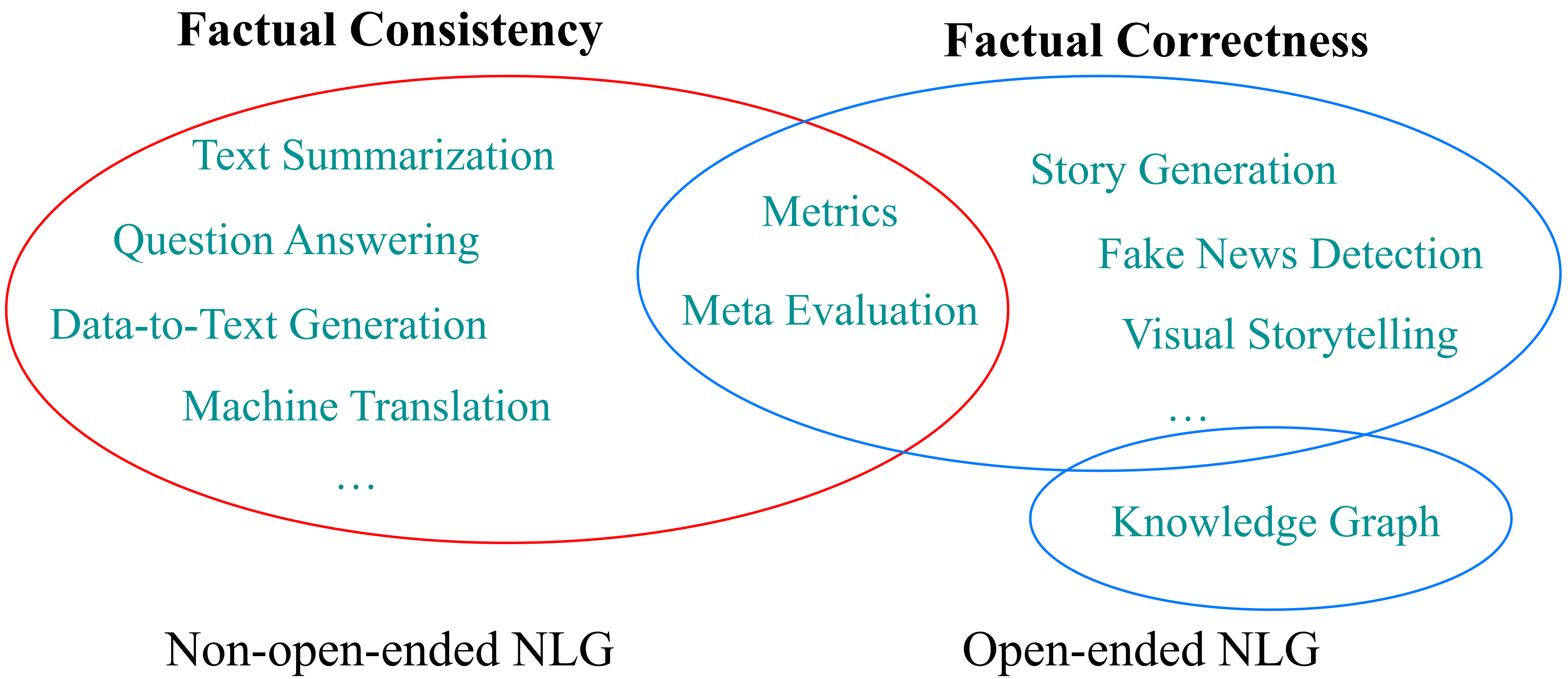
Does the Generator Mind its Contexts? An Analysis of Generative Model Faithfulness under Context Transfer

