap A_{inf}$: *Oak Island, NS*

$a_2 \in A_{inf}$: *Nova Scotia, Canada*

All ~~animals~~ answers are equal, but some ~~animals~~ answers are more equal than others.

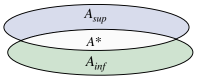
— George Orwell (1945)

More accurate QA evaluation, validated using EVOUNA

- Ground truth: **manual judgment** of NaturalQuestions and TrivialQA answers from 5 QA systems. [Wang et al. 2023]
- Baselines: 4 automatic QA evaluators, including the best LLM prompting strategy as an oracle. [Wang et al. 2023]
- Using entailment, answer correctness are judged **similarly to human judges**.
- Out-of-the-box entailment models **outperform prompt engineering**.

Method	F_1	Acc
Llama-2 (SFT)	94.6	92.3
Llama-2 + NLI (SFT)	94.8	92.6
CVI	84.7	73.5
Entailment (0-shot)	93.5	90.2

- As a zero-shot method, our method is **comparable to fine-tuned** evaluators; adding **entailment feature also helps** with fine-tuning QA evaluators.



$A_{sup} \oplus A_{inf}$: valid answers missed by default evaluators

6.5% - 10.1% underestimation of QA accuracy

	DPR-FiD		InstructGPT		ChatGPT		GPT-4		BingChat	
Evaluator	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
Lexical Match†	92.0	89.7	86.9	84.8	85.0	80.3	87.6	82.5	87.8	82.3
BERTScore†	83.5	75.1	77.6	69.5	81.2	72.8	84.3	76.0	77.5	67.5
GPT-3.5†	**95.3**	**93.6**	87.2	84.1	86.9	82.2	86.8	80.9	77.3	69.5
Entailment	94.8	92.5	**92.7**	**90.2**	**92.6**	**88.9**	**93.8**	**90.1**	**92.6**	**88.1**
Entailment (small)	91.5	88.5	88.0	85.4	87.7	83.2	89.9	85.0	87.8	82.0
GPT-3.5 (best prompting)†	95.5	93.9	88.3	84.5	89.4	84.5	91.2	86.0	87.1	80.4
Another Human†	97.4	96.3	97.8	96.8	96.5	95.6	97.9	96.6	97.2	95.5
on EVOUNA-NaturalQuestions										
Lexical Match†	91.8	94.7	94.8	92.3	95.2	92.3	94.8	91.1	94.1	89.8
BERTScore†	75.1	65.5	84.1	75.7	88.4	80.8	90.5	93.5	88.3	80.4
GPT-3.5†	**97.3**	**95.7**	94.2	91.2	95.5	92.5	95.7	92.4	88.2	80.9
Entailment	96.8	94.7	**96.6**	**94.2**	**96.6**	**94.2**	**97.4**	**95.3**	**95.9**	**92.5**
Another Human†	100	100	99.6	99.4	99.2	98.8	99.2	99.8	99.9	95.5
on EVOUNA-TriviaQA										

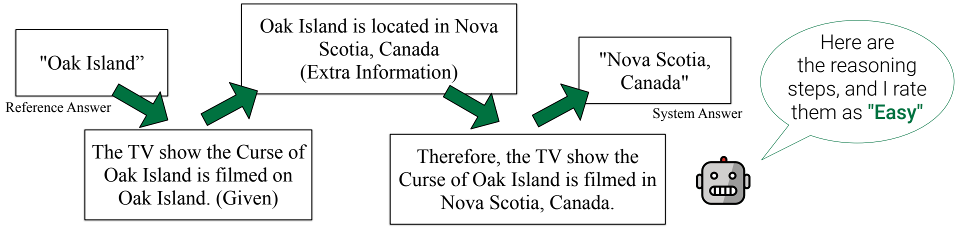
Explaining the entailment enables partial or bonus marks.

How far is the gap between reference answer and system answer?

- Let LLM make **verbal inferences** about why the entailment relation holds.
- Quantify the gap based on the verbal inferences, and use that as a **non-binary** and **unbounded** score for answer correctness!

Method	AUC
Inference + LLM Score	0.91
Inference + #Steps	0.91
LLM Score	0.88
F_1 Score	0.78

Table 3: Using LLM to explain the inference process behind how gold answers entail the system answers leads to higher AUROC in predicting human judgements on NQ, making it a good candidate for partial marks.



"Oak Island"

Reference Answer

Oak Island is located in Nova Scotia, Canada (Extra Information)

"Nova Scotia, Canada"

System Answer

The TV show the Curse of Oak Island is filmed on Oak Island. (Given)

Therefore, the TV show the Curse of Oak Island is filmed in Nova Scotia, Canada.

Here are the reasoning steps, and I rate them as "Easy"

Small Prints

Testing Entailment

- Question-answer pairs first rewritten as declarative statements using GPT-3.5 or Llama-2 (small).
- Entailment relations between statements judged by GPT-3.5 or DeBERTa-NLI (small).

Does the hierarchy really exist?

- The higher the rank, the more likely the system answer is correct.
- Tested by one-tailed Fisher's exact test.

Does LLM have stability?

- Likely yes. Cohen's Kappa shows almost perfect agreement across 4 random seeds.

SUPPORTED BY

Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada

Canada

Scotiabank.

Code | Data | Issues
U-Alberta / QA-partial-marks

Preprint
arXiv: 2405.16702

ACL 2024 Bangkok, Thailand