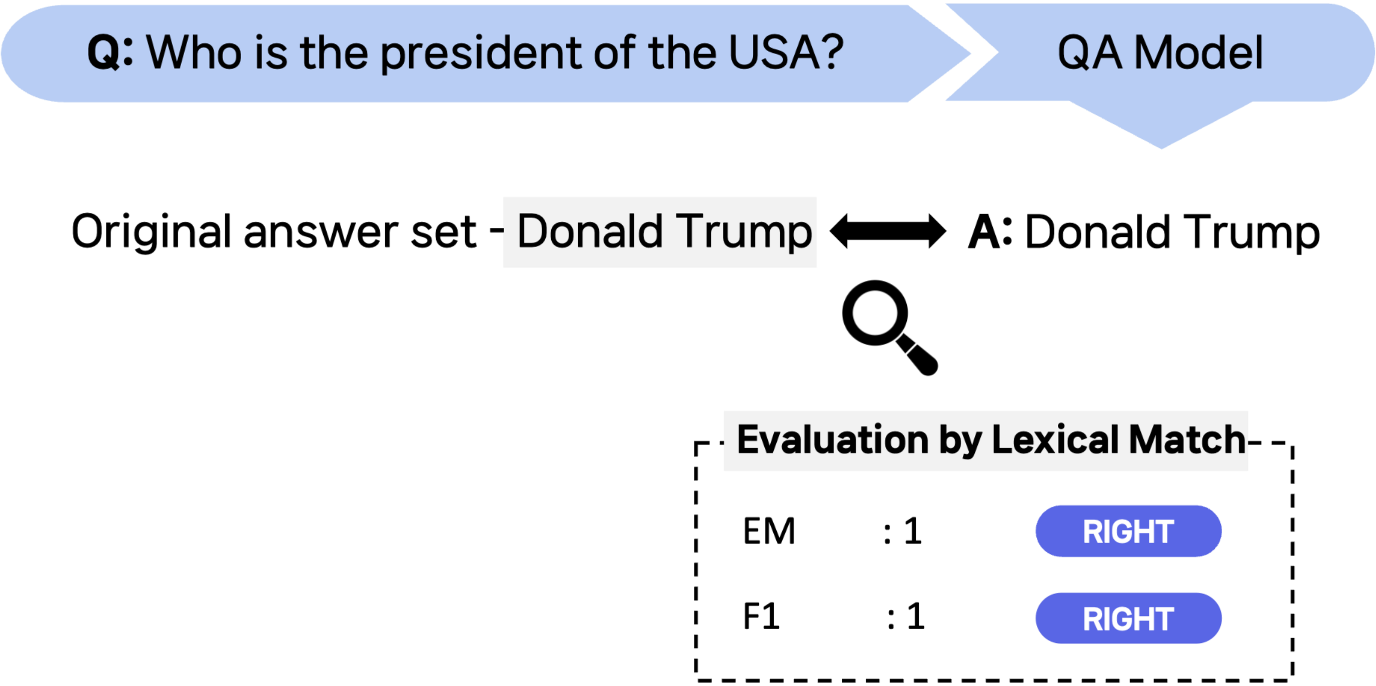
Return of EM: Entity-driven Answer Set Expansion for QA Evaluation

