Carnegie Mellon University School of Computer Science

NEULAB

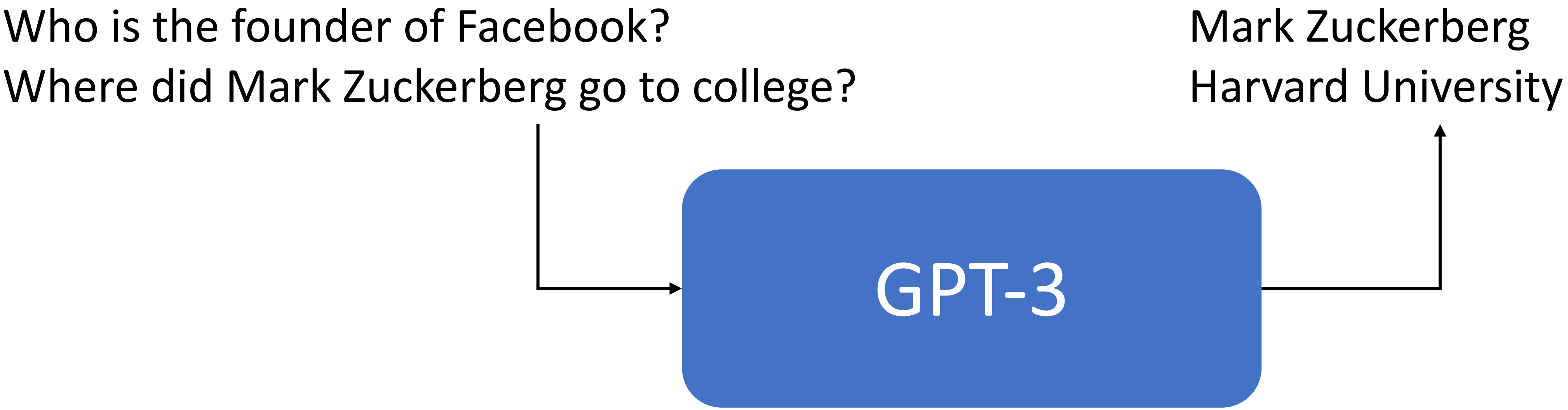
BOSCH Invented for life

Understanding and Improving Zero-shot Multi-hop Reasoning in Generative Question Answering

Zhengbao Jiang, Jun Araki, Haibo Ding, Graham Neubig

zhengbaj@cs.cmu.edu

Background



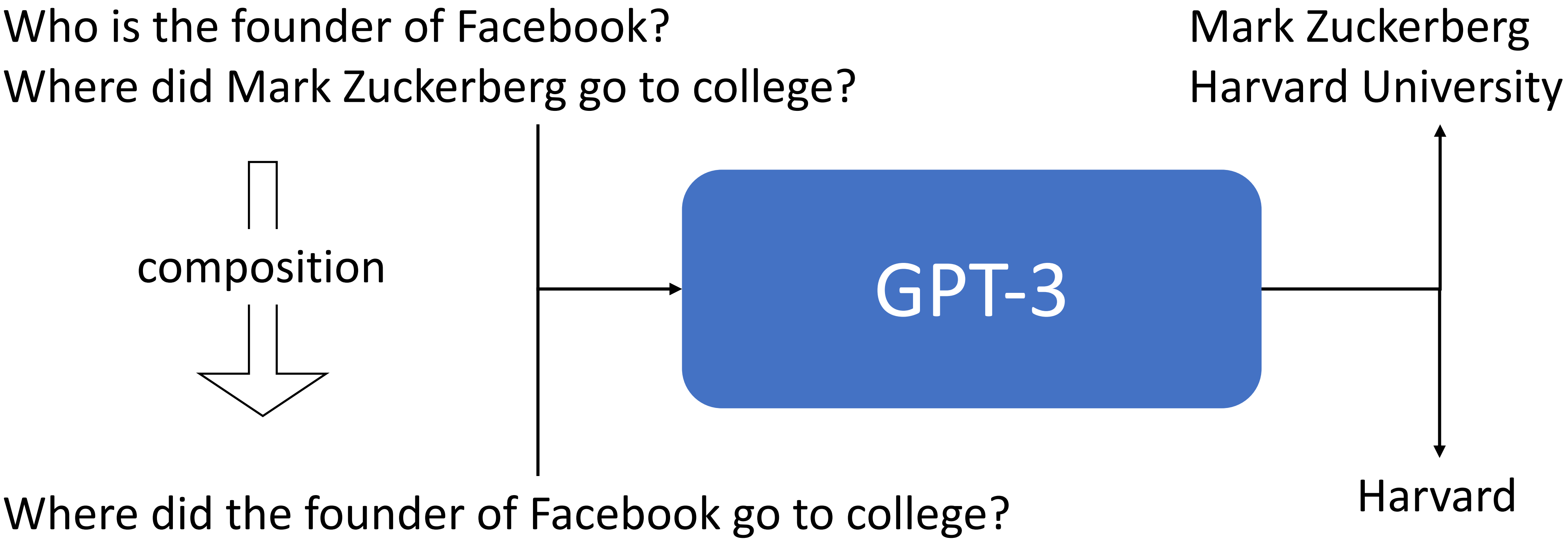
Who is the founder of Facebook?
Where did Mark Zuckerberg go to college?