Xinshuo Hu♠ , *Baotian Hu*♠ , *Dongfang Li*♠ , *Xiaoguang Li*♥ , *Lifeng Shang*♥

♠ *Harbin Institute of Technology, Shenzhen,* ♥ *Huawei Noah's Ark Lab*

Background

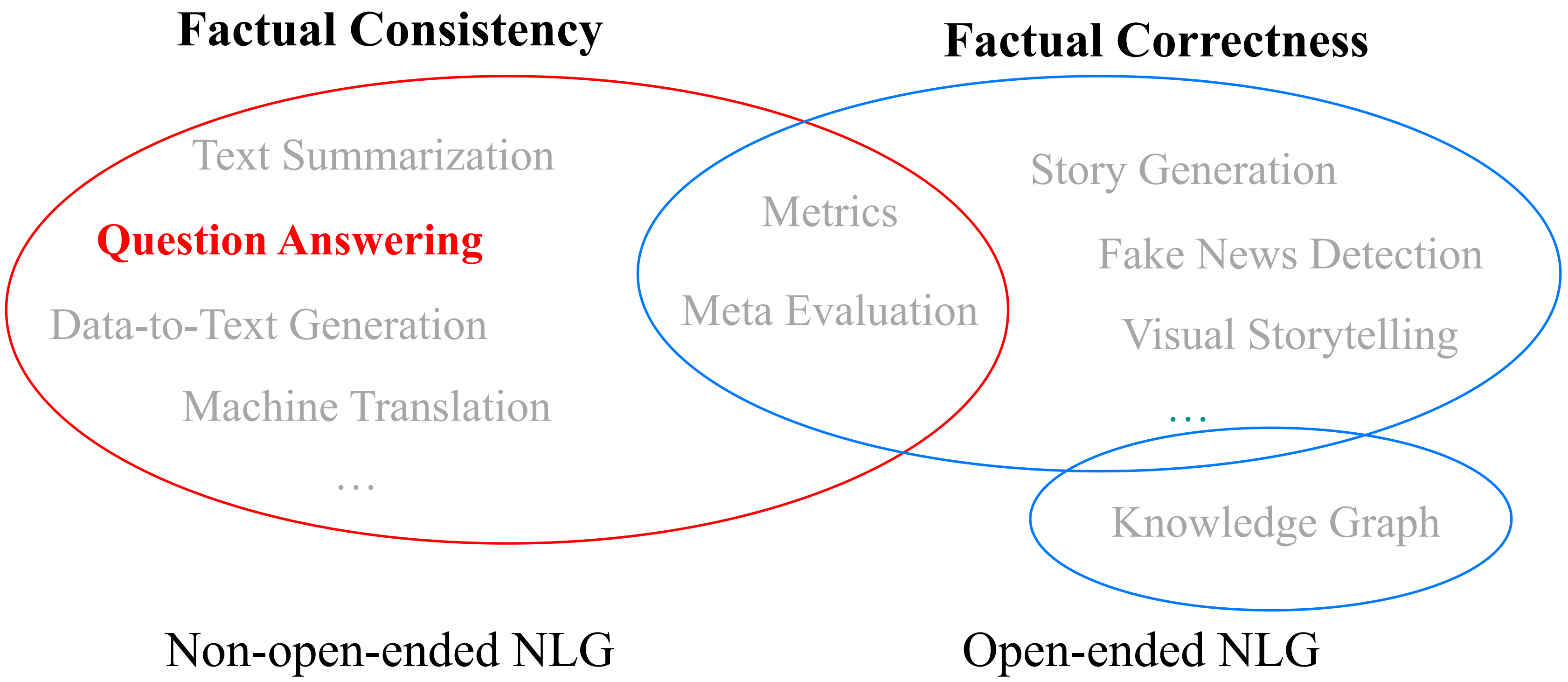
- Mountain to Climb for Generative Language Models:
  - The research framework on the faithfulness problem



Li W, Wu W, Chen M, et al. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods[J]. arXiv preprint arXiv:2203.05227, 2022.

Background

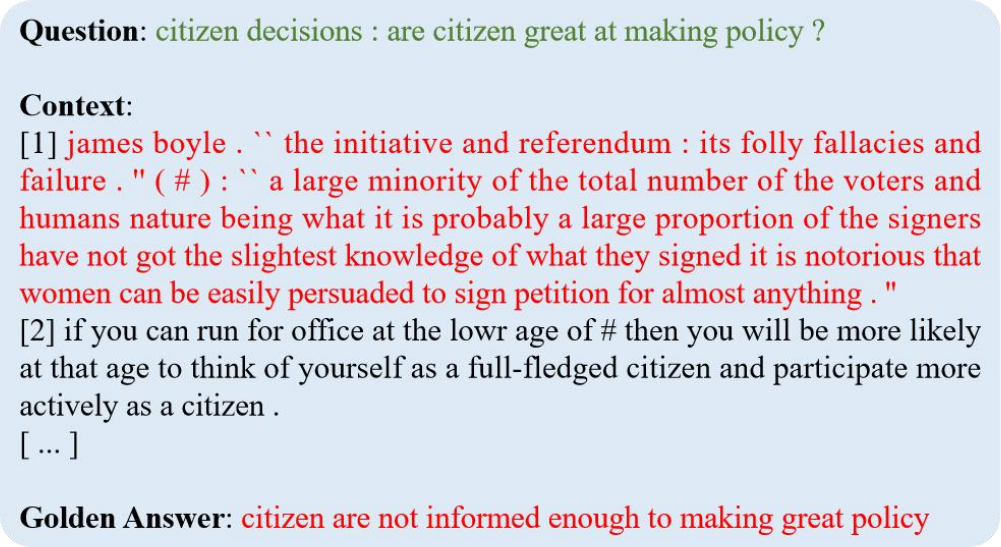
- Mountain to Climb for Generative Language Models:
  - The research framework on the faithfulness problem



Li W, Wu W, Chen M, et al. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods[J]. arXiv preprint arXiv:2203.05227, 2022.

Background

- Context Transfer in Question Answering:
  - Learning from the past, testing on the present
  - Training on old contextual documents, while testing on new ones (with the same question)

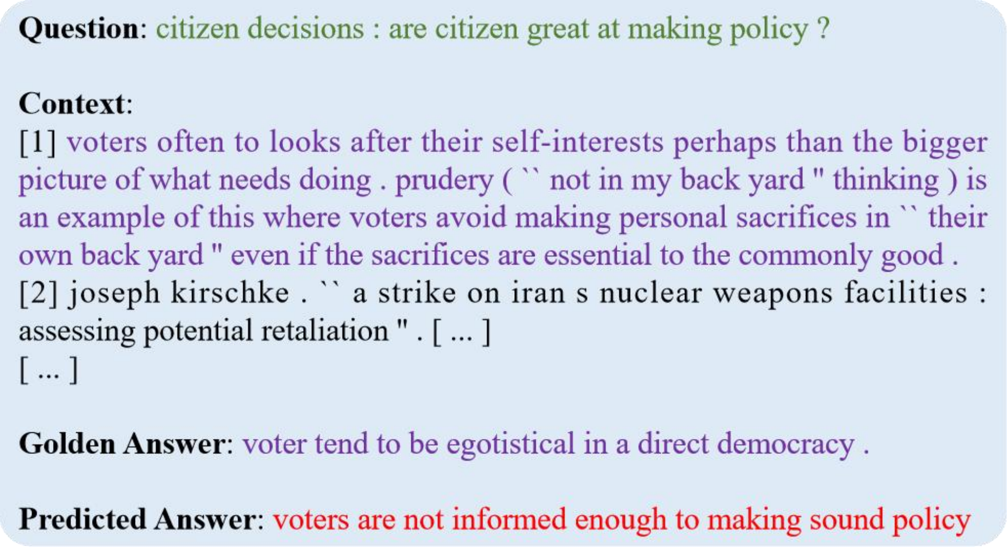


Question: citizen decisions : are citizen great at making policy ?