Dongryeol Lee, Minwoo Lee, Kyungmin Min, Joonsuk Park†, Kyomin Jung†

† = corresponding authors

COLING 2025

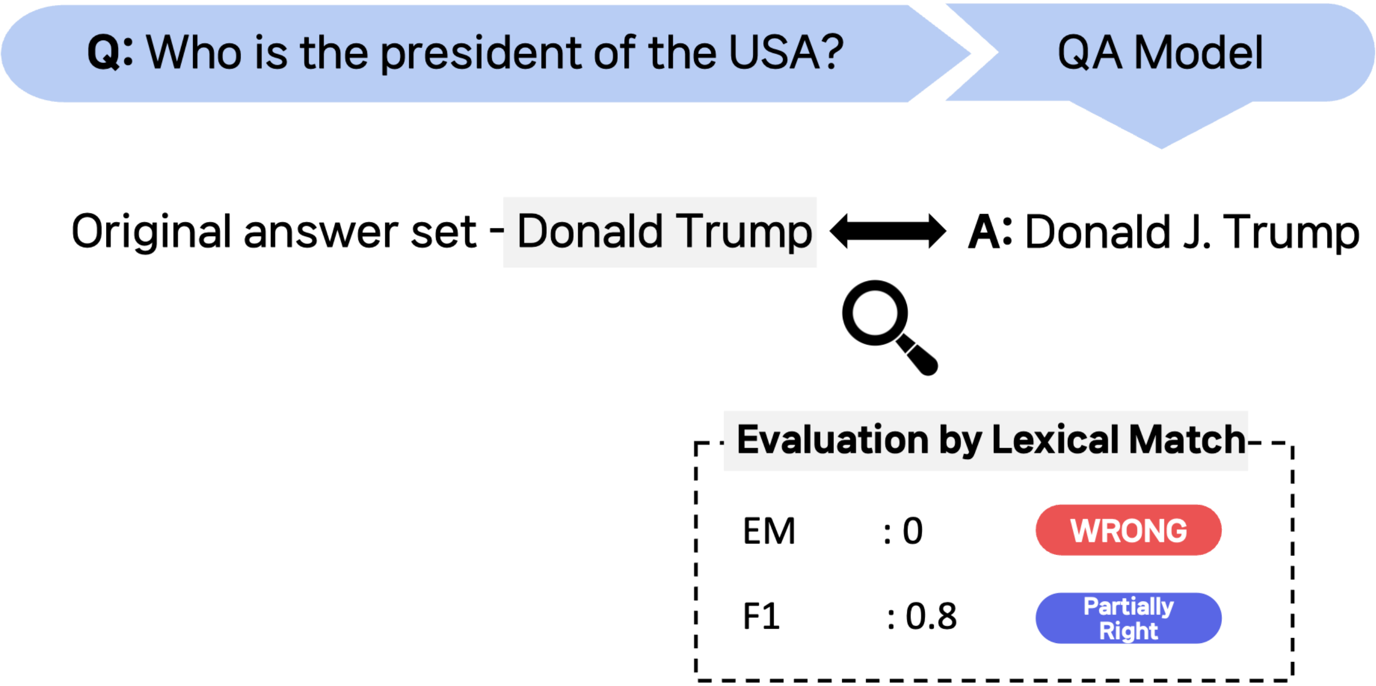
Introduction: QA Evaluation



- QA model outputs are typically evaluated using **lexical match metrics**, such as Exact Match (EM) or F1

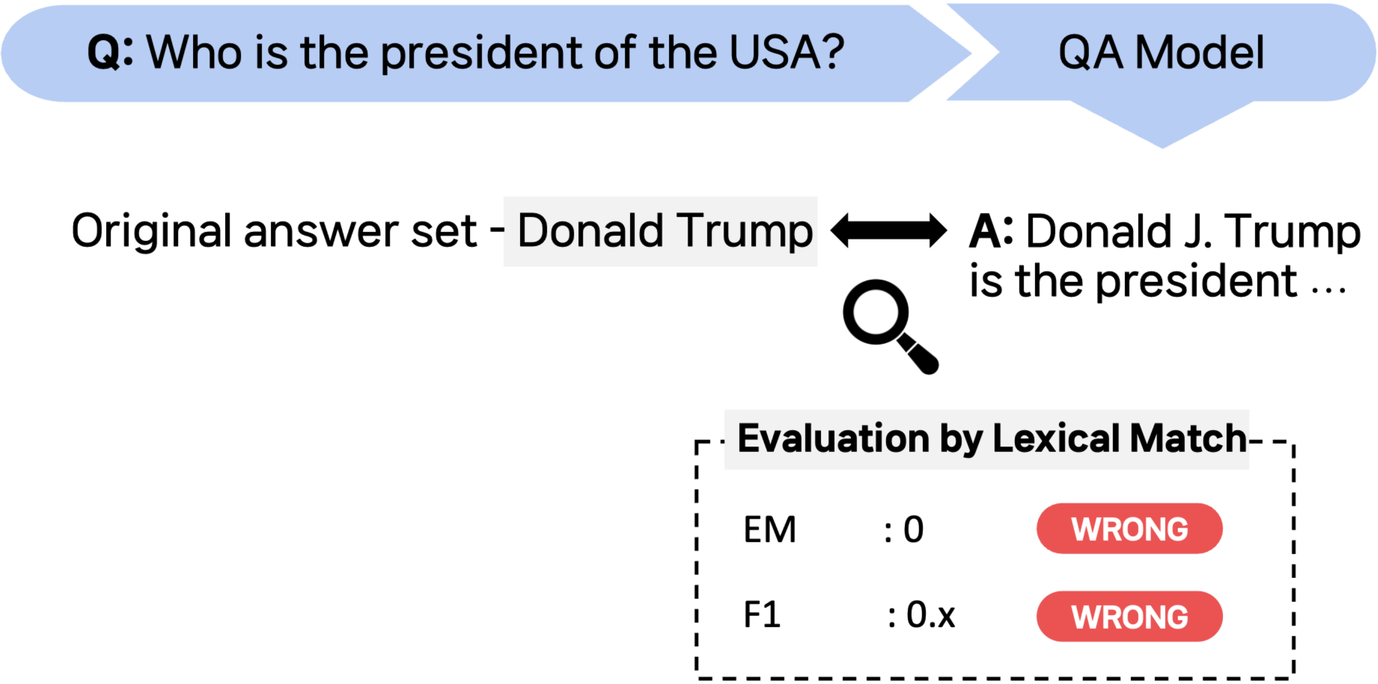
- These metrics compare the model's outputs with the provided answer set