GPT-3

Mark Zuckerberg
Harvard University

LMs have a decent amount of knowledge

Background



Who is the founder of Facebook?

Where did Mark Zuckerberg go to college?

composition

Where did the founder of Facebook go to college?

GPT-3

Mark Zuckerberg

Harvard University

Harvard

LMs seems to be able to understand and answer complex questions

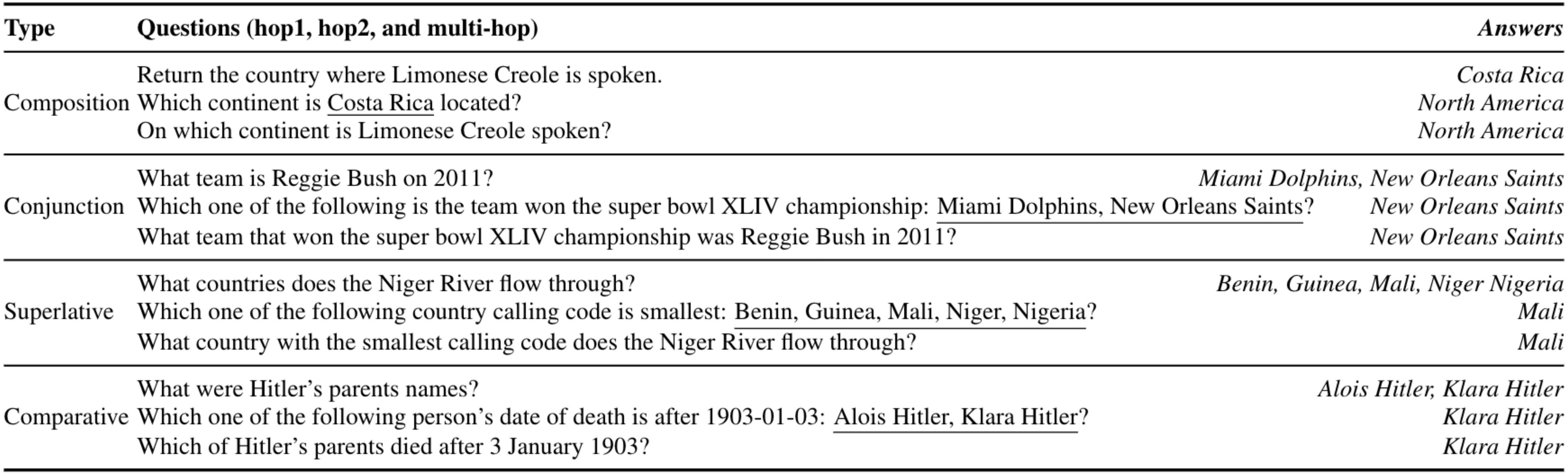
Motivations

- Understand the mechanism through which LMs answer complex questions
  - Correct on single-hop questions ←?→ correct on multi-hop questions?
  - Are answers to multi-hop question and chains of single-hop questions consistent?
  - Dose models trained on single-hop questions generalize to multi-hop questions?

- Improve models’ zero-shot multi-hop reasoning capacity
  - Train on concatenated single-hop questions.
  - Train on SPARQL queries.

Generative Question Answering

- Datasets: ComplexWebQuestions
  - Four types of multi-hop questions
    - Composition, conjunction, superlative, comparative
  - Decompose each multi-hop question q into two single-hop questions q_1 q_2.



Type	**Questions (hop1, hop2, and multi-hop)**	***Answers***
Composition	Return the country where Limonese Creole is spoken. Which continent is Costa Rica located? On which continent is Limonese Creole spoken?	*Costa Rica* *North America* *North America*
Conjunction	What team is Reggie Bush on 2011? Which one of the following is the team won the super bowl XLIV championship: Miami Dolphins, New Orleans Saints? What team that won the super bowl XLIV championship was Reggie Bush in 2011?	*Miami Dolphins, New Orleans Saints* *New Orleans Saints* *New Orleans Saints*
Superlative	What countries does the Niger River flow through? Which one of the following country calling code is smallest: Benin, Guinea, Mali, Niger, Nigeria? What country with the smallest calling code does the Niger River flow through?	*Benin, Guinea, Mali, Niger Nigeria* *Mali* *Mali*
Comparative	What were Hitler's parents names? Which one of the following person's date of death is after 1903-01-03: Alois Hitler, Klara Hitler? Which of Hitler's parents died after 3 January 1903?	*Alois Hitler, Klara Hitler* *Klara Hitler* *Klara Hitler*

Table 1: Four types of multi-hop questions and their decomposed single-hop questions. Intermediate answer is underlined.

