PRINCETON NLP

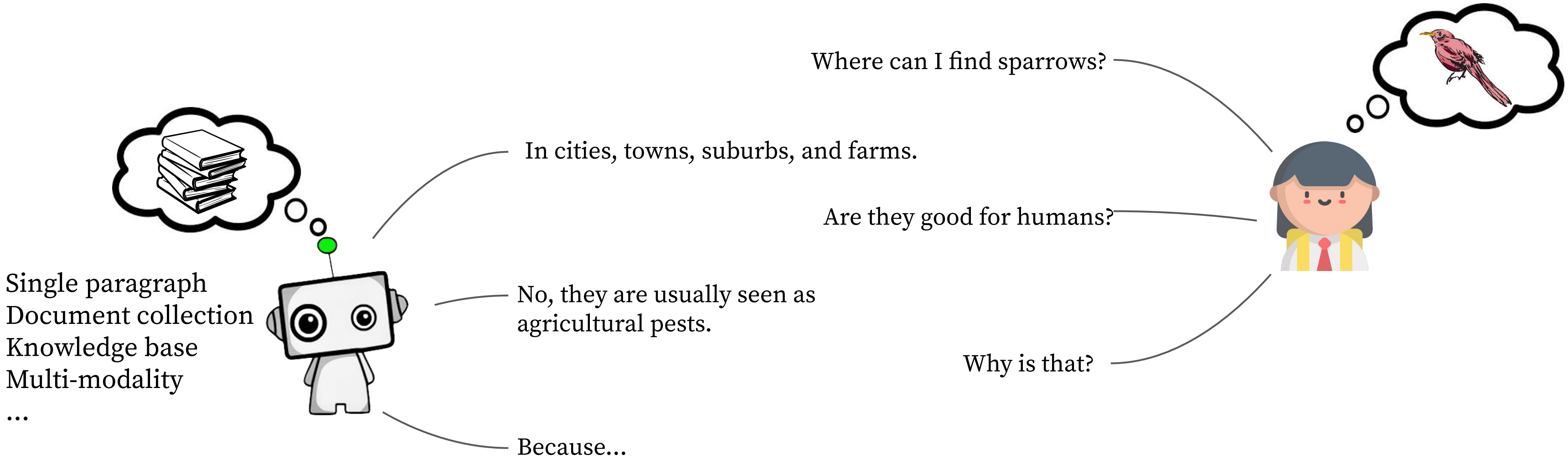
Ditch the Gold Standard: Re-evaluating Conversational Question Answering

Huihan Li*, Tianyu Gao*, Manan Goenka, Danqi Chen

ACL 2022

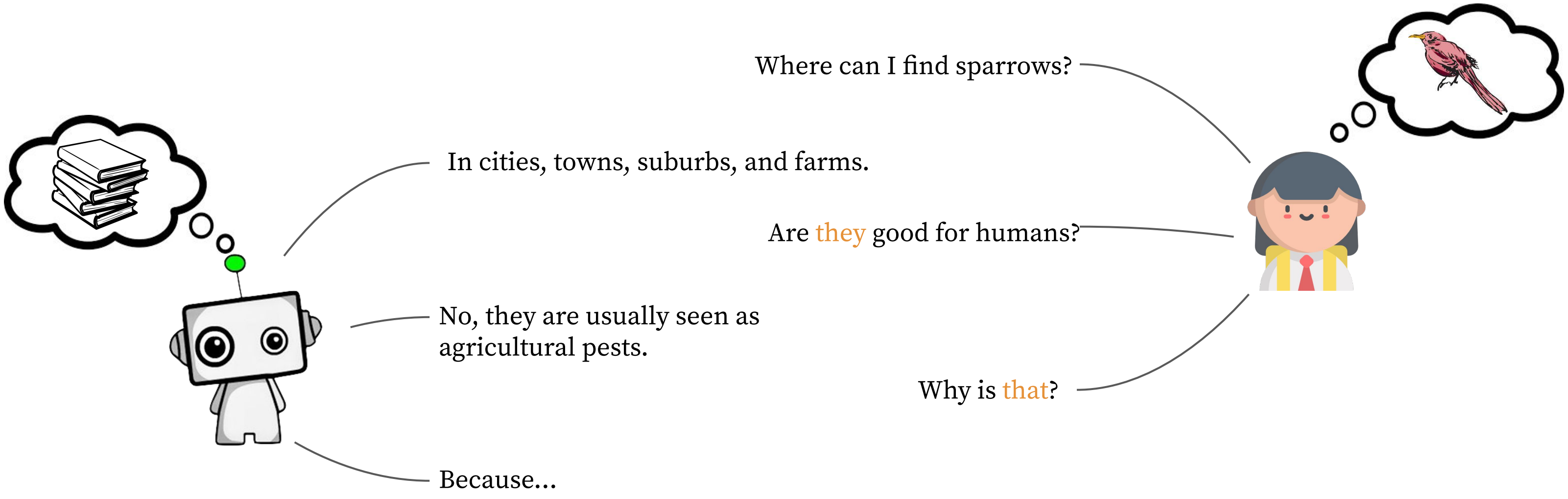
Background: Conversational Question Answering

Conversational QA aims to build machines that answer human questions in
information-seeking conversations



Background: Conversational Question Answering

Conversational QA aims to build machines that answer human questions in
information-seeking conversations



Challenge: Questions need to be understood from ***conversation history***

Conversational Question Answering Datasets

QuAC (Choi et al., 2018)
CoQA (Reddy et al., 2019)
DoQA (Campos et al., 2020)

Single paragraph

OR-QuAC (Qu et al., 2020)
TopiOCQA (Adlakha et al., 2021)
QReCC (Anantha et al., 2021)

Document collection

Visual Dialog (Das et al., 2017)
ShARC (Saeidi et al., 2018)
CSQA (Saha et al., 2018)

Knowledge bases or other modalities

Flaws in Conversational QA Evaluation

Benchmarks consist of *pre-collected* human-human conversations
QuAC (Choi et al. 2018)

Topic: Spandau Ballet (English pop band)

What was the band's first success album at the international level?