Challenges in QA Evaluation



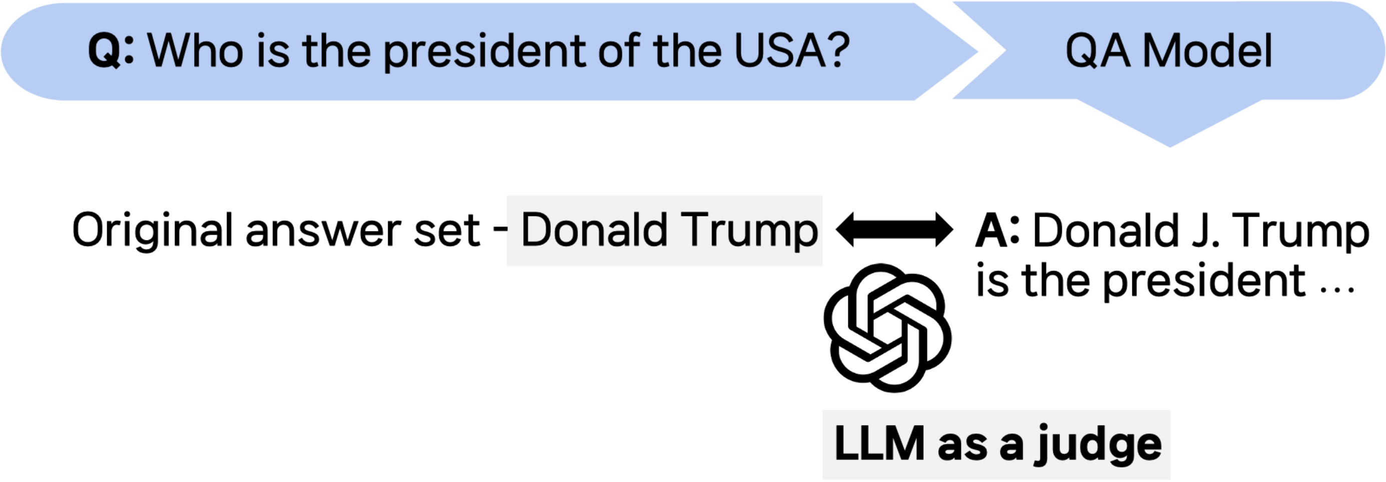
- Existing answer sets usually include only a **single answer**
- Answers can appear in **different surface formats**
  (e.g., Donald Trump vs. Donald J. Trump)

Challenges in QA Evaluation



- Recent studies utilize LLM itself as a QA model, usually resulting in **long-form answers with various surface formats**
- Lexical match metrics are overly strict, leading to False Negative evaluations