Generative Question Answering

- Experimental settings and models
  - Close-book QA: q → a
    - Model: UnifiedQA, T5 (3B) trained on multiple QA datasets in seq2seq format.
  - Open-book QA: q, c → a
    - Model: RAG, BART (base) model augmented with DPR as retriever.
    - Context of single-hop questions: 1 positive + 1 negative.
    - Context of multi-hop questions: concatenate the context of q_1 and q_2.

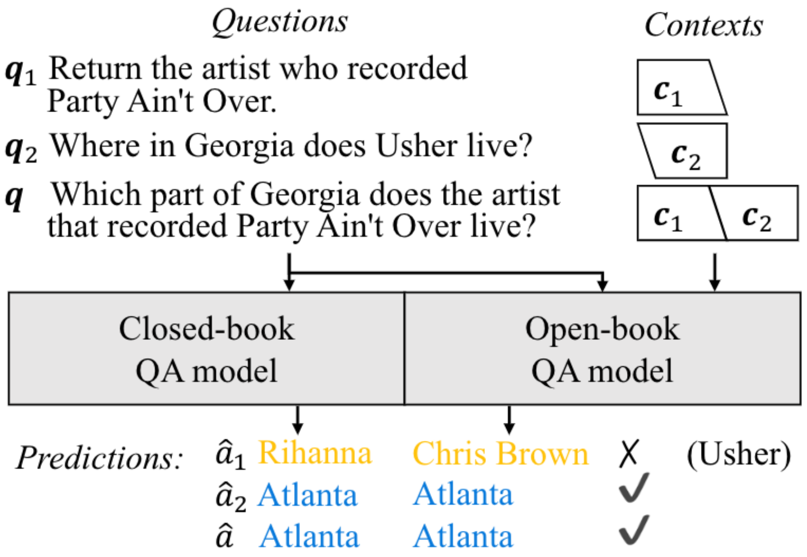
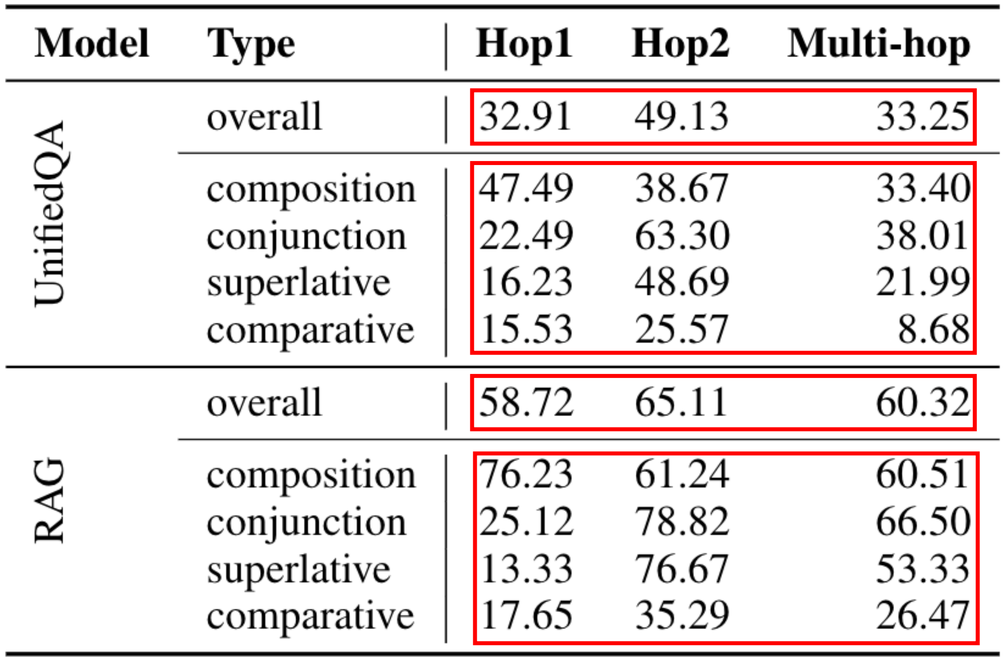


Figure 1: Close- and open-book experimental settings.