"Parade" from 1984.

What songs were in it?

"Only When You Leave".

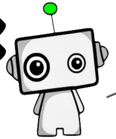
Flaws in Conversational QA Evaluation

Topic: Spandau Ballet (English pop band)

Benchmarks consist of *pre-collected* human-human conversations

What was the band's first success album at the international level?

Gold answers are always provided during evaluation even when predictions are wrong

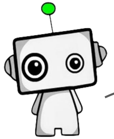


~~They achieved platinum status.~~

"Parade" from 1984.

What songs were in it?

Problem: Models *do not* have access to gold answers in real-world human-machine conversations!



"Only When You Leave".

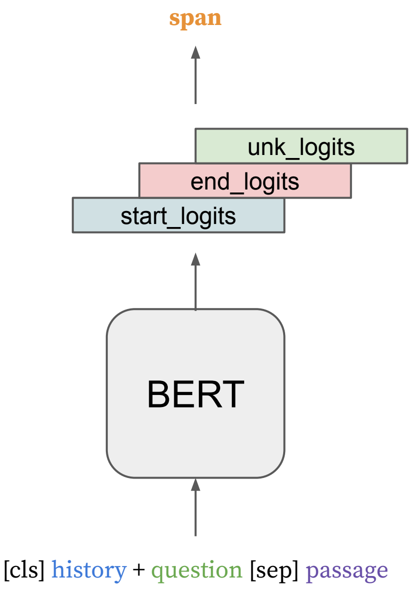
Questions of Interest

- How do current conversational QA models perform in **human-machine conversations**?
- Can current automatic evaluation **reflect human judgment**?
- How can we **improve current automatic evaluation**?
- What are important for a **good** conversational question answering **model**?

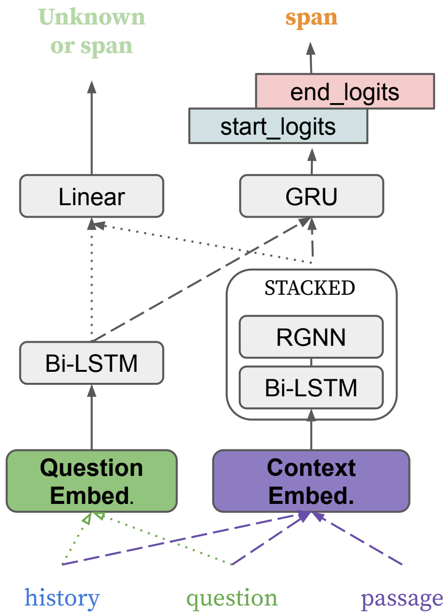
Models

We experiment with four state-of-the-art ConvQA models

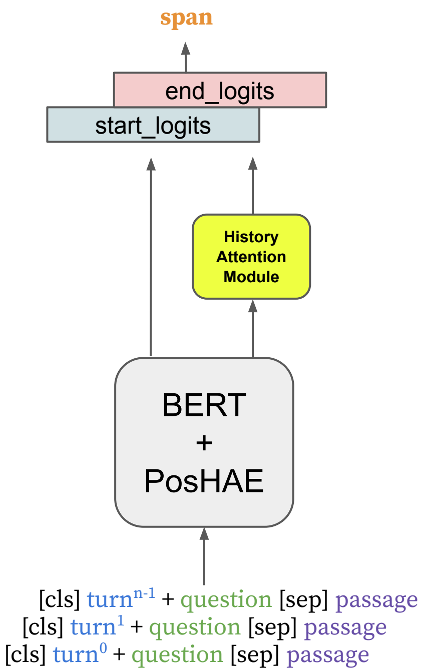
BERT (Devlin et al. 2019)



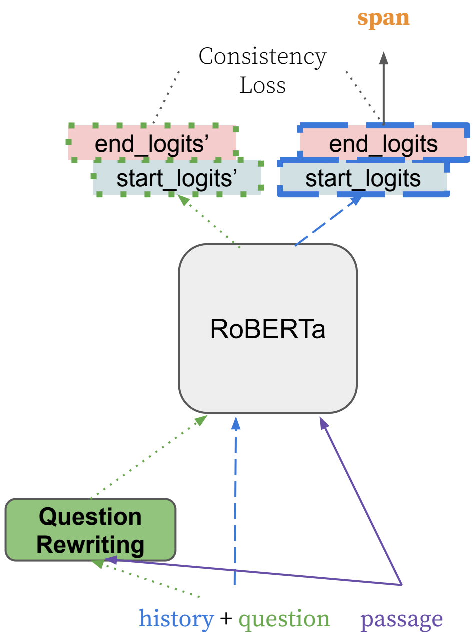
GraphFlow (Chen et al. 2019)



HAM (Qu et al. 2019)



ExCorD (Kim et al. 2021)



What do human-machine conversations look like?

Topic: Spandau Ballet (English pop band)

Humans can *adjust* the next question based on the model prediction.

Different models might result in **different** conversations.

What was the band's first success album at the international level?

They achieved platinum status.