Challenges in QA Evaluation



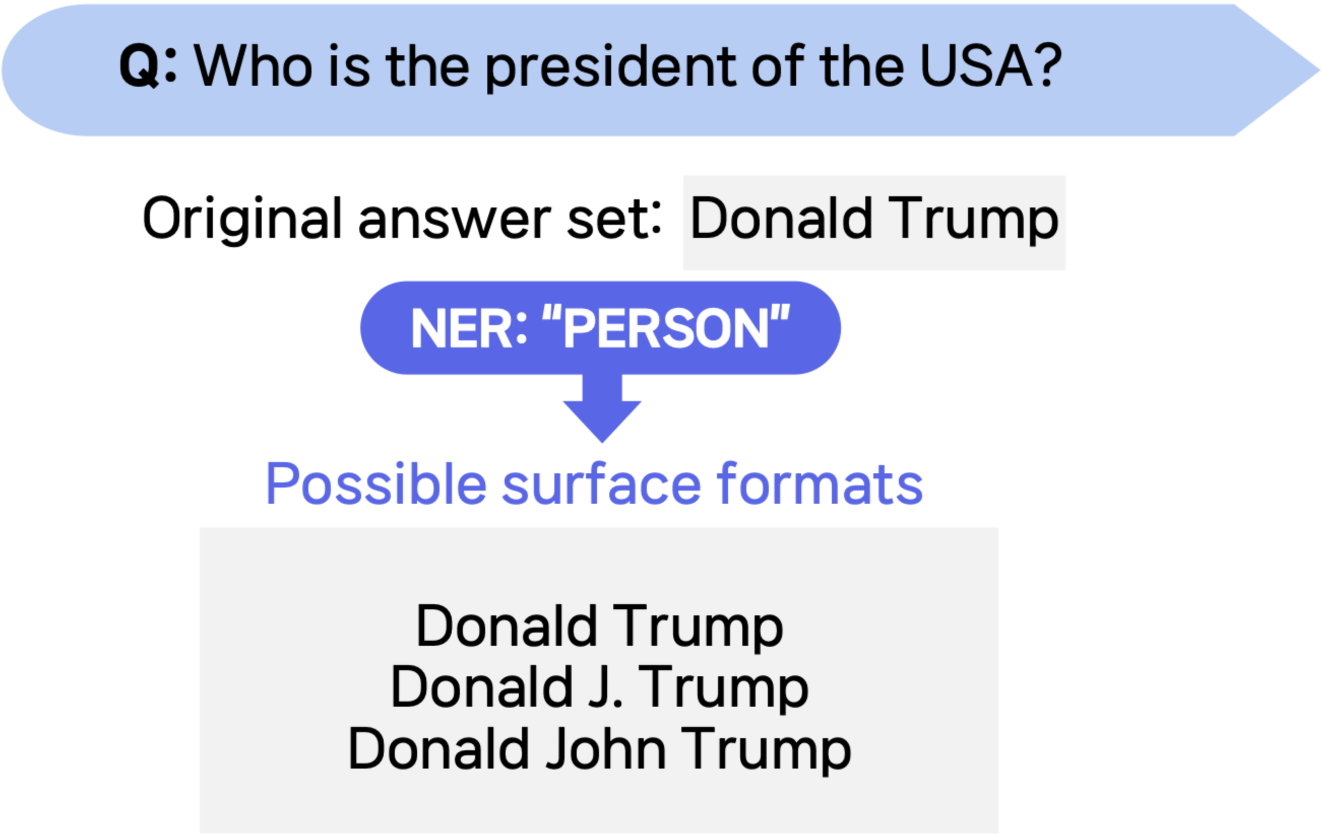
- **LLM as a judge**, directly prompting LLMs to evaluate outputs, has shown reliable performance

- However, it is expensive and suffers from poor interpretability, showing various biases

Challenges in QA Evaluation

Can we build a QA evaluation system that is cost-efficient and reliable?

Motivation: Correlation Between Surface Formats and Entity Types



- **Entity types drive surface form variations**

- For examples, **[PERSON]** entities may appear as abbreviations, last names, or full name

(e.g. Donald John Trump → Trump → Donald J. Trump)

Motivation: Correlation Between Surface Formats and Entity Types

Q: When was Donald Trump born?

Original answer set: June 14, 1946

NER: "DATE"

Possible surface formats

June 14, 1946
June 14th, 1946
14 June, 1946
14 Jun, 1946

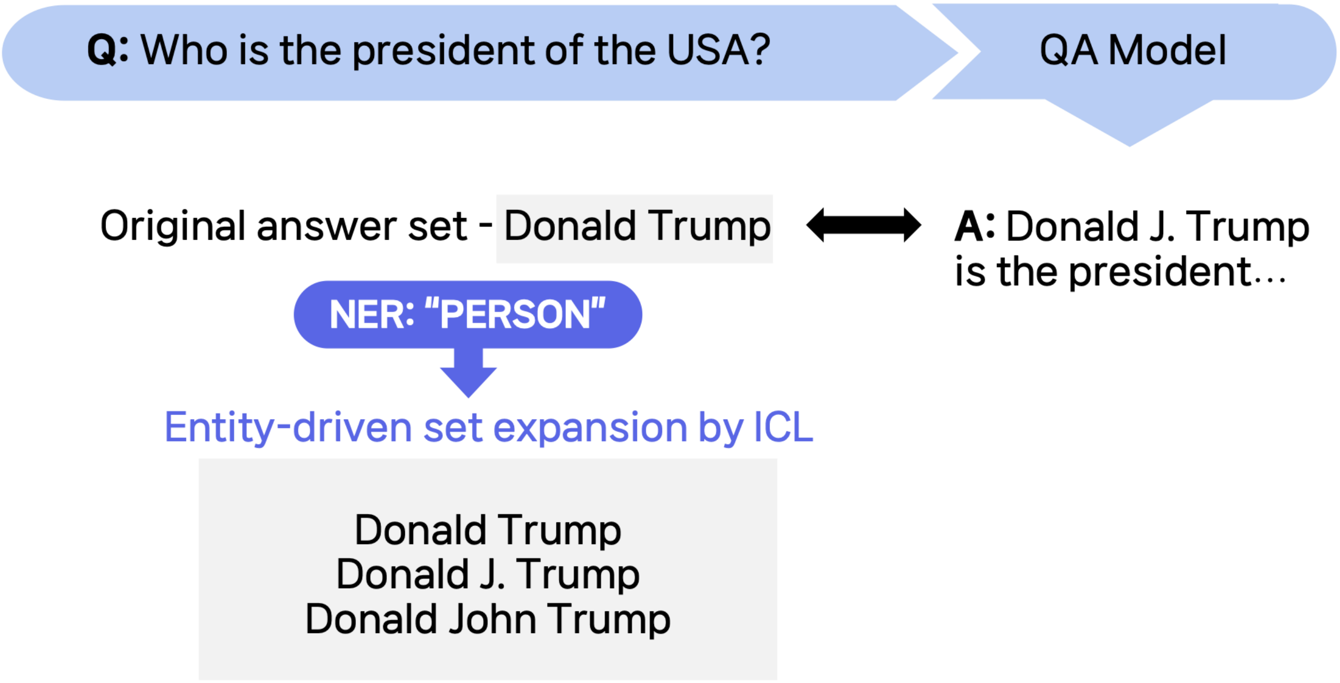
- **Entity types drive surface form variations**
- For examples, **[DATE]** entities may differ in order (e.g., June 14 → 14 June) or abbreviation (e.g., June → Jun).