Overall Performance

- Evaluation
  - Finetune models (UnifiedQA, RAG) on q_1, q_2 and q from the train split.
  - Test on q_1, q_2 and q from the test split using exact match as the metric.
- Observation
  - UnifiedQA (close-book) < RAG (open-book)
  - Hop2 > Multi-hop ≈ Hop1
  - Superlative and comparative are harder



Model	Type	Hop1	Hop2	Multi-hop
UnifiedQA	overall	32.91	49.13	33.25
	composition	47.49	38.67	33.40
	conjunction	22.49	63.30	38.01
	superlative	16.23	48.69	21.99
	comparative	15.53	25.57	8.68
RAG	overall	58.72	65.11	60.32
	composition	76.23	61.24	60.51
	conjunction	25.12	78.82	66.50
	superlative	13.33	76.67	53.33
	comparative	17.65	35.29	26.47

Table 2: Overall performance on ComplexWebQuestions.

Correlation of Correctness

- Notations
  - s_1, s_2 and s: correctness (0/1) of q_1, q_2 and q.
  - $P(s, s_1, s_2)$: percentage of a certain correctness
- Bucket all examples based on correctness s_1, s_2

$P(s = 0, s_1 s_2 = 11)$ $P(s = 1, s_1 s_2 = 11)$

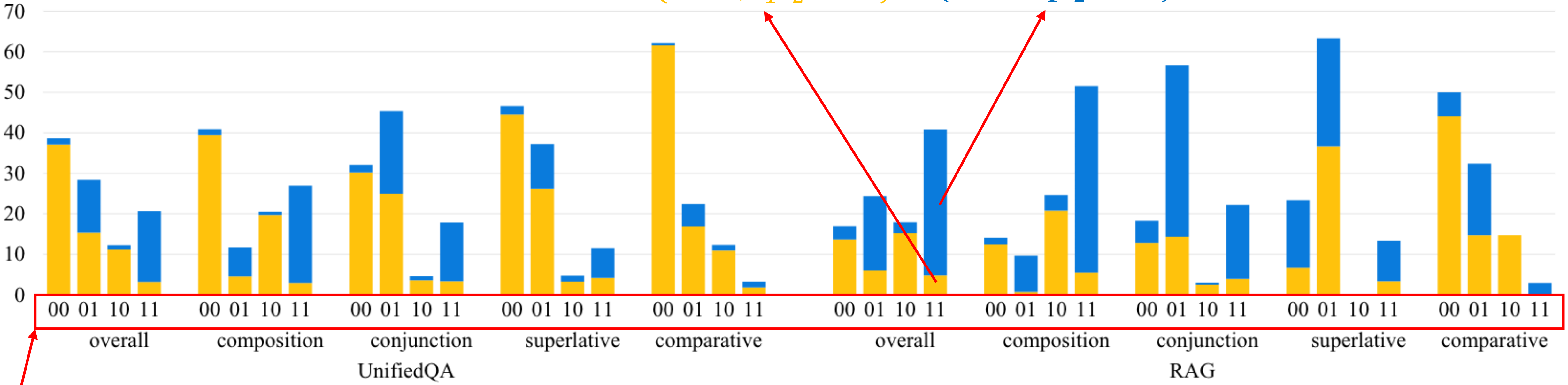


Figure 2: Correlation of correctness between single- and multi-hop questions.

correctness of single-hop questions s_1, s_2

Correlation of Correctness

- Observations
  - Success on single-hop questions does not always imply success on multi-hop questions.

