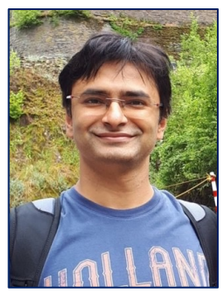
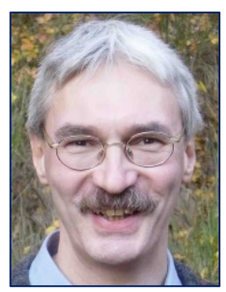
CompMix: A Benchmark for Heterogeneous Question Answering

Philipp Christmann, Rishiraj Saha Roy, Gerhard Weikum

Max Planck Institute for Informatics
Saarbrücken, Germany

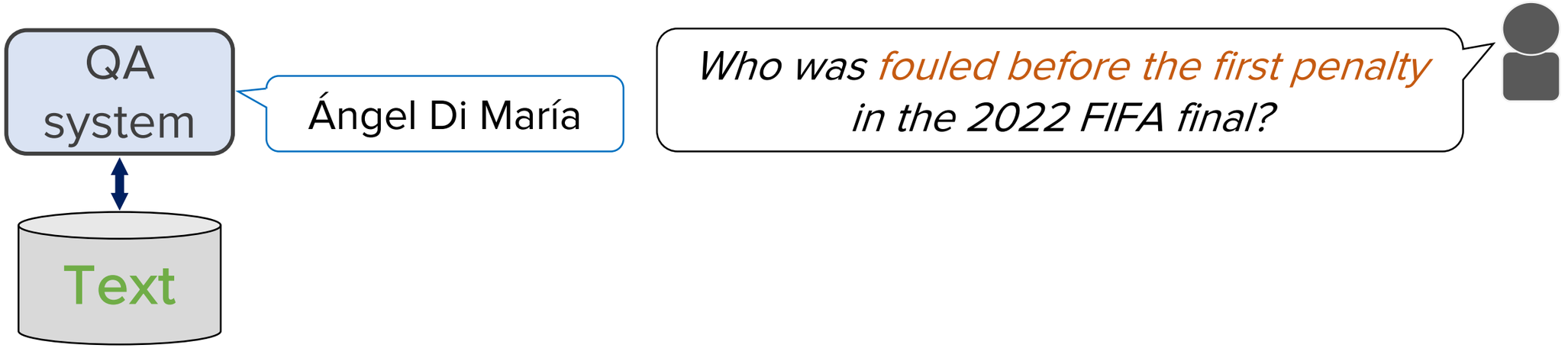




MAX PLANCK INSTITUTE FOR INFORMATICS

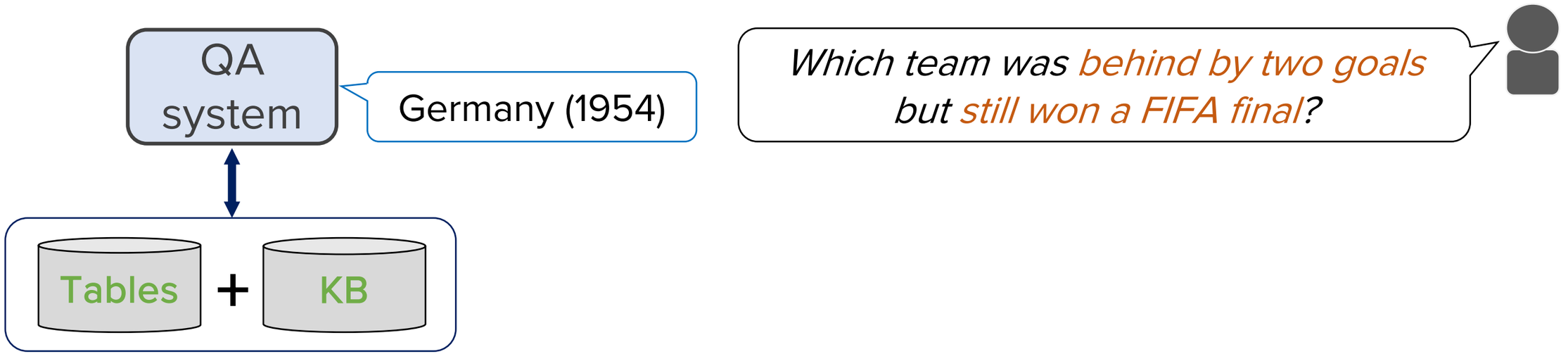
WWW 2024, Singapore

SIC Saarland Informatics Campus

Heterogeneous question answering

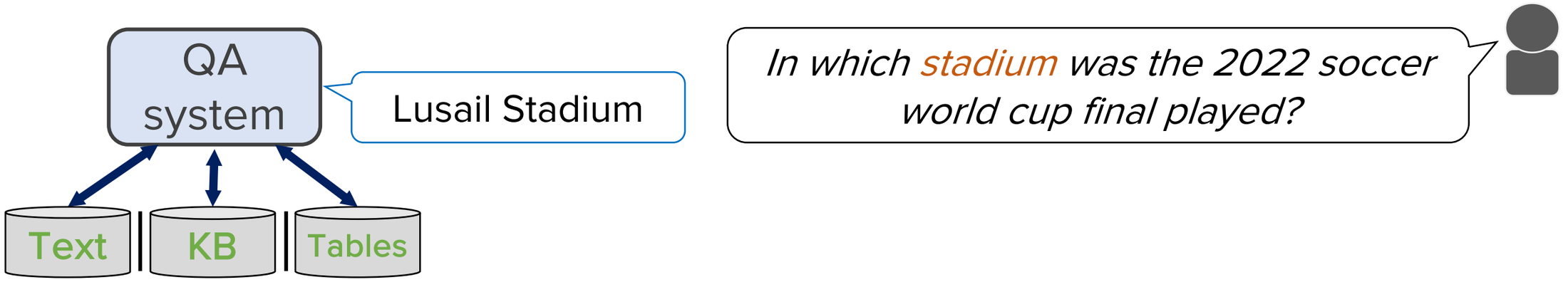


Heterogeneous question answering



⇒ QA systems operating over **heterogeneous sources as a result**
⇒ Integrate **multiple** sources…
⇒ …for **broader answer coverage**

Heterogeneous question answering



⇒ QA systems operating over **heterogeneous sources as a result**
⇒ Integrate **multiple** sources…
⇒ …for **broader answer coverage**
⇒ **…**to leverage answer redundancy

Benchmarks for heterogeneous QA

Existing benchmarks

✗ Questions **not human** generated

✗ Span **only two** sources

✗ **Domain-specific**

✗ **…**

Benchmarks for heterogeneous QA

Existing benchmarks

✗ Questions **not human** generated

✗ Span **only two** sources

✗ **Domain-specific**

✗ ...

⇒ Methods for heterogeneous QA typically **consider source-specific** benchmarks, and **artificially weaken** one of the input sources (e.g. dropping 50% of the KB)

CompMix aims to fill this gap

Existing benchmarks

✗ Questions **not human** generated

✗ Span **only two** sources

✗ **Domain-specific**

✗ ...

CompMix benchmark (ours)

✓ Questions **crowdsourced**

✓ Spans **four** kinds of sources
(KB, text, tables, infoboxes)

✓ Covers **five different** domains
(books, movies, music, tv series, soccer)

✓ ...

⇒ Collate **completed questions** from **ConvMix** (dataset for conversational QA)
⇒ CompMix has been **used** as a benchmark **already**