Method: Soft EM with Entity-Driven Answer Set Expansion

Entity type	Format types	Examples
Numeric - TIME - MONEY - QUANTITY - PERCENT - CARDINAL - DATE - ORDINAL	Numerals	**Q:** How many episodes are in season 2 of the handmades tale **Gold Answer:** 13 **Model Prediction:** The Season 2 of the Handmaid's Tale have thirteen episodes.
	Different Representation (symbols, abbrev., order)	**Q:** When was ye rishta kya kehlata hai started **Gold Answer:** January 12, 2009 **Model Prediction:** The Ye Rishta Kya Kehlata Hai started in 12 Jan., 2009.
	Specificity	**Q:** What's the population of fargo north dakota **Gold Answer:** 120,762 **Model Prediction:** The population of Fargo, North Dakota is about 120,000.
	Unit conversion	**Q:** How long is the movie son of god **Gold Answer:** 138 minutes **Model Prediction:** The movie Son of God is 2 hours and 18 minutes long.
Non-numeric - PERSON - GPE - ORG - Other	Different representation (symbols, abbrev., order)	**Q:** Where was the ncaa football championship game played 2018 **Gold Answer:** Atlanta, Georgia **Model Prediction:** The 2018 NCAA Football Championship Game was played in Atlanta, GA.
	Specificity	**Q:** Who played lionel in all in the family **Gold Answer:** Michael Evans **Model Prediction:** Mike Evans played Lionel Jefferson in All in the Family.
N/A	Contextual Paraphrase	**Q:** The pectoralis minor is located deep to which muscle **Gold Answer:** beneath the pectoralis major **Model Prediction:** under the pectoralis major muscle

- We categorize the **surface format variation** of each **entity type**
- Spacy's NER is used to classify answer set into 19 categories (18 predefined by Spacy + an additional N/A category)

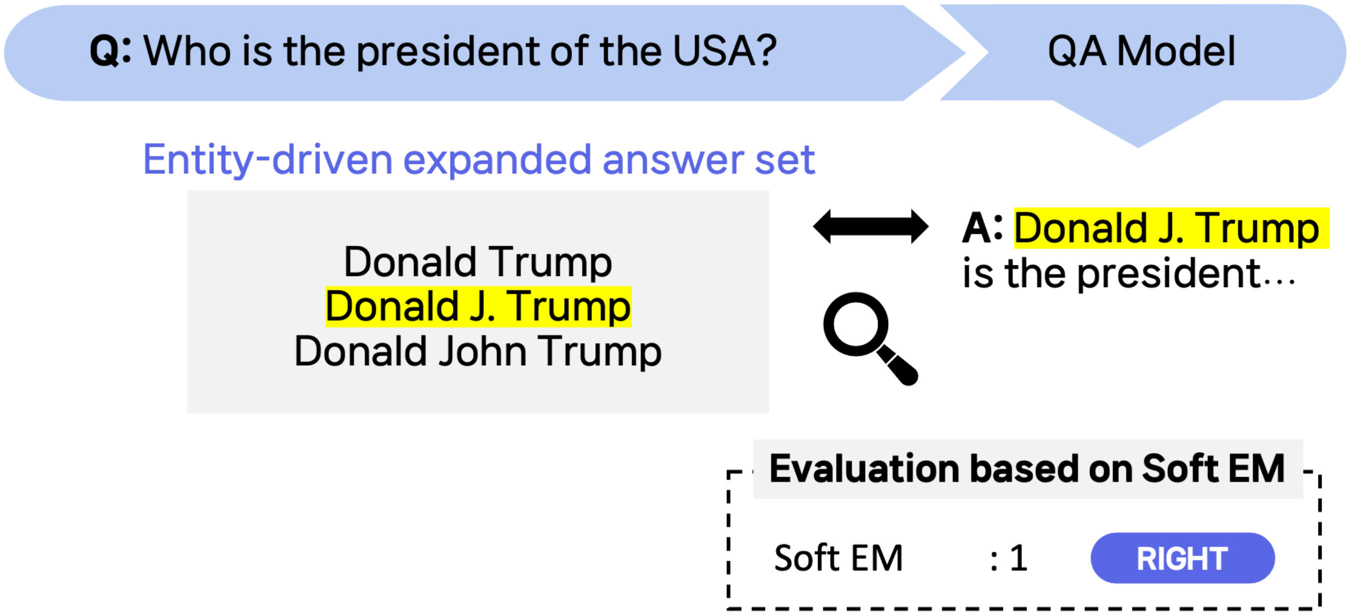
Method: Soft EM with Entity-Driven Answer Set Expansion