$P(s = 0, s_1 s_2 = 11) > 0$

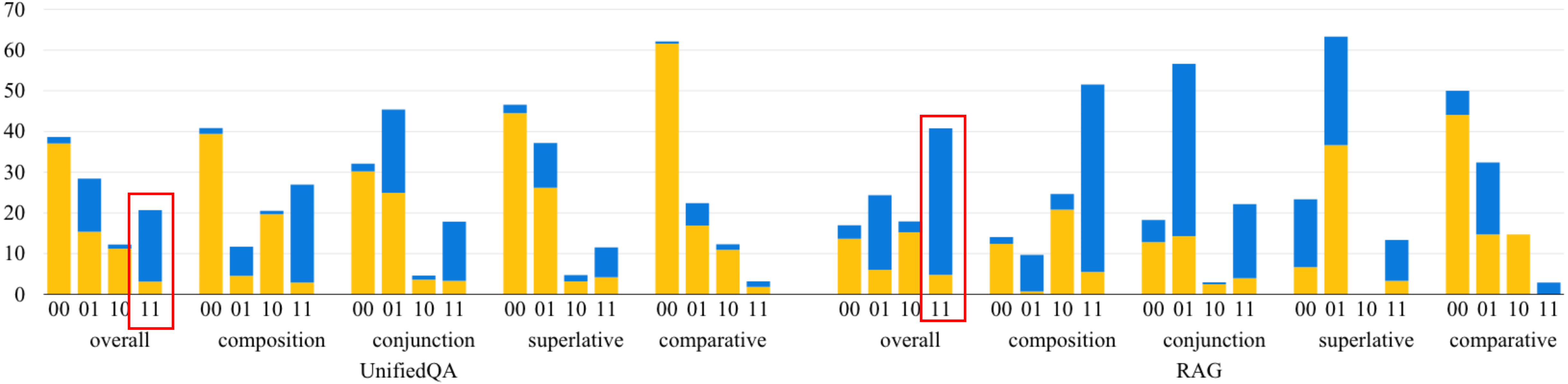


Figure 2: Correlation of correctness between single- and multi-hop questions.

Correlation of Correctness

- Observations
  - Success on single-hop questions does not always imply success on multi-hop questions.
  - Failure on single-hop questions does not always imply failure on multi-hop questions.

$P(s = 1, s_1 s_2 = 11) > 0$

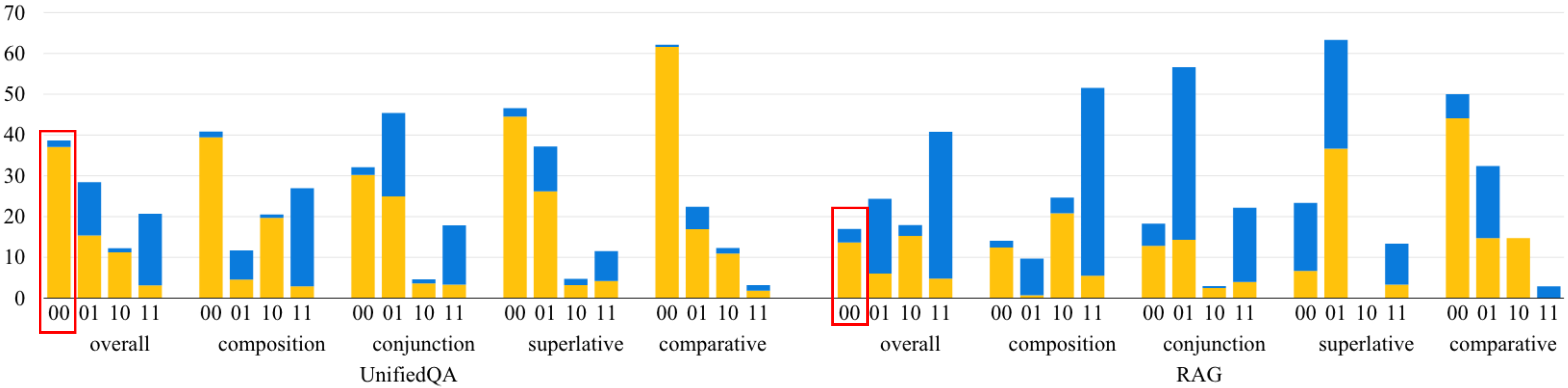


Figure 2: Correlation of correctness between single- and multi-hop questions.

Correlation of Correctness

- Observations
  - Success on single-hop questions does not always imply success on multi-hop questions.
  - Failure on single-hop questions does not always imply failure on multi-hop questions.
  - Multi-hop success is correlated with last-hop success, i.e., short cuts.