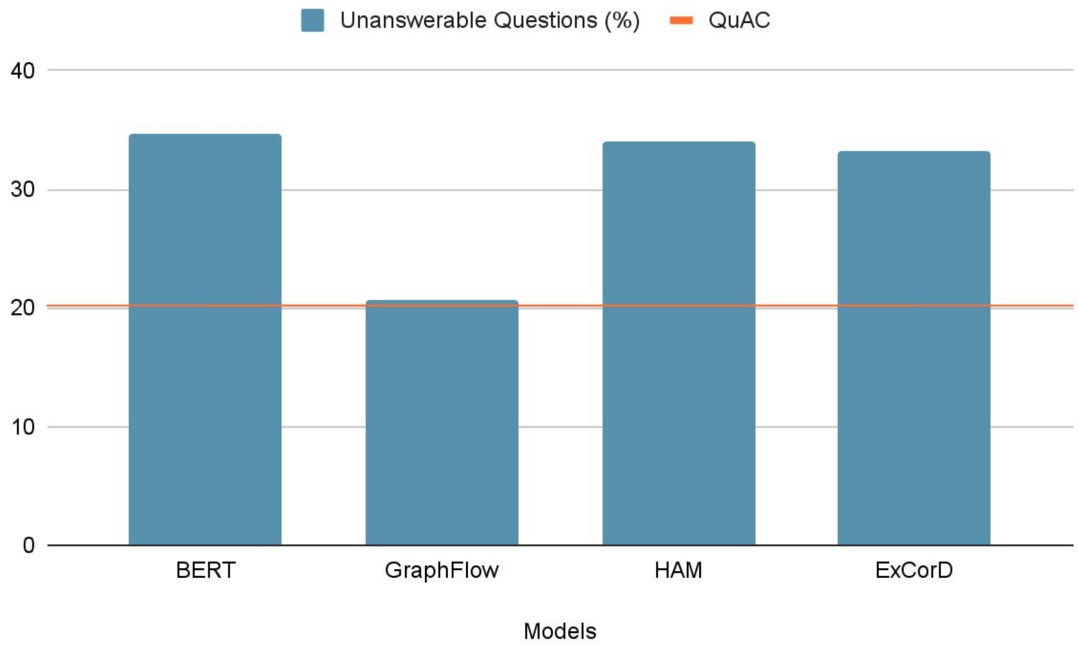
What year did this happen?

1985.

Human Evaluation

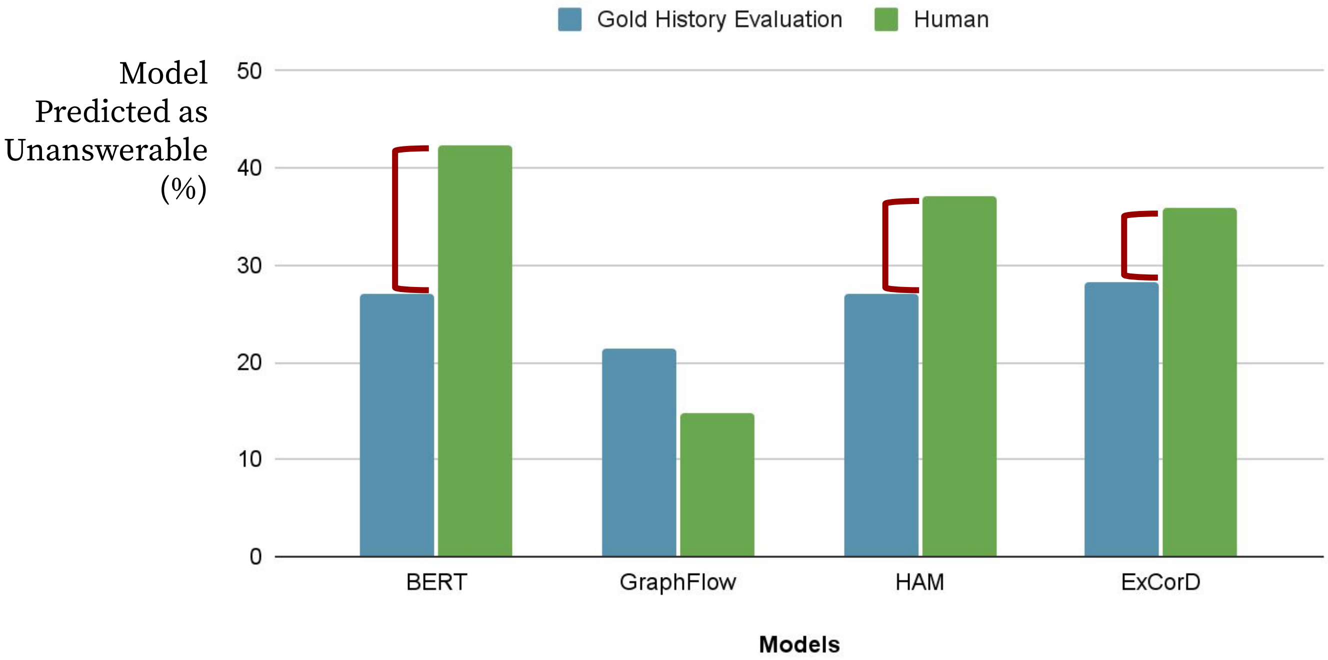
- 100 QuAC (Choi et al. 2018) dev set evidence passages
- 1,446 human-machine conversation
- 15,059 question-answer pairs
- Released on `https://github.com/princeton-nlp/EvalConvQA/`

1. Are *human-machine* conversations similar to *human-human* conversations?



Humans ask **more unanswerable questions** in *human-machine conversations* than in human-human conversations

1. Are *human-machine* conversations similar to *human-human* conversations?



Models **predict more questions as unanswerable** in *human-machine conversations* than in gold history evaluation

1. Are *human-machine* conversations similar to *human-human* conversations?

Title: Superstar Billy Graham
Section title: Disputes with the McMahons

Q1: What disputes did he have? ← We provide the first question from QuAC
A1: *CANNOTANSWER*