CompMix: Key statistics

★ 9,410 questions

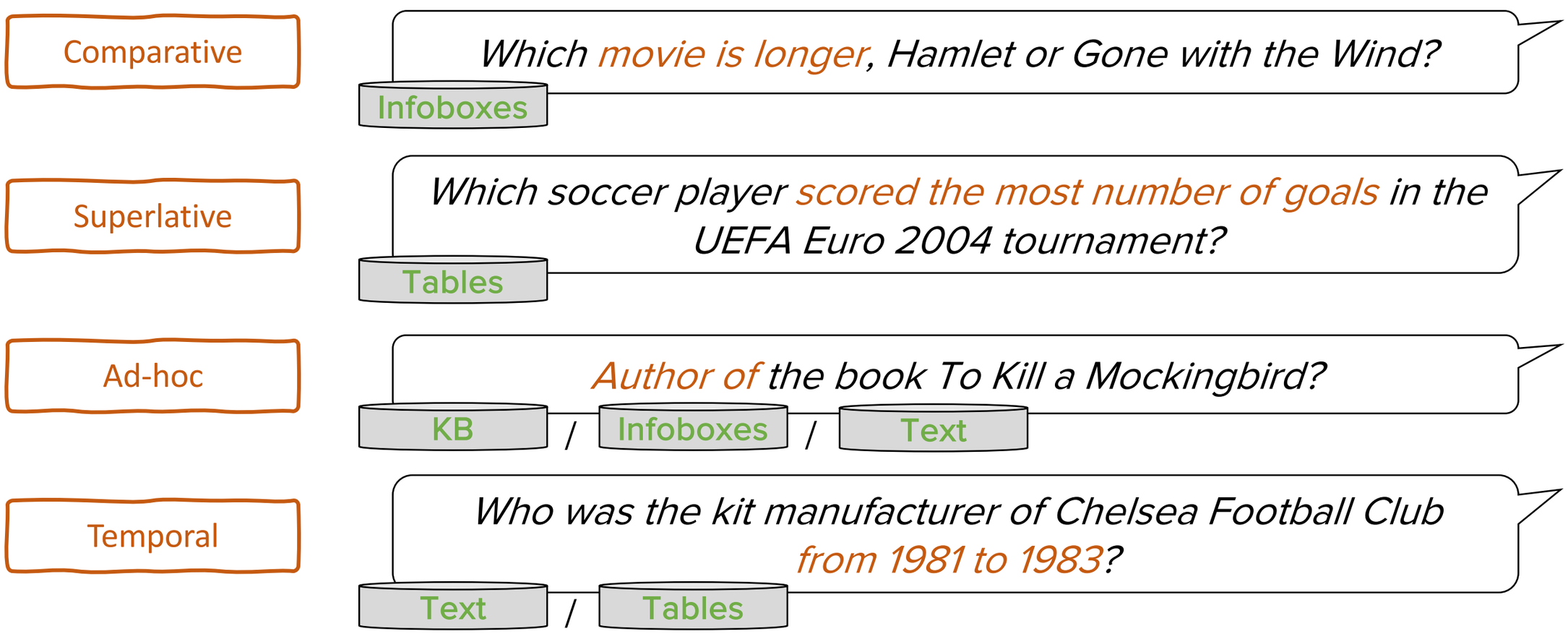
★ 9.19 words per question

★ 2.17 words per answer

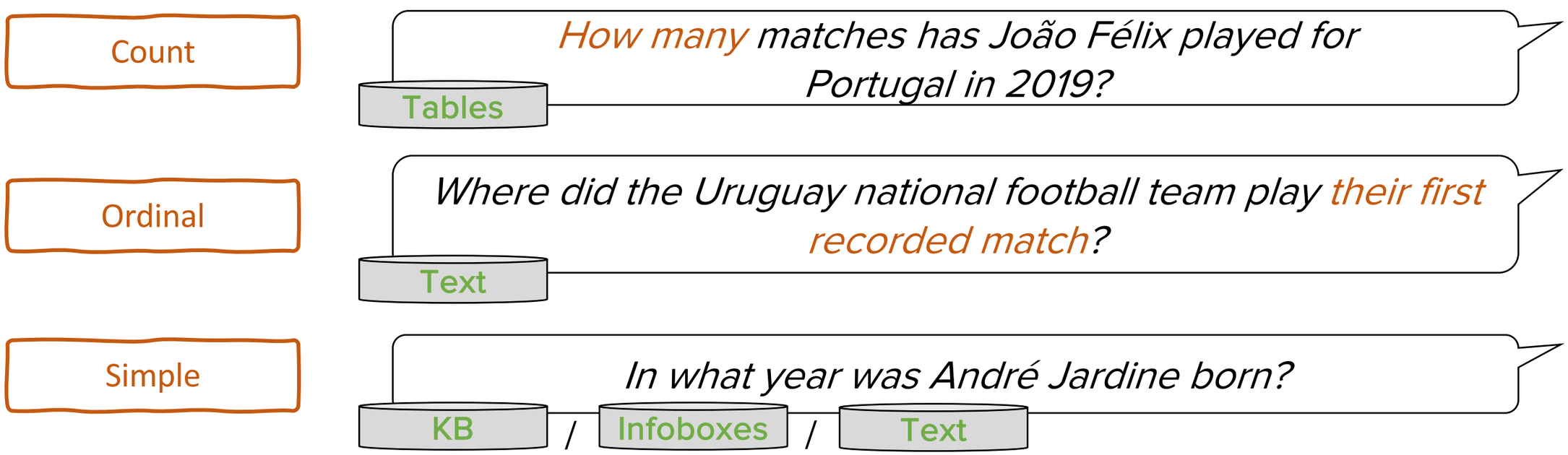
★ 5,413 entities

2,511 long-tail entities (<50 KB-facts)