$$P(s = 1, s_1 s_2 = 01) > P(s = 1, s_1 s_2 = 10)$$

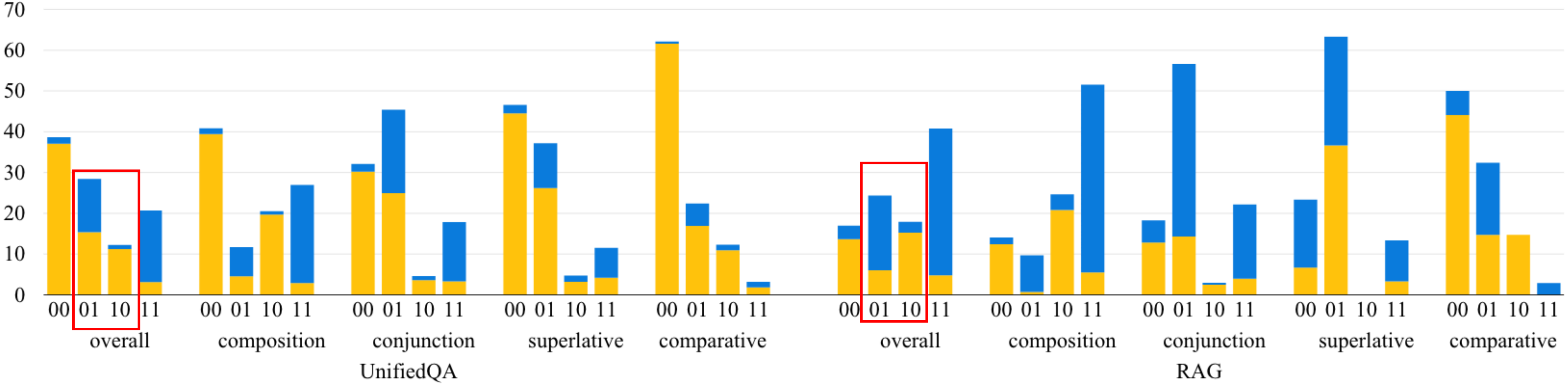


Figure 2: Correlation of correctness between single- and multi-hop questions.

Prediction Consistency

- Experimental settings
  - Query models using:
    - following single-hop questions, where the generate answer to q_1 is filled into the q_2.
    - multi-hop question q.
  - Whether the final generate answer is the same.
- Observation
  - Consistency is relatively low especially for the close-book UnifiedQA model.
  - Harder questions (superlative/comparative) as less consistent.

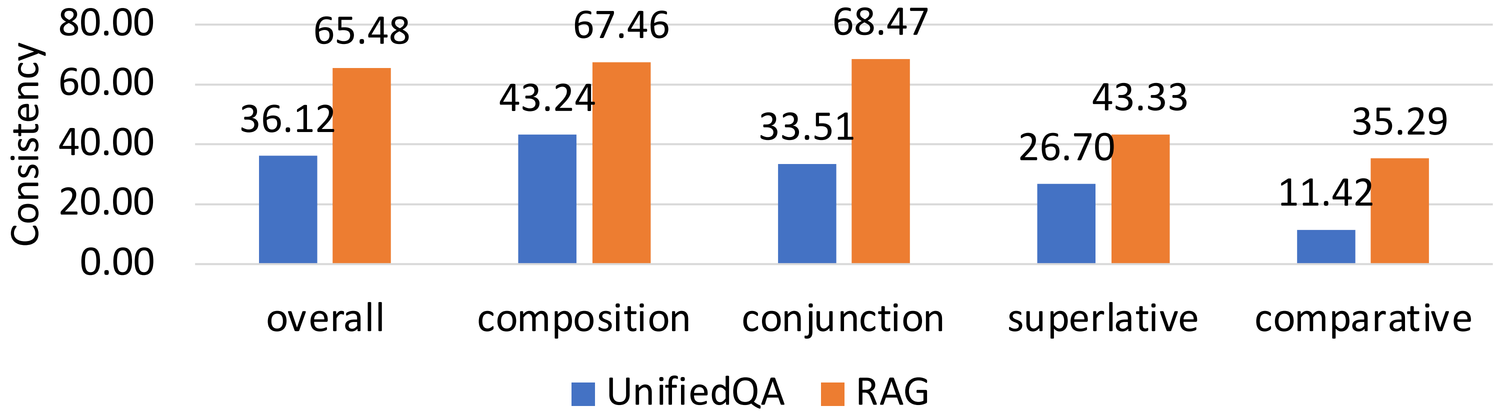


Figure 3: Prediction consistency.

Poor Zero-shot Multi-hop Performance

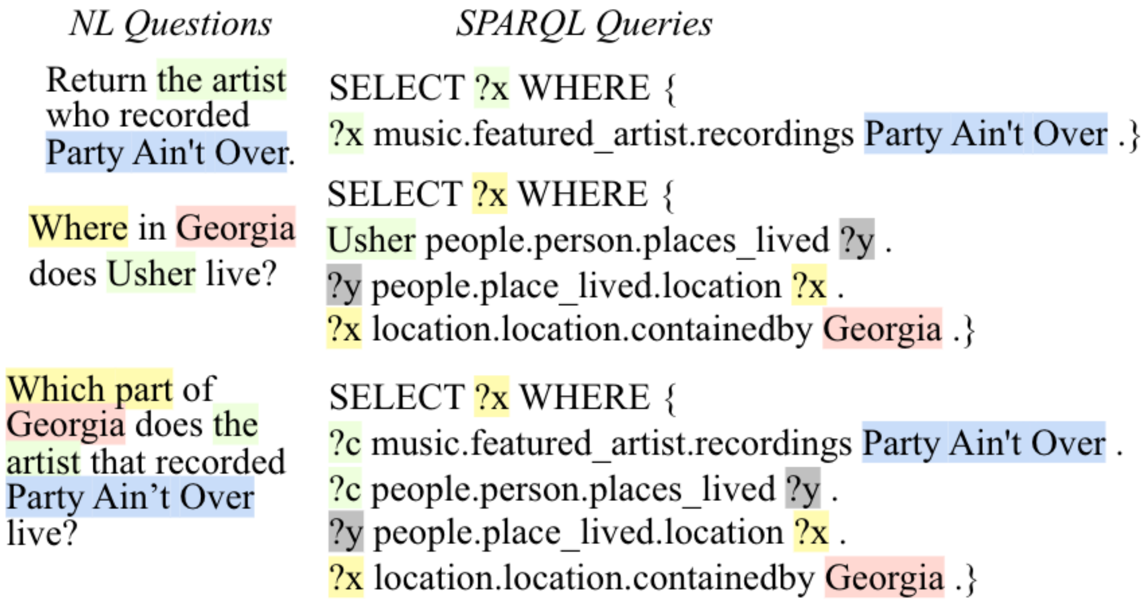
- Multi-hop question performance (UnifiedQA/RAG)
  - Train on **both** single- and multi-hop question: 33.25/60.32.
  - Train on **only** single-hop questions (zero-shot): 17.02/34.03.