Context:
[1] james boyle . `` the initiative and referendum : its folly fallacies and failure . " ( # ) : `` a large minority of the total number of the voters and humans nature being what it is probably a large proportion of the signers have not got the slightest knowledge of what they signed it is notorious that women can be easily persuaded to sign petition for almost anything . "
[2] if you can run for office at the lowr age of # then you will be more likely at that age to think of yourself as a full-fledged citizen and participate more actively as a citizen .
[...]

Golden Answer: citizen are not informed enough to making great policy

Training



Question: citizen decisions : are citizen great at making policy ?

Context:
[1] voters often to looks after their self-interests perhaps than the bigger picture of what needs doing . prudery (`` not in my back yard " thinking ) is an example of this where voters avoid making personal sacrifices in `` their own back yard " even if the sacrifices are essential to the commonly good .
[2] joseph kirschke . `` a strike on iran s nuclear weapons facilities : assessing potential retaliation " . [...]
[...]

Golden Answer: voter tend to be egotistical in a direct democracy .

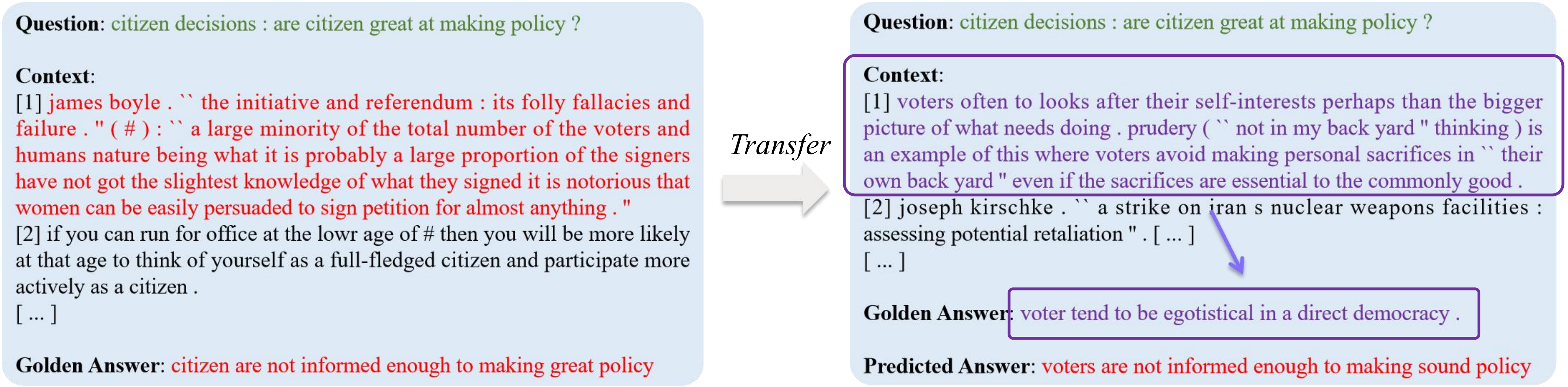
Predicted Answer: voters are not informed enough to making sound policy

Testing

Background

- Context Transfer in Question Answering:
  - Memory hallucination:

  Disregard the transferred contextual knowledge and generate an out-of-date answer in training data, when answering the same question



Question: citizen decisions : are citizen great at making policy ?

Context:
[1] james boyle . `` the initiative and referendum : its folly fallacies and failure . " ( # ) : `` a large minority of the total number of the voters and humans nature being what it is probably a large proportion of the signers have not got the slightest knowledge of what they signed it is notorious that women can be easily persuaded to sign petition for almost anything . "
[2] if you can run for office at the lowr age of # then you will be more likely at that age to think of yourself as a full-fledged citizen and participate more actively as a citizen .
[...]

Golden Answer: citizen are not informed enough to making great policy

Training

Transfer

Question: citizen decisions : are citizen great at making policy ?

Context:
[1] voters often to looks after their self-interests perhaps than the bigger picture of what needs doing . prudery (`` not in my back yard " thinking ) is an example of this where voters avoid making personal sacrifices in `` their own back yard " even if the sacrifices are essential to the commonly good .
[2] joseph kirschke . `` a strike on iran s nuclear weapons facilities : assessing potential retaliation " . [...]
[...]

Golden Answer: voter tend to be egotistical in a direct democracy .