CompMix: Complex phenomena

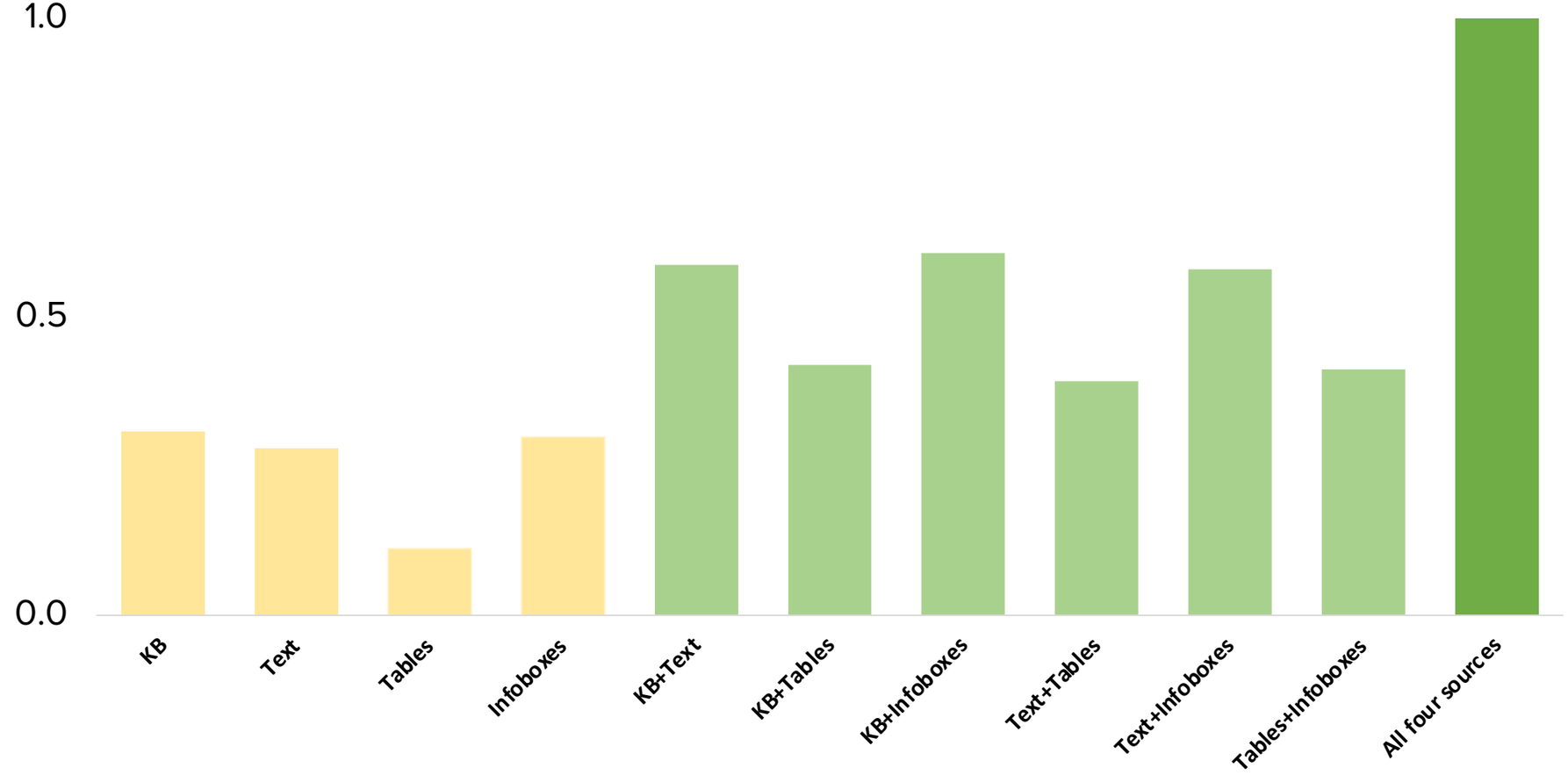


CompMix: Complex phenomena

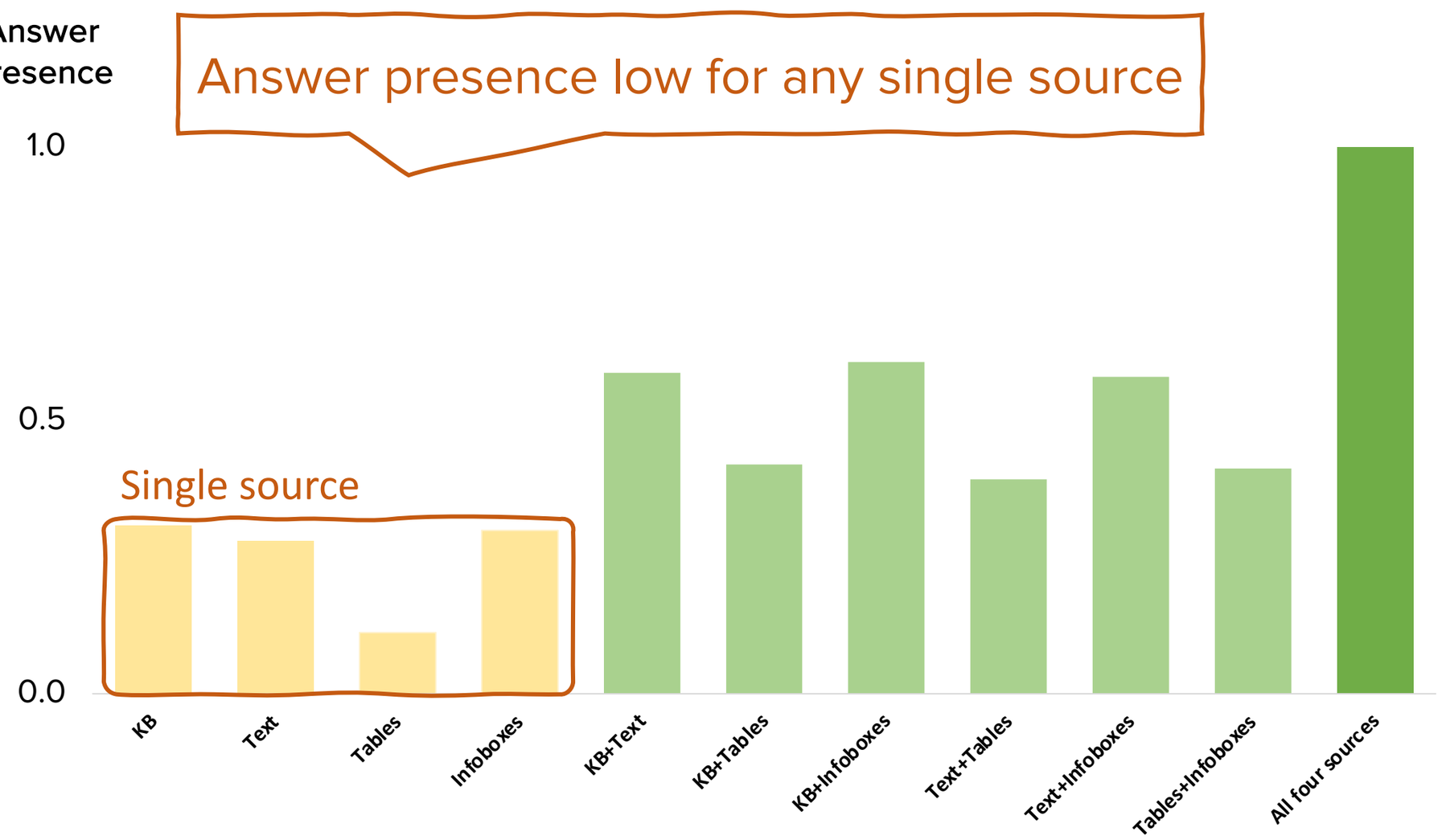