Improve Zero-shot Multi-hop Reasoning

- Approximate multi-hop questions
    - (1) Simply concatenating two single-hop questions
        - Motivation: LMs can identify semantically similar expressions
        - Example
            - q_1: Return the artist who recorded Party Ain’t Over.
            - q_2: Where in Georgia dose #1 live?
            - q: Which part of Georgia does the artist that recorded Party Ain’t Over live?
            - Concatenation: *Return the artist who recorded Party Ain’t Over. Where in Georgia dose #1 live?*

Improve Zero-shot Multi-hop Reasoning

- Approximate multi-hop questions
  - (2) Use SPARQL as pseudo questions and train LMs to “execute” them
    - Motivated by TAPEX (Liu et al., 2021): training on structured language endows LMs with reasoning capabilities.
    - Example



NL Questions	*SPARQL Queries*

Return the artist who recorded Party Ain't Over.	SELECT ?x WHERE {
	?x music.featured_artist.recordings Party Ain't Over .}

Where in Georgia does Usher live?	SELECT ?x WHERE {
	Usher people.person.places_lived ?y .
	?y people.place_lived.location ?x .
	?x location.location.containedby Georgia .}

Which part of Georgia does the artist that recorded Party Ain’t Over live?	SELECT ?x WHERE { ?c music.featured_artist.recordings Party Ain't Over . ?c people.person.places_lived ?y . ?y people.place_lived.location ?x . ?x location.location.containedby Georgia .}

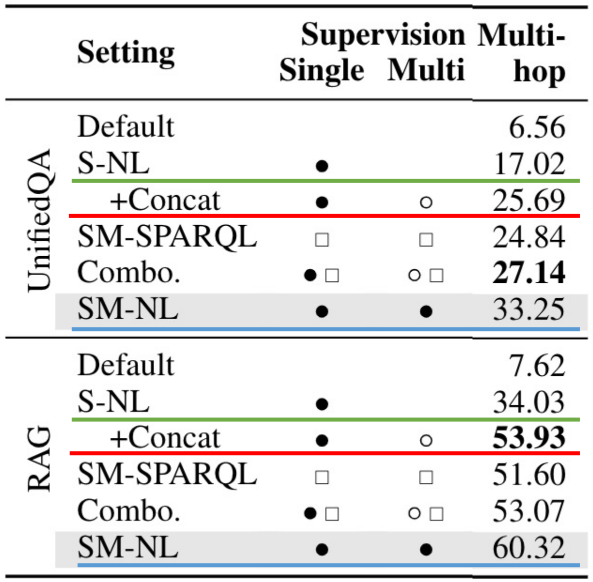
Figure 4: SPARQL queries of single- and multi-hop questions.

Improve Zero-shot Multi-hop Reasoning

- Experimental settings
  - Notations
    - **S**ingle-hop question, **M**ulti-hop question
    - ●, ○, □ denotes NL question, concatenation, and SPARQL
  - 5 experimental settings:
    - S-NL (zero-shot): single-hop NL question.
    - S-NL + concat.: single-hop NL question + concatenation.
    - SM-SPARQL: single- and multi-hop SPARQL queries
    - S-NL + concat + SM-SPARQL (combo): all above
    - SM-NL (upper bound): use both single- and multi-hop NL questions.

Improve zero-shot multi-hop reasoning capacity

- Conclusion
  - Concatenation is a good approximation of multi-hop questions (red > green by 7-20%).