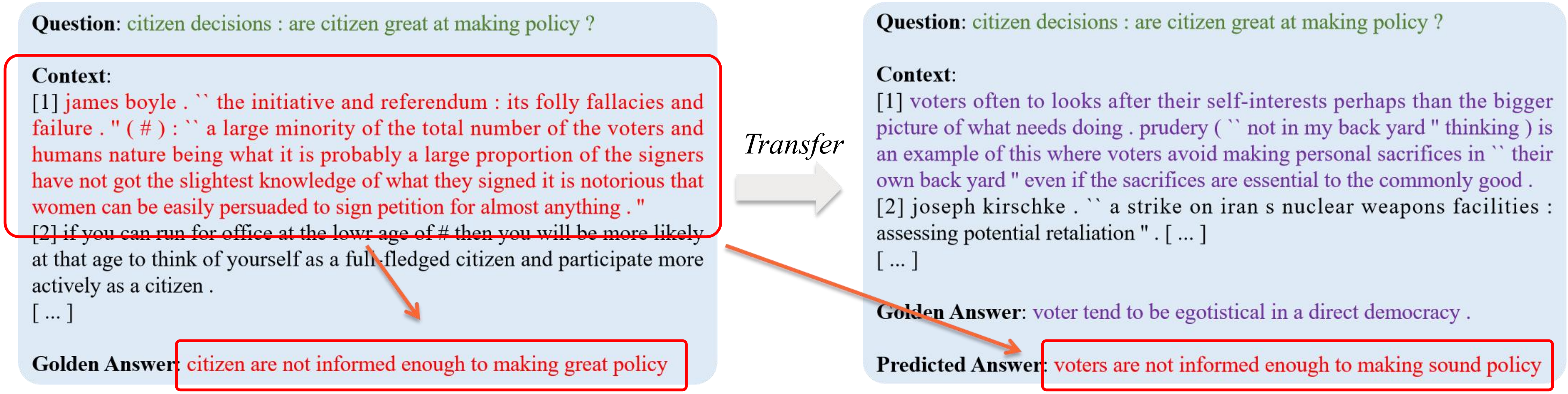
Predicted Answer: voters are not informed enough to making sound policy

Testing

Background

- Context Transfer in Question Answering:
  - Memory hallucination:

  Disregard the transferred contextual knowledge and generate an out-of-date answer in training data, when answering the same question



Question: citizen decisions : are citizen great at making policy ?

Context:
[1] james boyle . `` the initiative and referendum : its folly fallacies and failure . " ( # ) : `` a large minority of the total number of the voters and humans nature being what it is probably a large proportion of the signers have not got the slightest knowledge of what they signed it is notorious that women can be easily persuaded to sign petition for almost anything . "
[2] if you can run for office at the lowr age of # then you will be more likely at that age to think of yourself as a full-fledged citizen and participate more actively as a citizen .
[...]

Golden Answer: citizen are not informed enough to making great policy

Training

Question: citizen decisions : are citizen great at making policy ?

Context:
[1] voters often to looks after their self-interests perhaps than the bigger picture of what needs doing . prudery (`` not in my back yard " thinking ) is an example of this where voters avoid making personal sacrifices in `` their own back yard " even if the sacrifices are essential to the commonly good .
[2] joseph kirschke . `` a strike on iran s nuclear weapons facilities : assessing potential retaliation " . [...]
[...]

Golden Answer: voter tend to be egotistical in a direct democracy .

Predicted Answer: voters are not informed enough to making sound policy

Testing

Research Questions

1

RQ1

To what extent does the generative model exhibit faithfulness under context transfer?

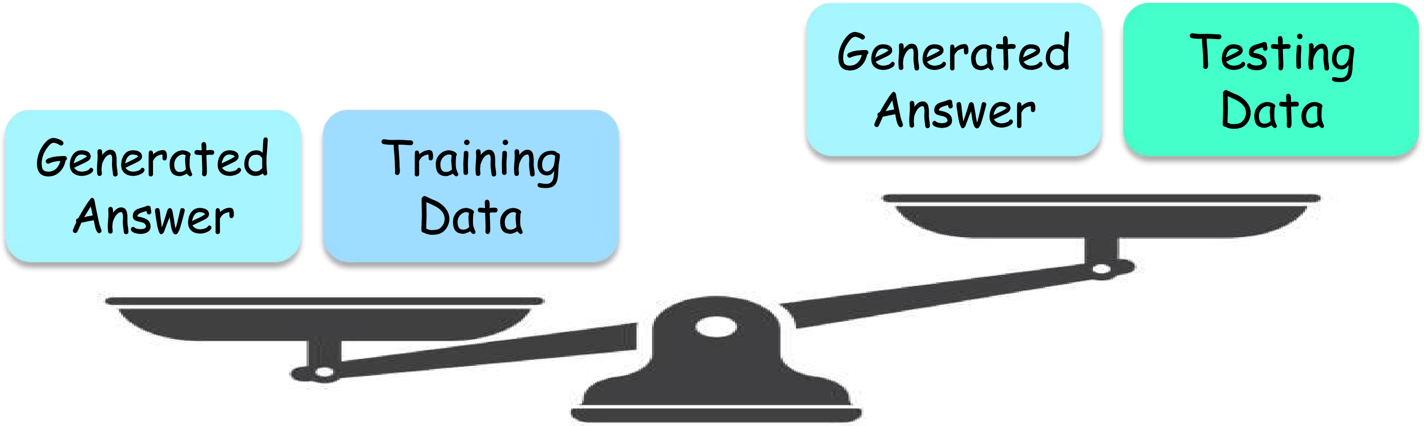
2

RQ2

What are the underlying reasons for the occurrence of memory hallucination?