Answer presence per source

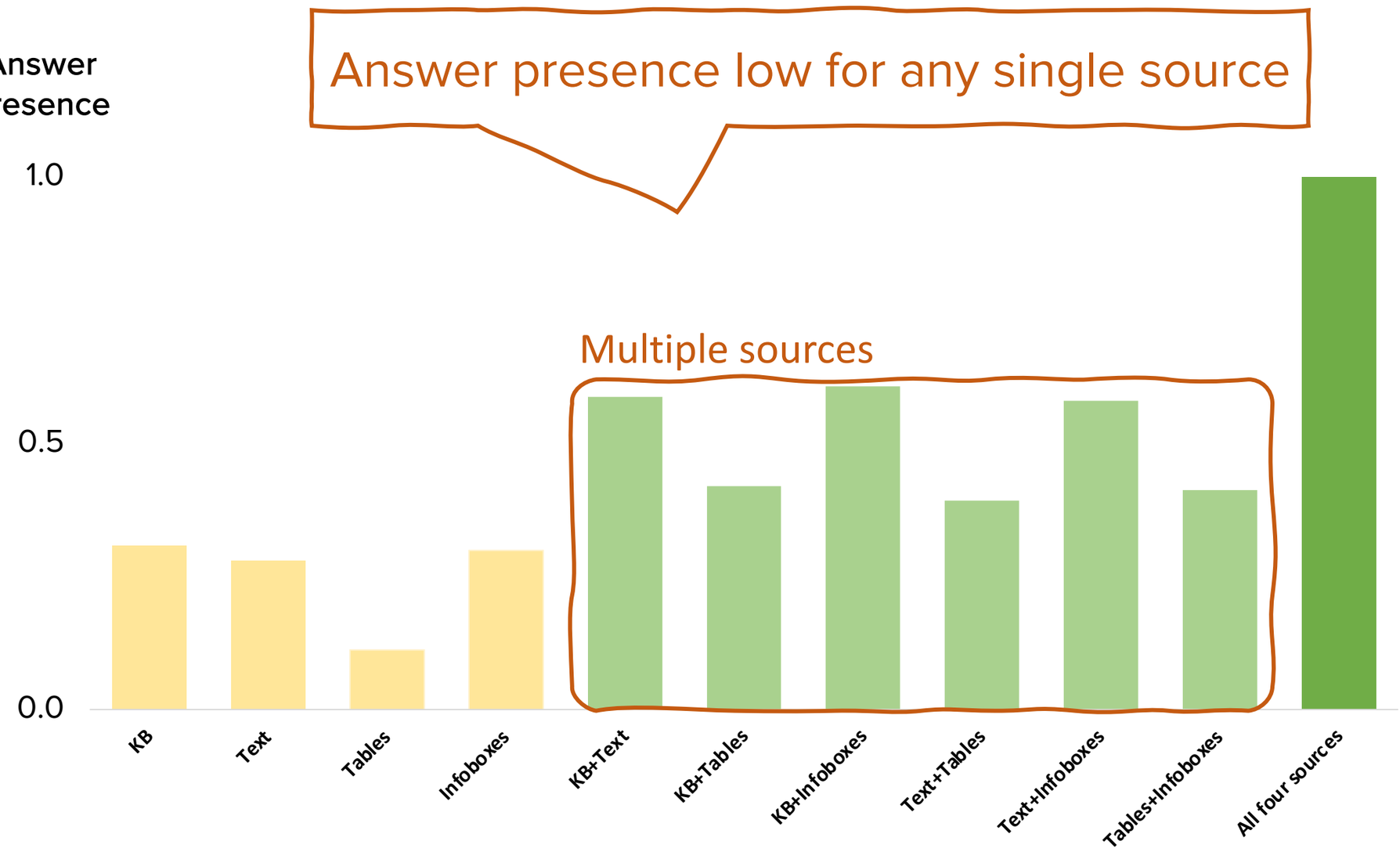
How often the answer is present for any of the sources / source combinations