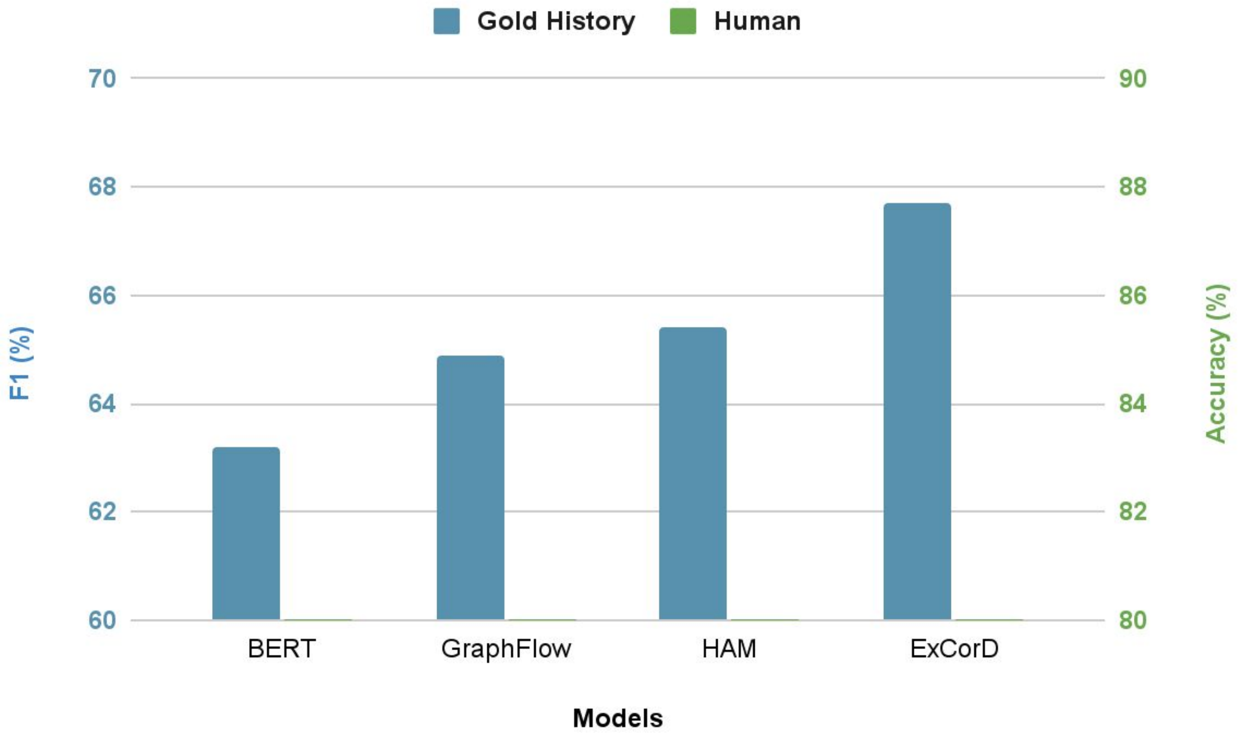
Q2: Where is Billy from? ← Asks an unanswerable question
A2: *CANNOTANSWER*

Q3: What else is interesting about this article? ← Asks an open question
A3: *Graham personally sued Zahorian and the WWF*

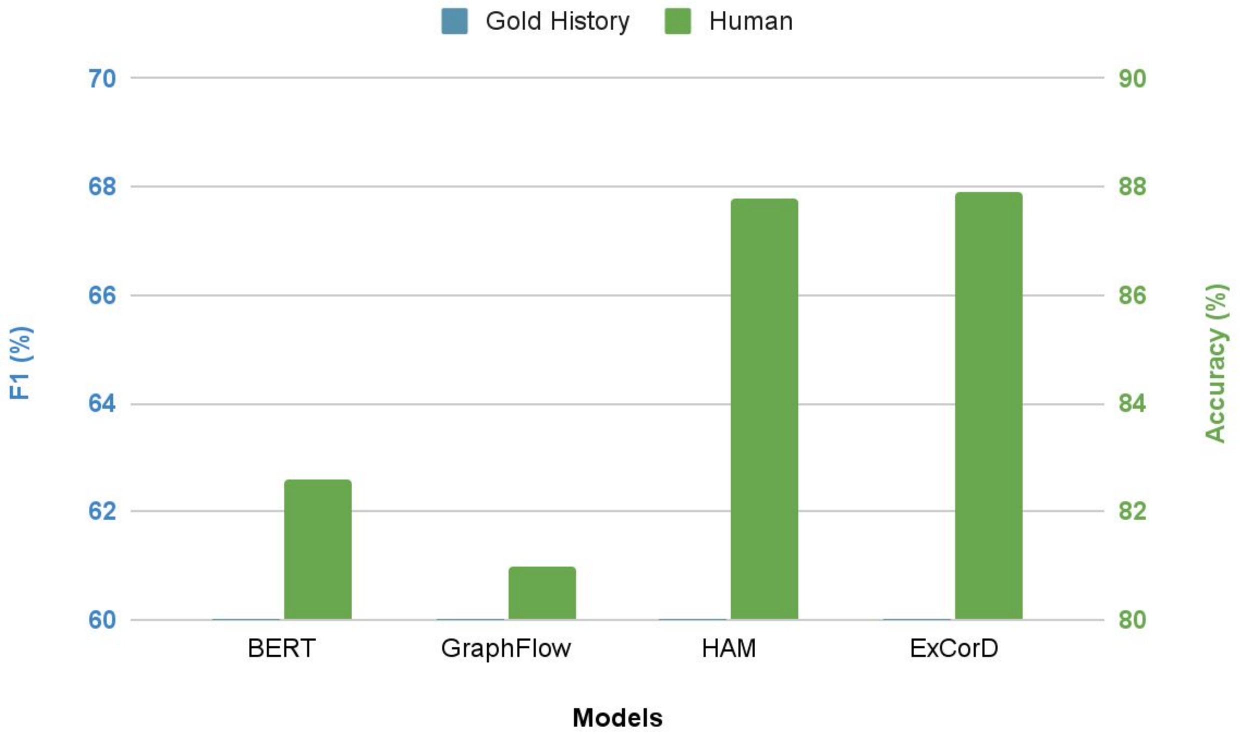
Because of *low-quality model answers*, humans ask *more* **unanswerable questions** and **open questions**