Methodology

- How to measure such problem



Generated Answer

Training Data

Generated Answer

Testing Data

Methodology

- How to measure such problem
  - **M**argin grounding **F**ailure of context transfer:

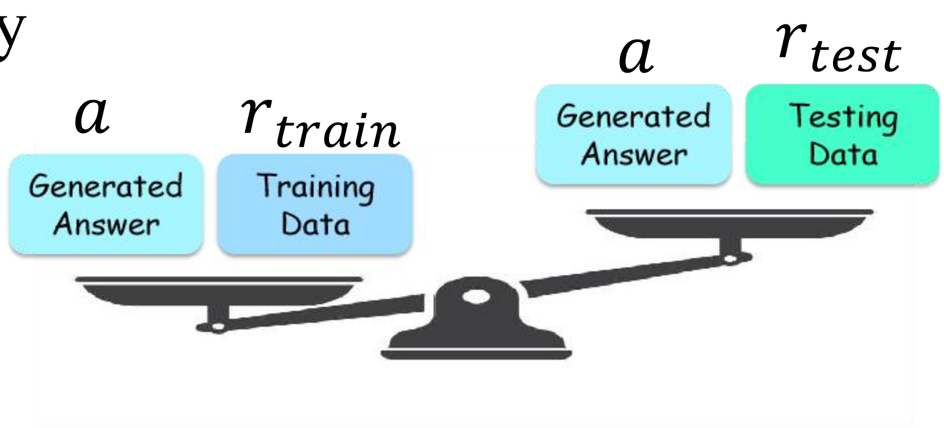
$$MF(\Phi) = \begin{cases} 1, & \Phi(a, r_{train}) > m \times \Phi(a, r_{test}) \\ 0, & \Phi(a, r_{train}) \leq m \times \Phi(a, r_{test}) \end{cases}$$

a: generated answer

r_{train}: reference in training data (answer or context)

Φ: any basic metric to measure similarity

m: margin

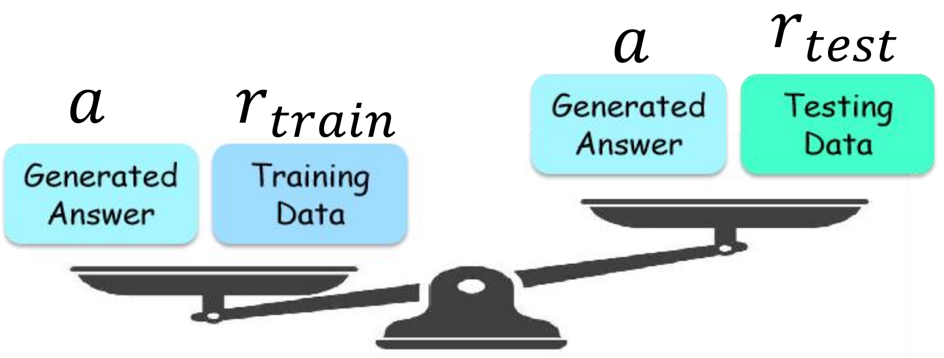


Methodology

- How to measure such problem
  - Specifically,

$$MF(\text{BertScore}) = \begin{cases} 1, & \text{BertScore}(a, a_{train}) \; > \; 1.25 \times \text{BertScore}(a, a_{test}) \\ 0, & \text{BertScore}(a, a_{train}) \; \leq \; 1.25 \times \text{BertScore}(a, a_{test}) \end{cases}$$

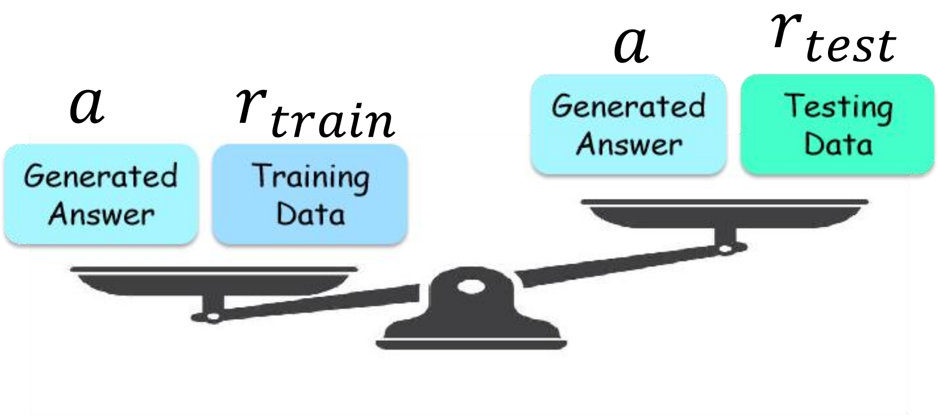
In this work we use BertScore to measure the similarity between generated answer and reference answer (from training or testing)



Methodology

- How to measure such problem
  - **M**argin **F**ailure **R**ate is defined as the percentage of grounding failure:

$$MFR(\text{BertScore}) = \frac{1}{N}\sum_{i=1}^{N} MF_i(\text{BertScore})$$



Experimental Settings

1

Evaluation Dataset

having examples where a question is paired with several different context and answer:

- □ *Debatepedia*

2

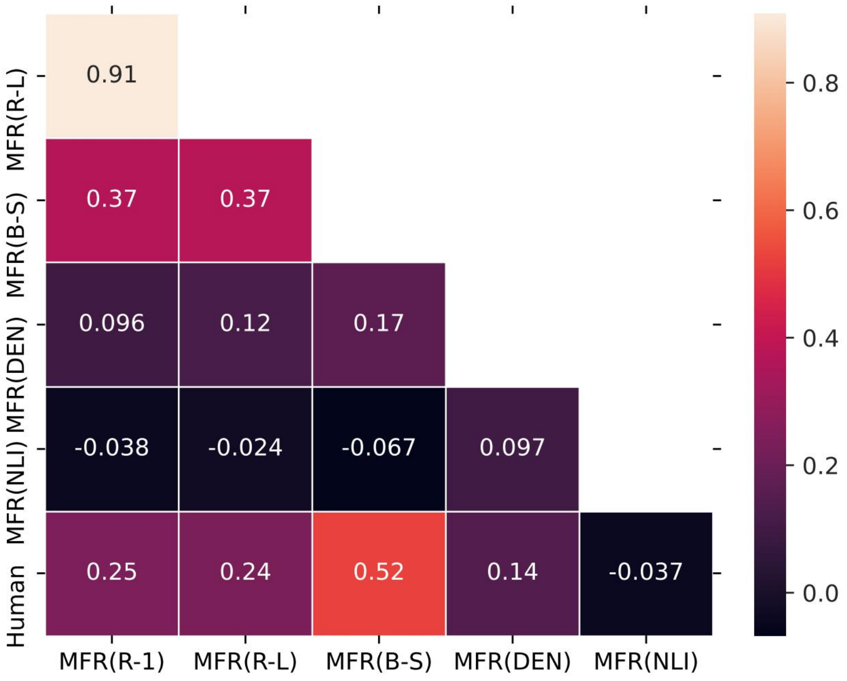
Evaluation Models

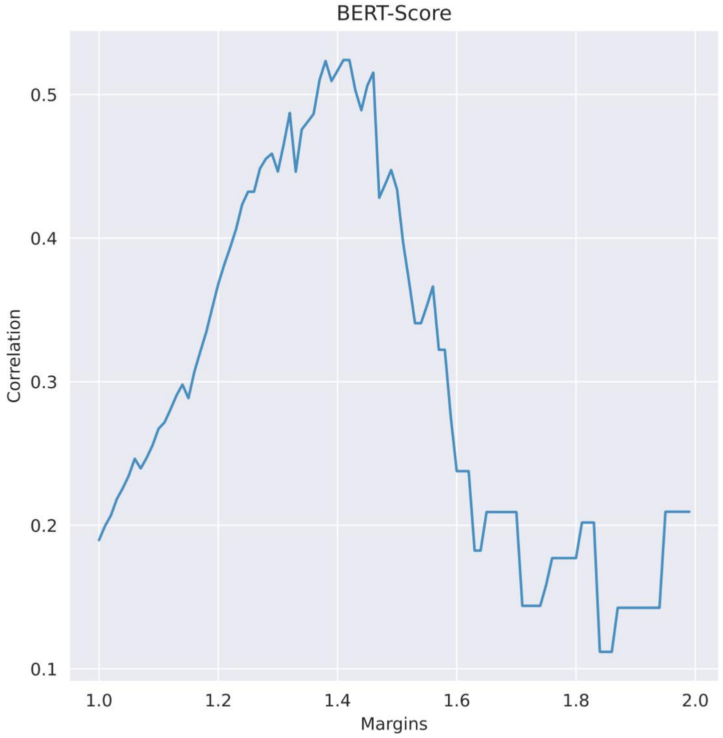
Generative Models in QA:

- □ *T5*
- □ *BART*
- □ *FiD(T5)*
- □ *FiD(BART)*

Experiments

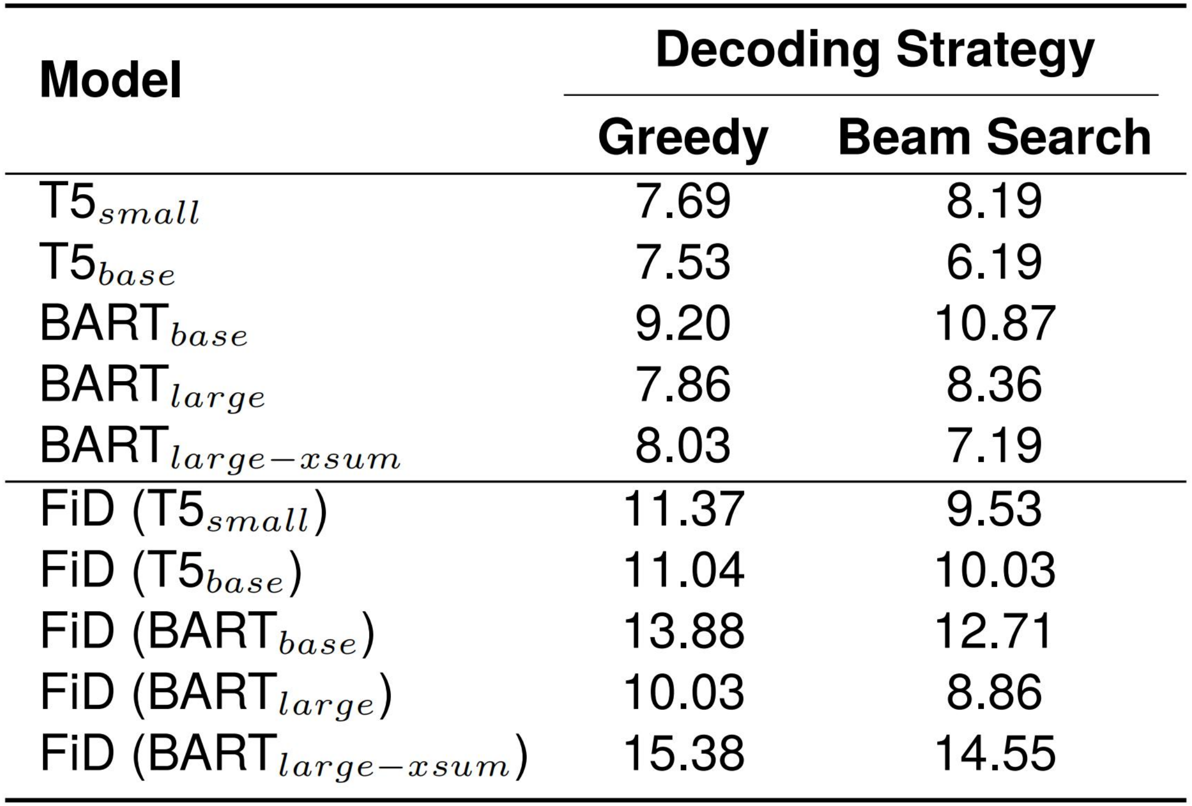
- Meta Evaluation of *MFR* on annotated dev set
  - BertScore (B-S) has the best Pearson Correlation with human labels
  - Setting $m = 1.25$ gets a great correlation