Step 1: Entity-Driven Answer Set Expansion

- Manually create **few-shot expanded answer set** for **each entity type**
- Leverage InstructGPT (GPT-3.5-turbo-instruct) with **In-Context Learning (ICL) for expansion**

Method: Soft EM with Entity-Driven Answer Set Expansion



Step 2: Evaluation with Soft EM

- Evaluate QA model outputs using the expanded answer set
- **Soft EM** marks a candidate as correct if it includes any answer from the expanded set

Research Questions

RQ #1: Is our method effective compared to other answer set expansion approaches? (e.g. knowledge-base methods)

RQ #2: Is our method reliable compared to other QA evaluation metrics?

Research Question #1

RQ #1: Is our method effective compared to other answer set expansion approaches? (e.g. knowledge-base methods)

RQ #2: Is our method reliable compared to other QA evaluation metrics?

Experiment Setup

Dataset

- 3,020 instances from **Natural Questions** (Kwiatkowski 2019)
- 1,938 instances from **TriviaQA** (Joshi et al., 2017)
- Responses from **5 QA models** are evaluated – Fusion in Decoder (FiD), GPT 3.5, ChatGPT, GPT4, BingChat
- **Human judgment annotation** from EVOUNA (Wang et al., 2023) used as a reference

Evaluation

- **Accuracy against human judgment**

Experiment Setup

Baselines

Answer set expansion method using knowledge base

- **Freebase**: Expansion using Freebase knowledge base (Bollacker et al., 2008)
- **Wiki**: Expansion using Wikipedia knowledge base

Answer set expansion method using InstructGPT (GPT-3.5-turbo-instruct)

- **Inst-zero**: Expansion with **zero-shot** example
- **Inst-random**: Expansion with **random few-shot examples** regardless of entity type
- **Inst-entity (Ours):** Expansion with **entity type-driven few-shot examples**