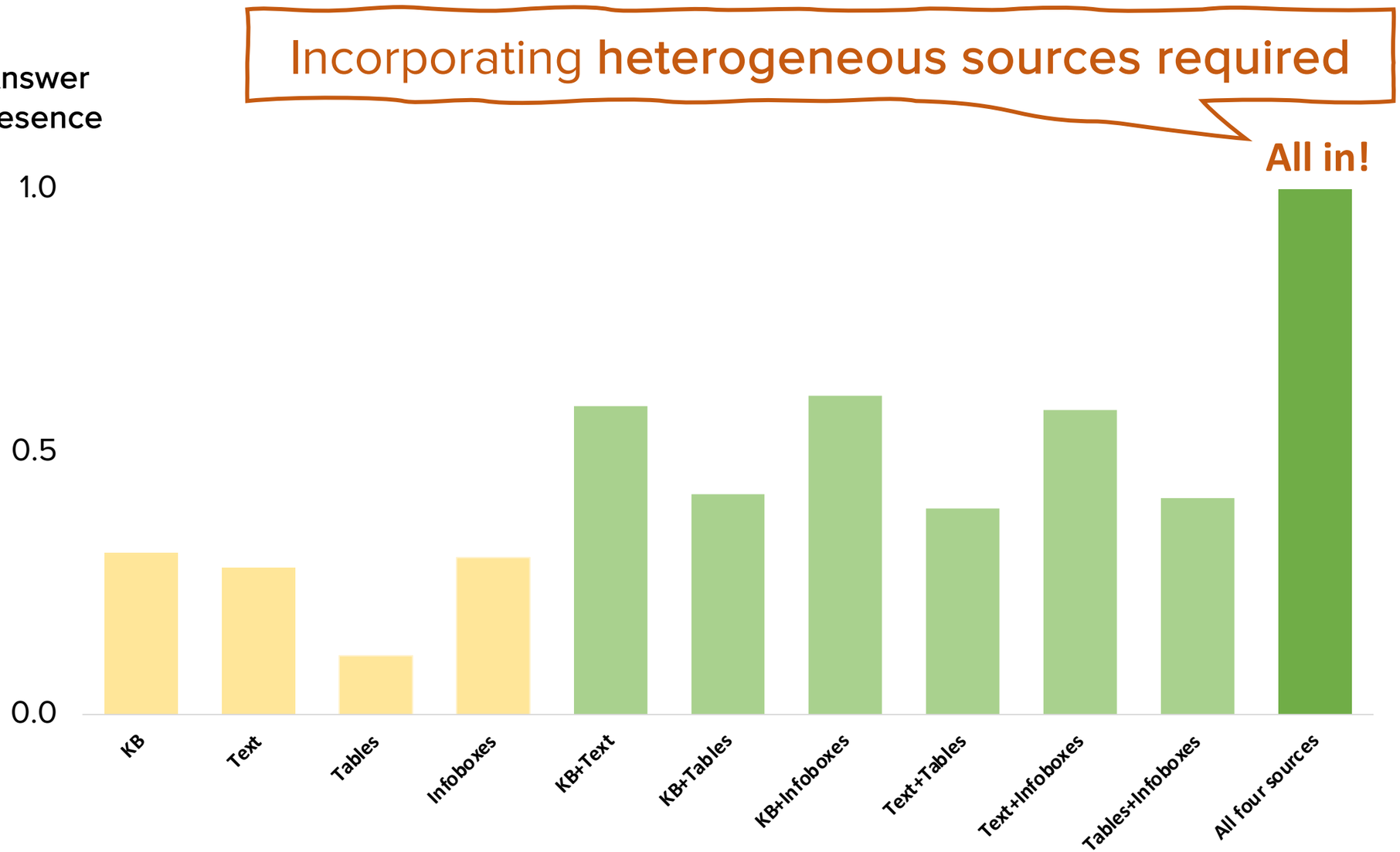
Answer presence per source



Answer presence per source

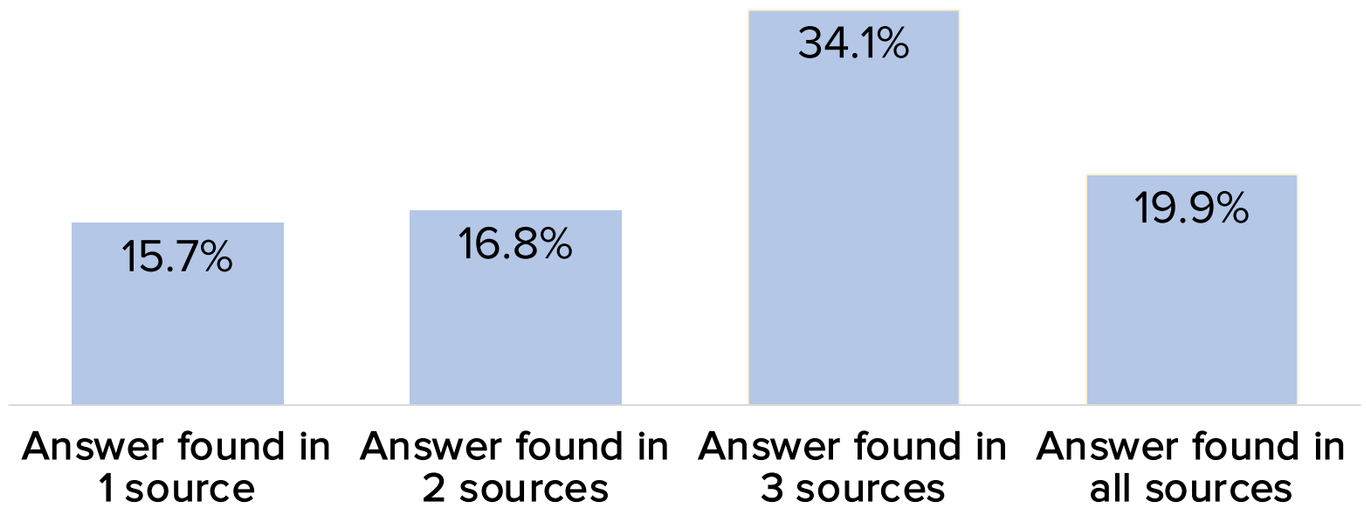


Answer presence per source



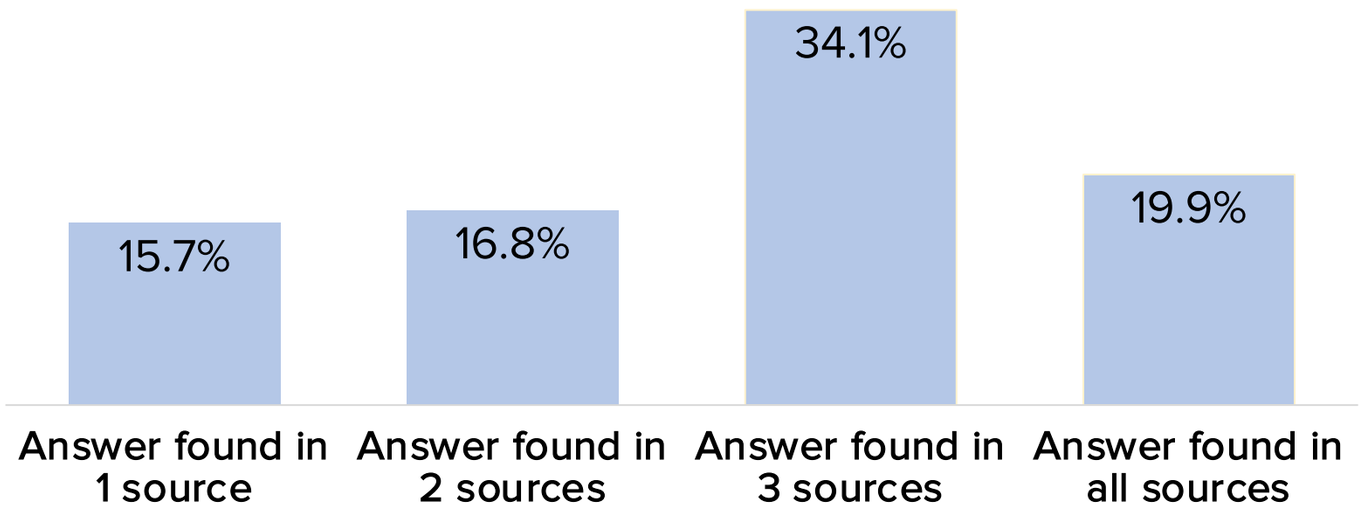
Answer redundancy

In how many information sources each answer can be found



Answer redundancy

For many questions, answer can be found in ≥3 different sources.
=> Leveraging **answer redundancy** can pay off**!**