Experiments

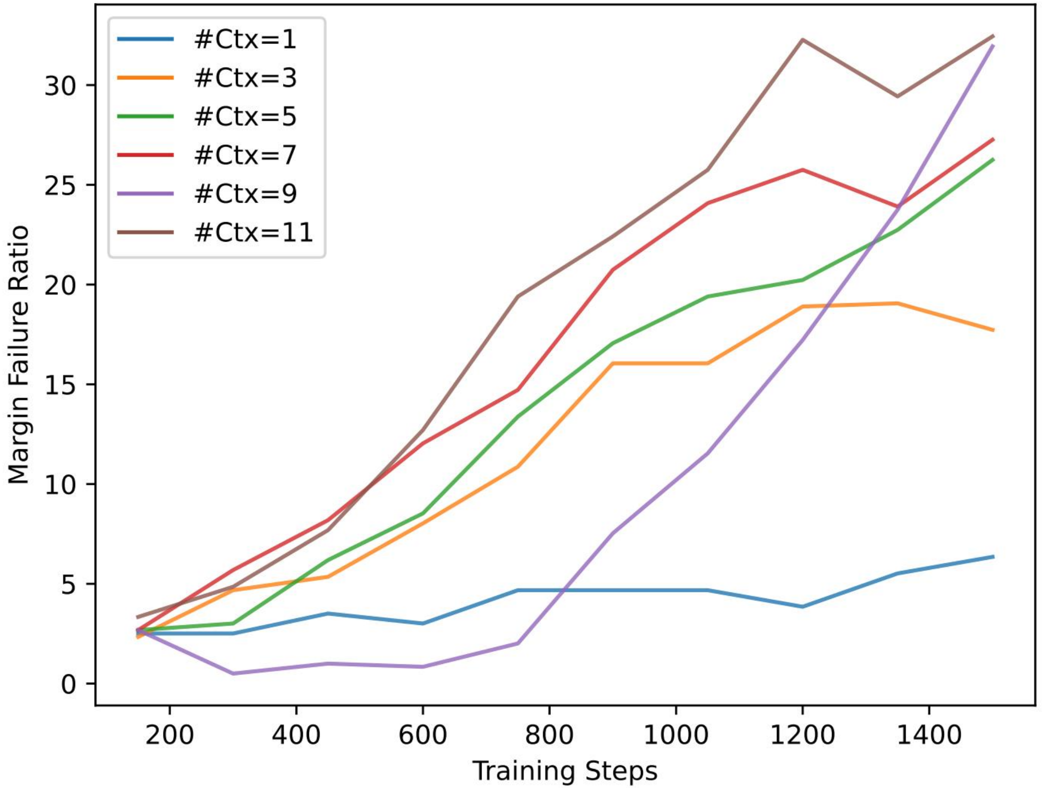
- RQ1: All models have memory hallucination under context transfer



Model	Decoding Strategy	
	Greedy	Beam Search
T5$_{small}$	7.69	8.19
T5$_{base}$	7.53	6.19
BART$_{base}$	9.20	10.87
BART$_{large}$	7.86	8.36
BART$_{large-xsum}$	8.03	7.19
FiD (T5$_{small}$)	11.37	9.53
FiD (T5$_{base}$)	11.04	10.03
FiD (BART$_{base}$)	13.88	12.71
FiD (BART$_{large}$)	10.03	8.86
FiD (BART$_{large-xsum}$)	15.38	14.55

Experiments

- RQ2: Impact of Contextual Knowledge Scale
  Memory hallucination increases proportionally with the expansion of the context scale