2. Does gold history evaluation agree with human judgement?

Averaged
word-level F1
from QuAC

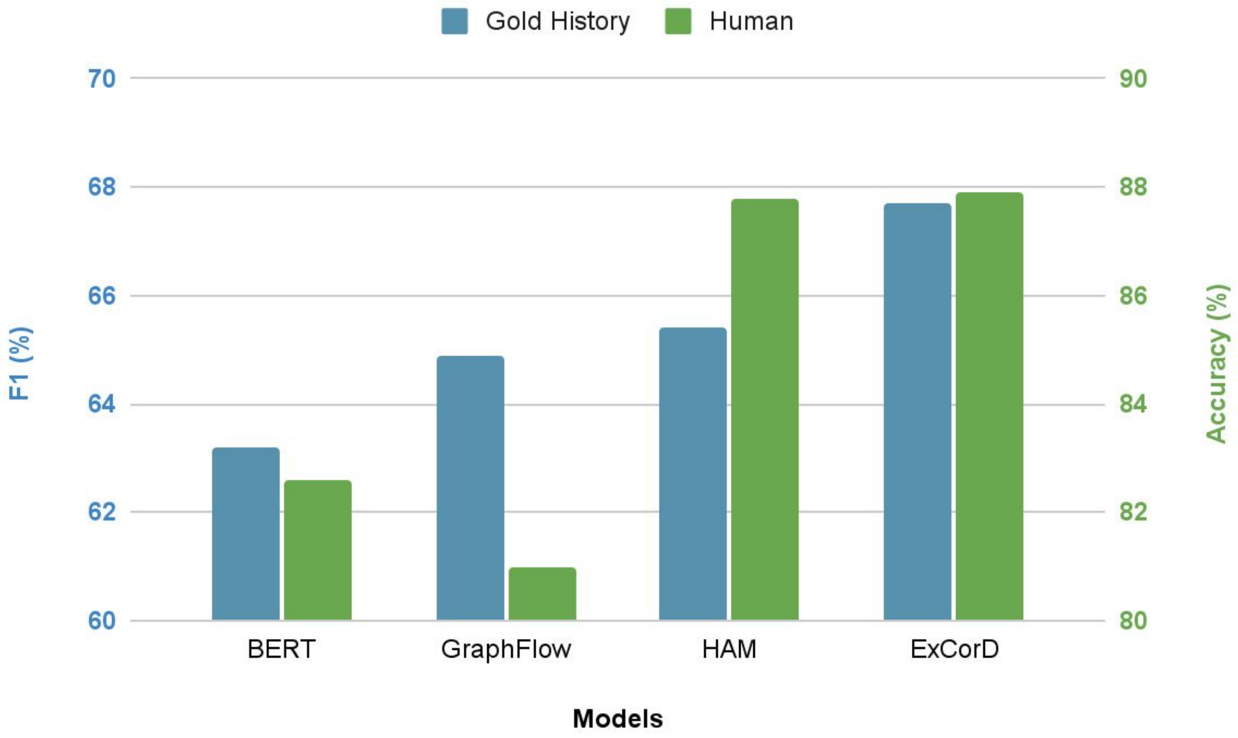


2. Does gold history evaluation agree with human judgement?



Human marks either “correct” or “incorrect”

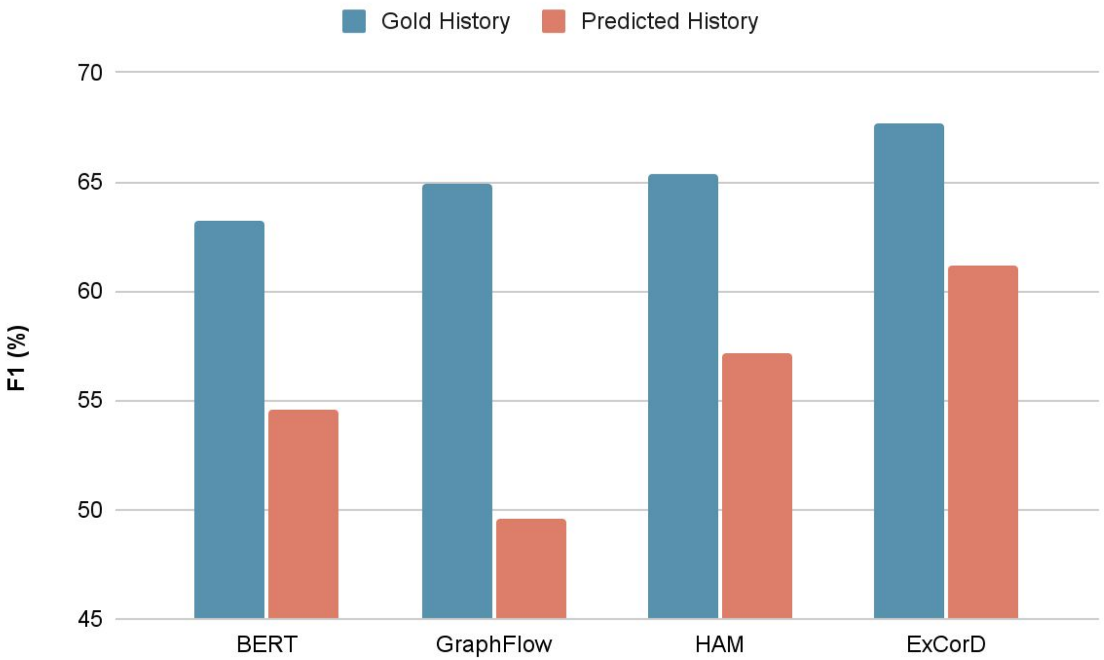
2. Does gold history evaluation agree with human judgement?



Gold history evaluation ranks models **differently** from human judgement!

Can we do better in automatic evaluation?

1. Can we simply use the models' prediction in history? (Mandya et al., 2020; Siblini et al., 2021)



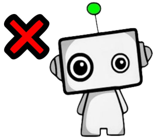
Predicted history evaluation has a **large performance drop** from gold history evaluation

… which is expected from low-quality model predictions

Simply using model predictions may **invalidate** the next question

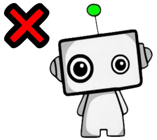
Topic: Spandau Ballet (English pop band)

What was the band's first success album at the international level?



They achieved platinum status.

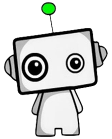
What songs were in it?



...??

Unresolvable Coreference

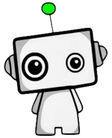
What was their first song?



CANNOTANSWER

Gold: *Punch in the Face*

How did $\boxed{\textit{it}}$ fair?



...??

A pronoun or a definite article reference (eg. *the film*) that is **not resolvable** from the conversation history

Rewrite question with unresolvable coreference

Topic: Spandau Ballet (English pop band)

What was the band's first success album at the international level?