	Setting	Supervision Single	 Multi	Multi- hop
UnifiedQA	Default			6.56
	S-NL	●		17.02
	+Concat	●	○	25.69
	SM-SPARQL	□	□	24.84
	Combo.	● □	○ □	**27.14**
	SM-NL	●	●	33.25
RAG	Default			7.62
	S-NL	●		34.03
	+Concat	●	○	**53.93**
	SM-SPARQL	□	□	51.60
	Combo.	● □	○ □	53.07
	SM-NL	●	●	60.32

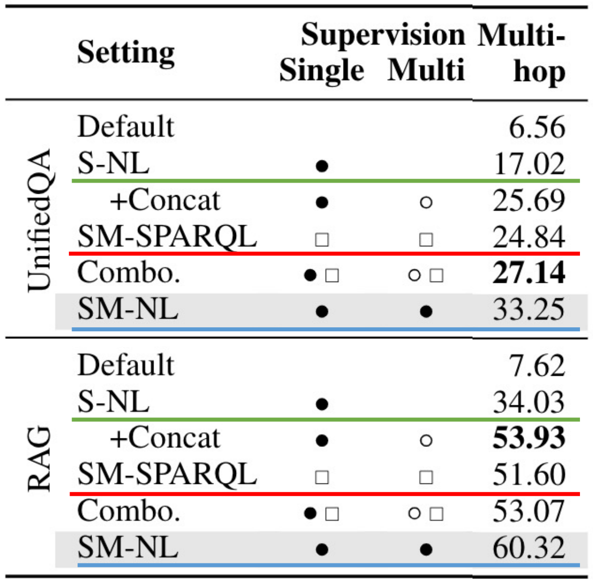
●, ○, □

NL question, concatenation, SPARQL

Table 3: Performance of different multi-hop question approximation methods. Green is baseline and blue is upper bound.

Improve zero-shot multi-hop reasoning capacity

- Conclusion
  - Concatenation is a good approximation of multi-hop questions.
  - Models can generalize from SPARQL to NL questions (red > green by 7-17%).



	Setting	Supervision Single	Supervision Multi	Multi-hop
UnifiedQA	Default			6.56
	S-NL	●		17.02
	+Concat	●	○	25.69
	SM-SPARQL	□	□	24.84
	Combo.	● □	○ □	**27.14**
	SM-NL	●	●	33.25
RAG	Default			7.62
	S-NL	●		34.03
	+Concat	●	○	**53.93**
	SM-SPARQL	□	□	51.60
	Combo.	● □	○ □	53.07
	SM-NL	●	●	60.32

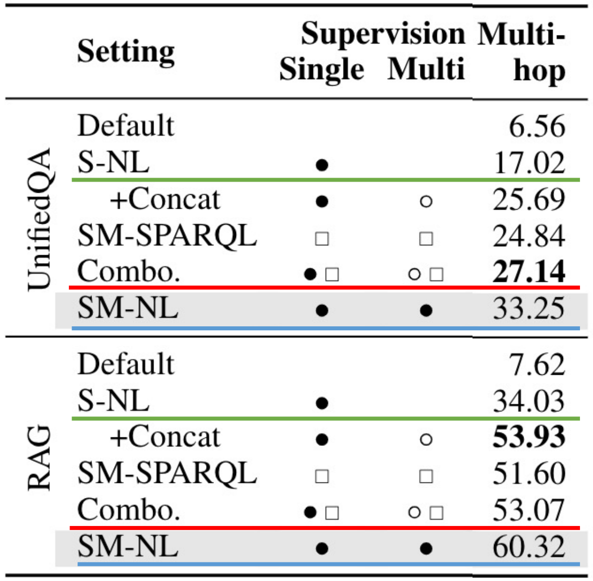
●, ○, □

NL question, concatenation, SPARQL

Table 3: Performance of different multi-hop question approximation methods. Green is baseline and blue is upper bound.

Improve zero-shot multi-hop reasoning capacity

- Conclusion
  - Concatenation is a good approximation of multi-hop questions.
  - Models can generalize from SPARQL to NL questions.
  - Combining both further improves on UnifiedQA (24.84 → 27.14).



	Setting	Supervision		Multi-
		Single	Multi	hop
UnifiedQA	Default			6.56
	S-NL	●		17.02
	+Concat	●	○	25.69
	SM-SPARQL	□	□	24.84
	Combo.	● □	○ □	**27.14**
	SM-NL	●	●	33.25
RAG	Default			7.62
	S-NL	●		34.03
	+Concat	●	○	**53.93**
	SM-SPARQL	□	□	51.60
	Combo.	● □	○ □	53.07
	SM-NL	●	●	60.32

●, ○, □

NL question, concatenation, SPARQL

Table 3: Performance of different multi-hop question approximation methods.
Green is baseline and blue is upper bound.

Future Work

- Examine larger language models such as OPT, GPT-3, and PaLM.
- Develop better multi-hop question approximation methods.

Questions?