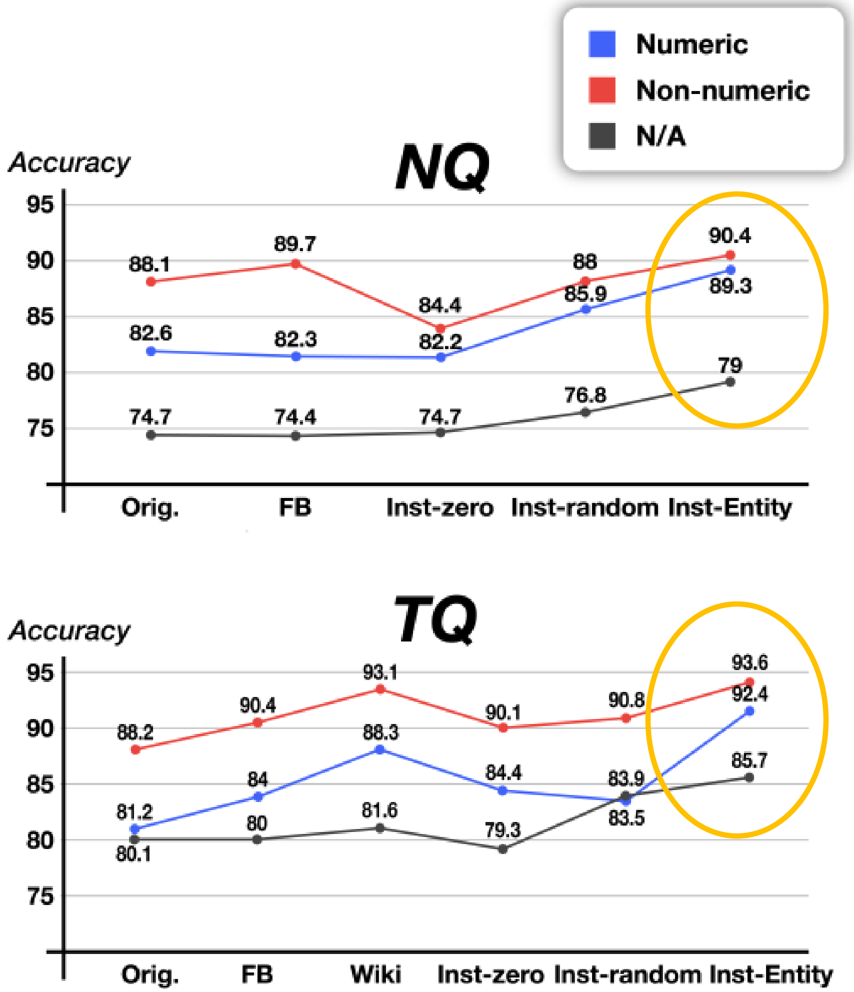
Result

Natural Questions						
Evaluation Method	FiD	GPT3.5	ChatGPT3.5	ChatGPT4	BingChat	Avg.
Soft EM with Answer Set expansion						
Freebase	<u>89.8</u>	<u>85.5</u>	81.7	83.9	83.9	85.0
Inst-zero	85.4	79.4	79.3	82.0	83.8	82.0
Inst-random	88.1	83.8	<u>82.2</u>	<u>86.0</u>	<u>86.6</u>	<u>85.3</u>
Inst-entity (Ours)	**91.0**	**86.8**	**85.7**	**88.2**	**87.7**	**87.9**
TriviaQA						
Evaluation Method	FiD	GPT3.5	ChatGPT3.5	ChatGPT4	BingChat	Avg.
Soft EM with Answer Set Expansion						
Freebase	90.6	89.4	89.0	88.4	87.0	88.9
Wiki	<u>92.0</u>	<u>92.2</u>	<u>92.3</u>	<u>91.2</u>	90.1	<u>91.6</u>
Inst-zero	88.1	86.1	88.6	89.7	90.3	88.6
Inst-random	89.3	87.4	89.4	90.3	<u>91.2</u>	89.5
Inst-entity (Ours)	**92.6**	**92.5**	**93.3**	**93.0**	**92.4**	**92.8**

Table 13: Reliability (accuracy w.r.t. human verdicts) of evaluation methods tested on the output of five QA models. **Bold** indicates the highest score, and <u>underline</u> indicates the second highest score.

- Our method (Inst-Entity) demonstrates the **highest reliability across 5 QA models and 2 datasets**

Result



- We separately report the accuracy based on entity types (Numeric, Non-numeric, N/A)
- Our method (Inst-Entity) demonstrates the **highest reliability regardless of entity types**
- Our method (Inst-Entity) is especially effective in **numeric entity type**

Research Questions

RQ #1: Is our method effective compared to other answer set expansion approaches? (e.g. knowledge-base methods) – Yes!

RQ #2: Is our method reliable compared to other QA evaluation metrics?

Research Question #2

RQ #1: Is our method effective compared to other answer set expansion approaches? (e.g. knowledge-base methods) – Yes!

RQ #2: Is our method reliable compared to other QA evaluation metrics?

Experiment Setup

Dataset

- 3,020 instances from **Natural Questions** (Kwiatkowski 2019)
- 1,938 instances from **TriviaQA** (Joshi et al., 2017)
- Responses from **5 QA models** are evaluated – Fusion in Decoder (FiD), GPT 3.5, ChatGPT, GPT4, BingChat
- **Human judgment annotation** from EVOUNA (Wang et al., 2023) used as a reference

Evaluation

- **Accuracy against human judgment**

Experiment Setup

Baselines

Lexical Matching-based with original answer set

- **Hard Exact Match (Hard EM)**: Candidate is correct if it exactly matches the gold answer
- **Soft Exact Match (Soft EM)**: Candidate is correct if it contains the gold answer
- **F1**: Measure the token overlap between the reference answer and prediction

Model-based

- **BEM** (Bulian et al., 2022): Pre-trained BERT model for answer equivalence
- **Insteval**: Directly prompt InstructGPT (GPT-3.5-turbo-instruct) to evaluate response