They achieved platinum status.

Gold Answer: "Parade" from 1984.

What songs were in it?

What songs were in "Parade"?

First single "Only When You Leave".

Gold Answer: "Only When You Leave"

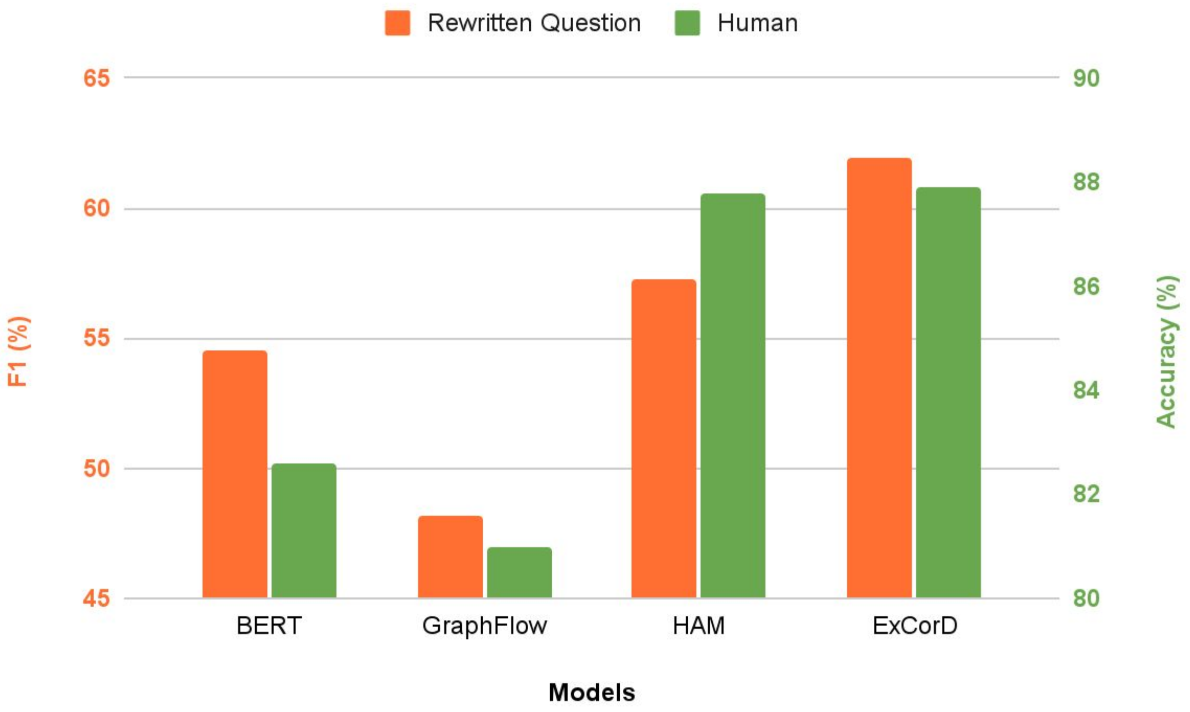
How did it do on the charts?

Became the band's last American hit.

- Use coreference resolution model (Lee et al. 2018) to resolve the entity-of-interest in current question using the *gold history* and *predicted history*, separately
- If the resolutions do not match, substitute the entity-of-interest with its mention in *gold history*
- If the resolutions match, no need to rewrite

2. Is rewritten question evaluation closer to human judgement?

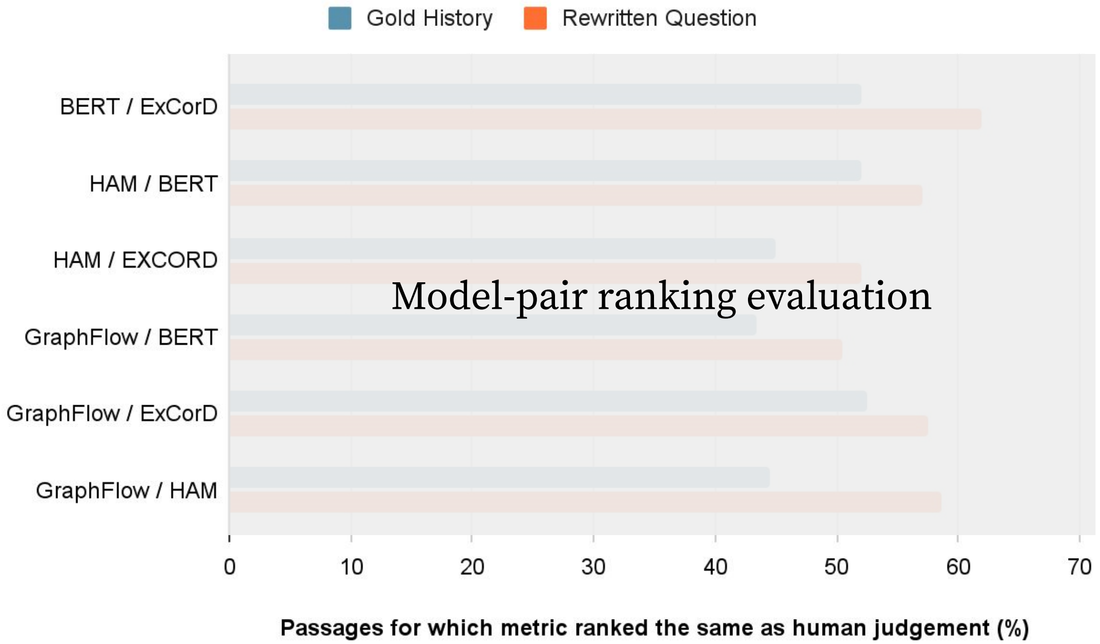
Averaged word-level F1 from QuAC



Human marks either “correct” or “incorrect”