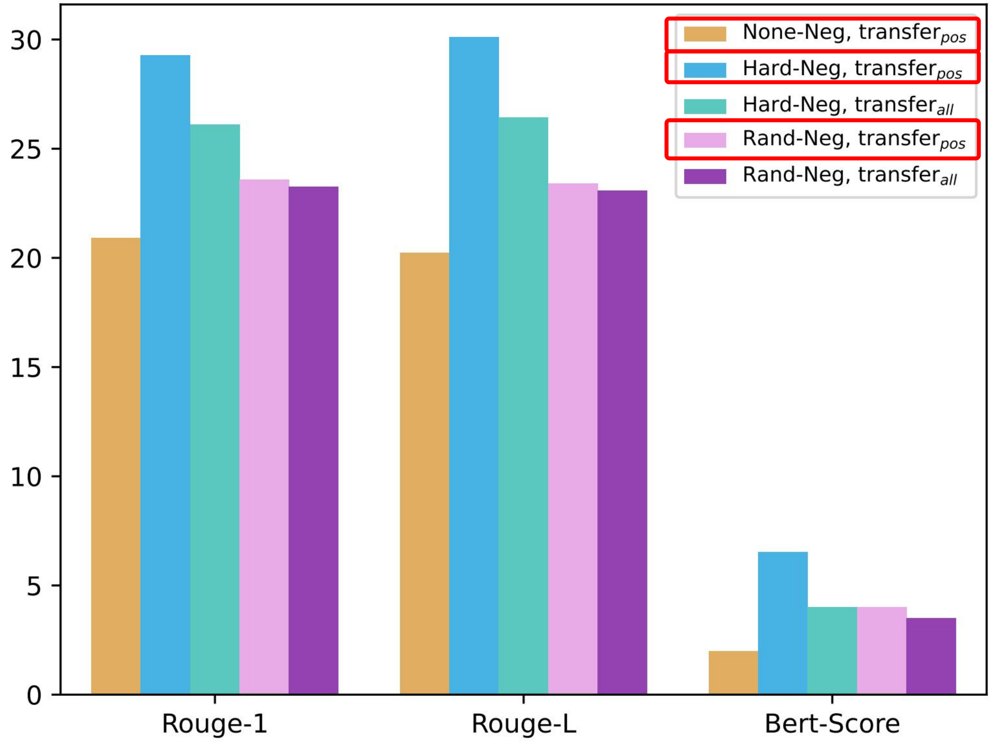


Experiments

- RQ2: Impact of Irrelevant Noisy Context
  - Different negative context settings:
    - □ None Negative Contexts (None-Neg)
    - □ Hard Negative Contexts (Hard-Neg):
      retrieved negative contexts by BM25
    - □ Random Negative Contexts (Rand-Neg):
      randomly sampled negative contexts
  - Different context transfer settings:
    - □ transfer_{pos}: transferring only the positive context
    - □ transfer_{all}: transferring both the positive and negative context

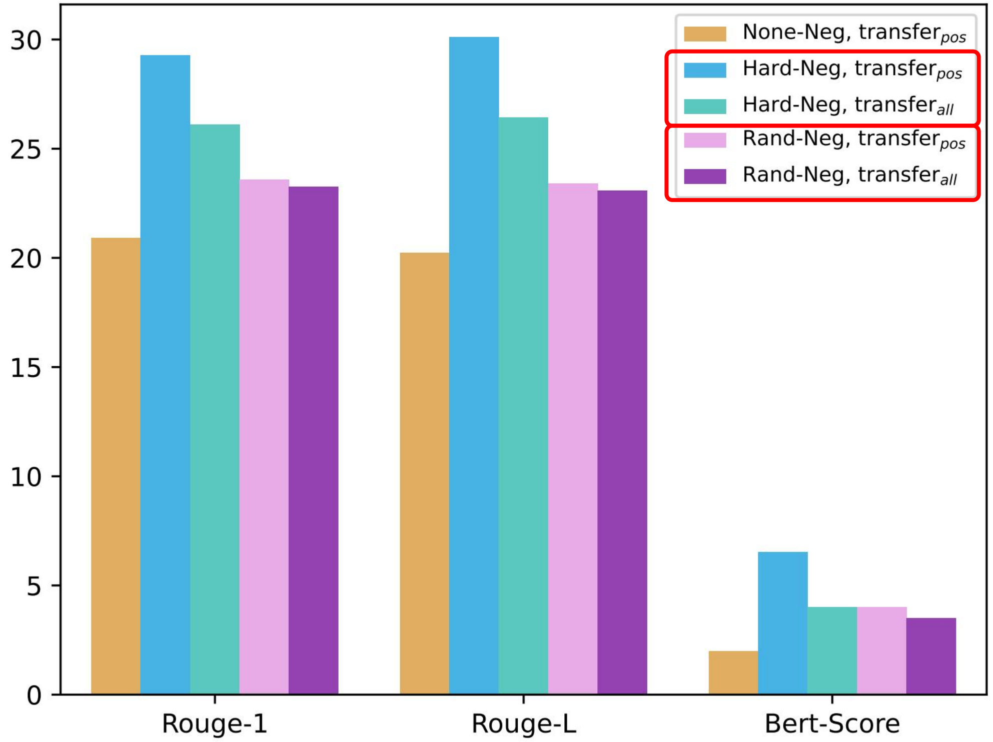
Experiments

- RQ2: Impact of Irrelevant Noisy Context
  - During training phase, encourage model to establish spurious correlations.
  - During testing phase, disperse the model's attention on the answers



Experiments

- RQ2: Impact of Irrelevant Noisy Context
  - During training phase, encourage model to establish spurious correlations.
  - During testing phase, disperse the model's attention on the answers



Takeaways

- Conclusion
  - Examing multiple models, unveiling their potential deficiencies faithfully align contextual knowledge.
  - Emphasizing the pivotal role of (negative-) context in the manifestation of hallucinations during both training and testing phases.

- Future Work
  - Investigation in large language models
  - Effective solution for memory hallucination

Thanks for Listening!

Contact: yanshek.woo@gmail.com