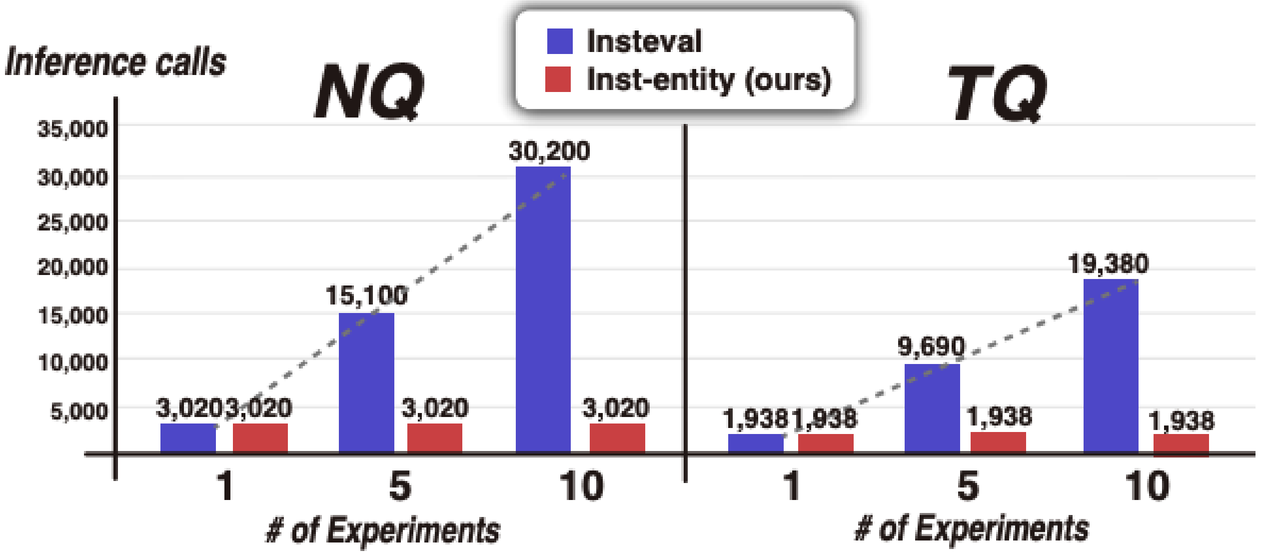
Result

Evaluation Method	FiD	GPT3.5	ChatGPT3.5	ChatGPT4	BingChat	Avg.
Natural Questions						
Model-based						
BEM	93.5	73.6	77.9	82.1	84.0	82.2
Insteval	91.8	85.2	**86.2**	**89.2**	**88.0**	**88.1**
Lexical Matching-based						
Soft EM	89.7	84.9	80.5	82.9	82.7	84.1
Hard EM	86.9	37.3	28.5	21.2	20.1	38.8
F1	**94.4**	40.2	31.5	23.4	20.5	42.0
Soft EM with Answer Set expansion						
Inst-entity (Ours)	91.0	**86.8**	85.7	88.2	87.7	87.9
TriviaQA						
Evaluation Method	FiD	GPT3.5	ChatGPT3.5	ChatGPT4	BingChat	Avg.
Model-based						
BEM	93.8	89.2	88.3	92.2	90.3	90.8
Insteval	**96.4**	**94.2**	**94.9**	**96.0**	**95.1**	**95.3**
Lexical Matching-based						
Soft EM	88.0	87.5	87.3	86.2	84.8	86.8
Hard EM	85.3	40.8	22.0	13.2	10.4	34.3
F1	93.0	50.9	26.3	20.6	10.6	40.3
Soft EM with Answer Set Expansion						
Inst-entity (Ours)	92.6	92.5	93.3	93.0	92.4	92.8

Table 14: Reliability (accuracy w.r.t. human verdicts) of evaluation methods tested on the output of five QA models. **Bold** indicates the highest score, and underline indicates the second highest score.

- Our method (Inst-Entity) achieves the **second-highest reliability** across 5 QA models and 2 datasets
- Insteval (LLM-as-a-judge) demonstrates the **highest reliability**

Result: Comparison Against Insteval



- **Insteval** **requires inference calls that scale linearly** with the number of evaluation
- In contrast, **our method requires only a single inference call** for evaluation while **maintaining comparative reliability**

Result: Comparison Against Insteval

Type	Examples
Nonsensical Evaluation (84%)	**Question:** who has played in the most masters tournaments **Answer:** [Gary Player] **Model prediction:** Jack Nicklaus has played in the most Masters Tournaments, with a total of 44 appearances. **Human judgement** on **Model prediction:** Incorrect **Insteval judgement** on **Model prediction:** Correct

- Insteval **suffers from poor interpretability**, with 84% of its errors **lacking understandable reasons**
- In contrast, **our method offers clearer justification** for evaluation outcomes

Research Questions

RQ #1: Is our method effective compared to other answer set expansion approaches? (e.g. knowledge-base methods) – Yes!

RQ #2: Is our method reliable compared to other QA evaluation metrics?
– Yes! Additionally, our method offers significant advantages in cost efficiency and interpretability

Takeaways

- Proposed to **expand QA answer sets based on entity type** and evaluate with **Soft EM**
- Achieved **high correlation with human judgments**, with **benefits in cost and interpretability**
- **Open-sourced the expanded answer set** for the community

Datasets

Paper

Contact: drl123@snu.ac.kr

Thank you