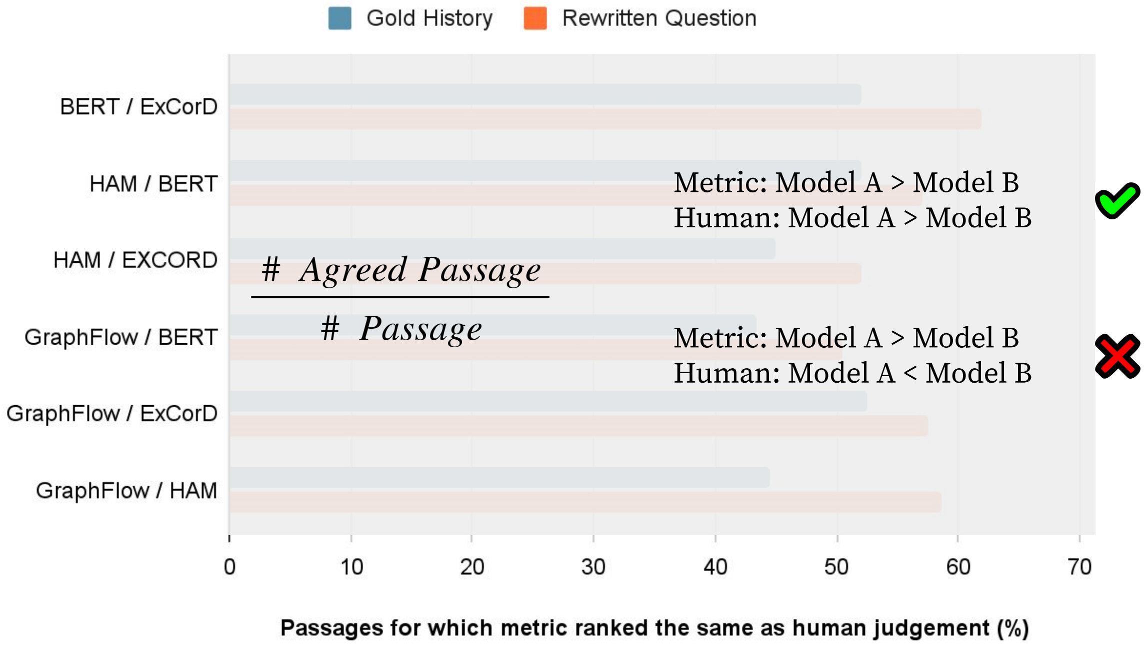
Rewritten question evaluation ranks the models **the same way** as human judgement

2. Is rewritten question evaluation closer to human judgement?

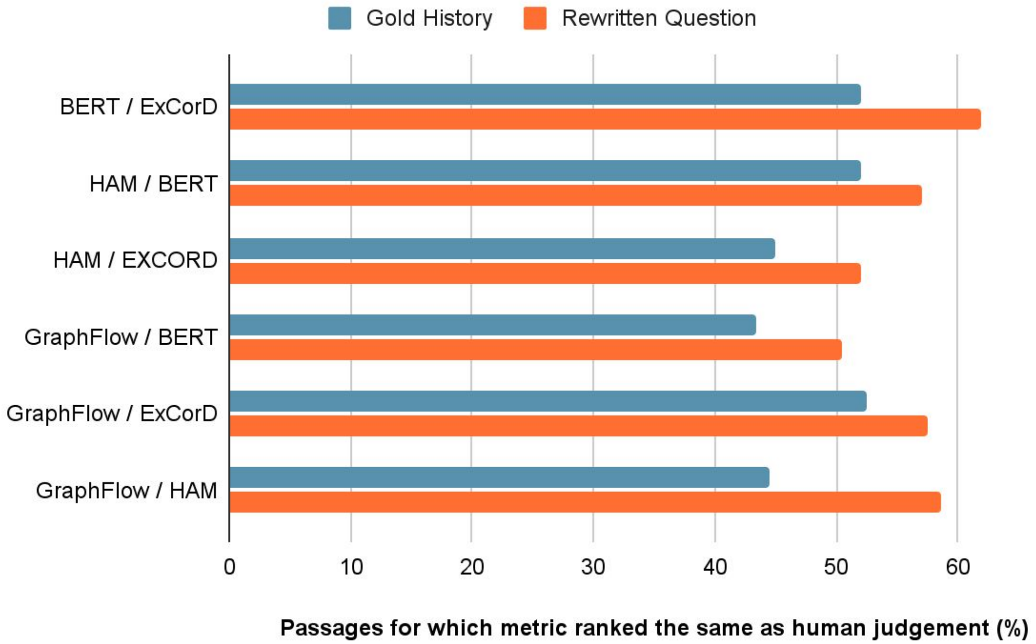


Model-pair ranking evaluation

2. Is rewritten question evaluation closer to human judgement?



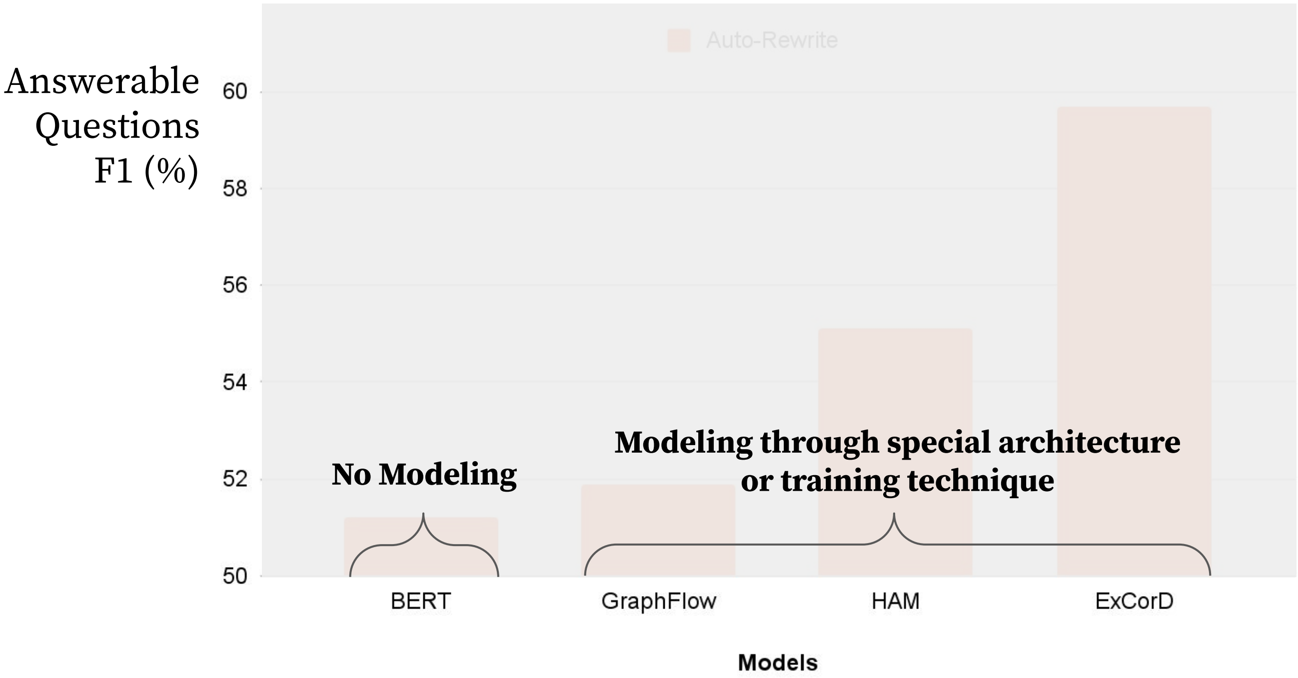
2. Is rewritten question evaluation closer to human judgement?



Rewritten question evaluation ranks the *passage-wise model performance* **more similarly to humans** than gold history evaluation

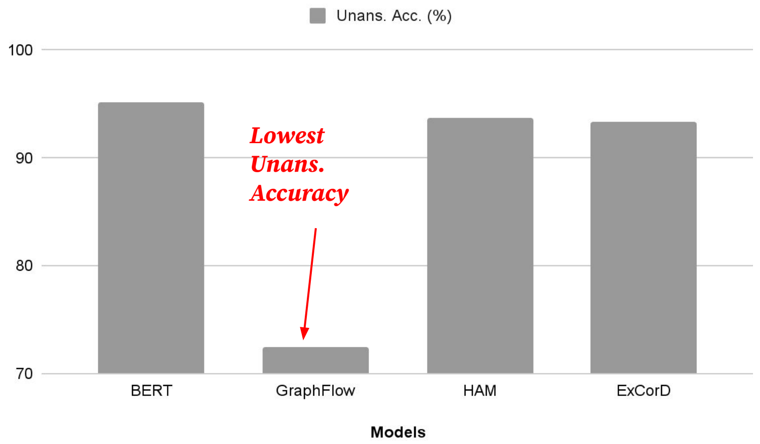
How to make a good conversational question answering model?

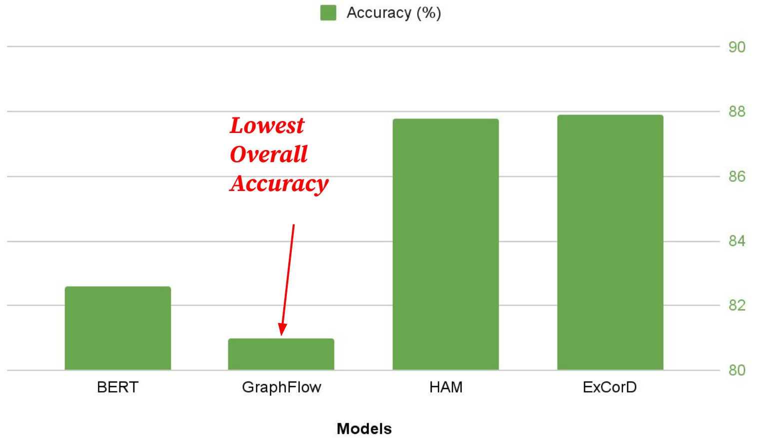
1. Modeling question dependencies on conversational context



Modeling question-history and question-context dependency helps with **span prediction**

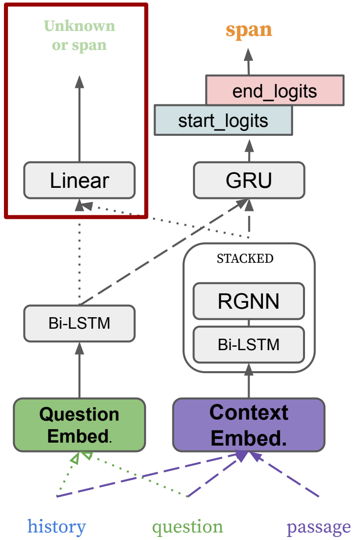
2. Calculate "unanswerable" probability together with span probability





GraphFlow uses a *separate network* for predicting answerability, which is *harder to calibrate*.

GraphFlow (Chen et al. 2019)



Summary

- We conduct the **first large-scale human evaluation** on conversational QA systems.

    - Human-machine conversations have *different question distribution* and *answer distribution* from human-human conversations.
    - Gold history evaluation of current benchmarks *does not agree* with human judgement in human-machine conversations.

Summary

- We propose **a new evaluation protocol** with question rewriting.

    - Simply using the model's prediction in history will result in *invalid questions* because of incoherent history.
    - Rewriting question evaluation *resolves* the invalid questions, and is *closer to human judgment*.

Summary

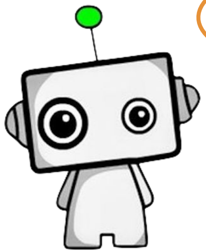
- We provide some insight on **better ConvQA Modeling**.

    - ❖ Modeling question dependencies on conversational context helps with *span prediction*.
    - ❖ Calculating “unanswerable” probability together with span probability helps with *answerability prediction*.

Future Direction

- Training model for the Rewritten Question Evaluation protocol
    - Train model using model's own prediction history (Mandya et al., 2020; Siblini et al., 2021)

Thank you for listening!



Code & Human Evaluation Data

https://github.com/princeton-nlp/EvalConvQA

Email

huihanl@princeton.edu