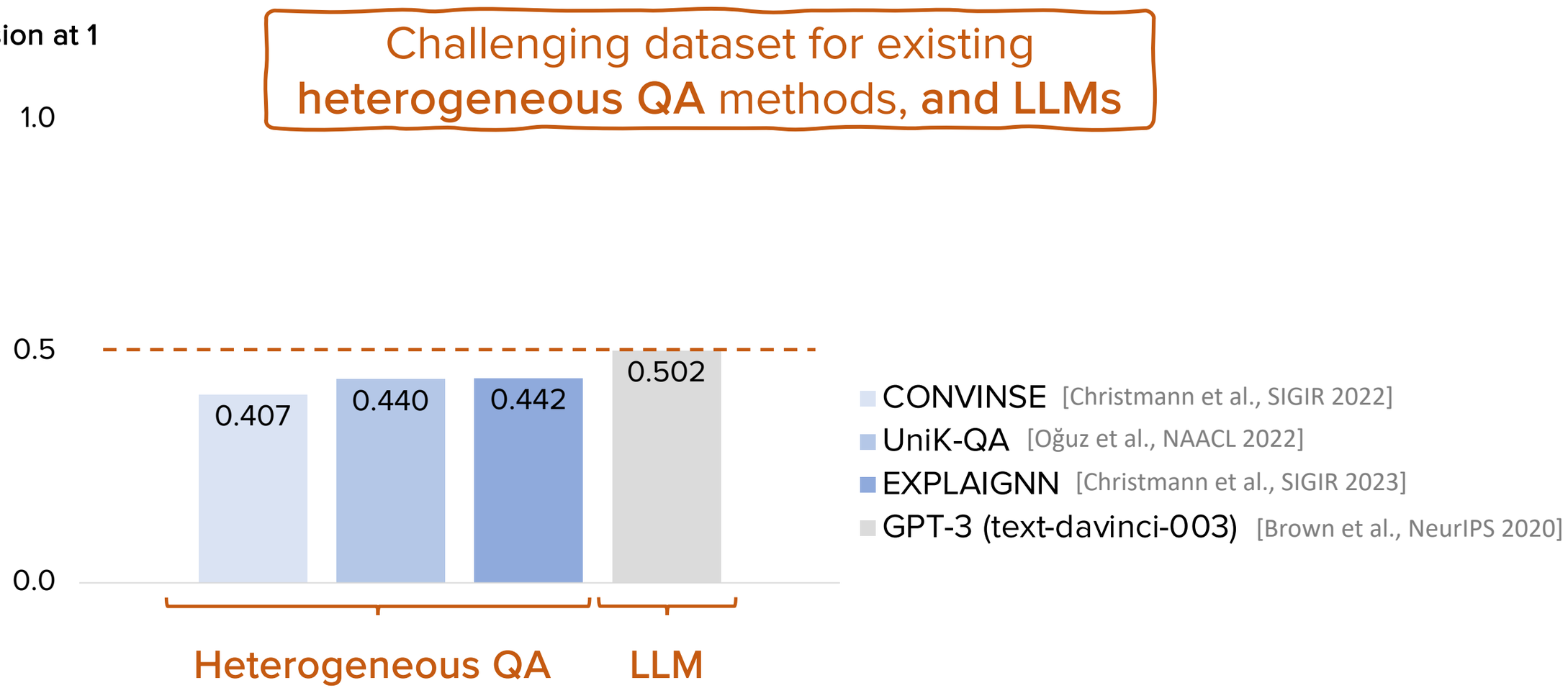
Experimental setup

★ Experiments with diverse range of QA systems
- ★ **Generative LLMs:** GPT-3 (text-davinci-003)
- ★ **Heterogeneous QA:** UNIQORN, Unik-QA, EXPLAIGNN

★ Metrics
- ★ Precision at 1 (**P@1**)
- ★ Mean reciprocal rank
- ★ Hits at 5

Evaluation with CompMix

Challenging dataset for existing **heterogeneous QA** methods, **and LLMs**



Failure cases

None of the methods was able to answer these!

Who played as adult Pi Patel in Life of Pi movie?

Who were the twin brothers who played soccer for Manchester United?

Require judicious combination of heterogeneous sources:
- Textual (“adult Pi Patel” / ”twin brothers”)
- Structured (movie cast / club membership)

Desiderata in heterogeneous QA

Desiderata:

★ Integration of heterogeneous sources

★ Grounding the answer to specific evidence

★ Traceability of the provided answer

EXPLAIGNN [Christmann et al. SIGIR 2023]

Desiderata:

★ Integration of heterogeneous sources

★ Grounding the answer to specific evidence

★ Traceability of the provided answer

Demo at: https://explaignn.mpi-inf.mpg.de

CompMix – Walkthrough

★ Download from website

https://qa.mpi-inf.mpg.de/compmix

★ Load from HuggingFace datasets

```
from datasets import load_dataset

dataset = load_dataset("pchristm/CompMix")
```

CompMix – Walkthrough


```
array [4966]
▼ 0 {8}
    question_id : 3642
    question : What is the genre of the tv series High Seas?
    domain : tvseries
  ▼ entities [1]
    ▼ 0 {2}
        id  : Q59591953
        label : High Seas
▼ answers [1]
  ▼ 0 {2}
        id  : Q186424
        label : detective fiction
  answer_src : kb
  answer_text : detective fiction
```


Question

CompMix – Walkthrough


```
array [4966]
▼ 0 {8}
      question_id : 3642
      question : What is the genre of the tv series High Seas?
      domain : tvseries
   ▼ entities [1]
   ▼ 0 {2}
         id  : Q59591953
         label : High Seas
▼ answers [1]
   ▼ 0 {2}
         id  : Q186424
         label : detective fiction
   answer_src : kb
   answer_text : detective fiction
```


Domain

CompMix – Walkthrough


```
array [4966]
▼ 0 {8}
    question_id : 3642
    question : What is the genre of the tv series High Seas?
    domain : tvseries
  ▼ entities [1]
    ▼ 0 {2}
        id  : Q59591953
        label : High Seas
  ▼ answers [1]
    ▼ 0 {2}
        id  : Q186424
        label : detective fiction
    answer_src : kb
    answer_text : detective fiction
```


Question entities

CompMix – Walkthrough


```
array [4966]
▼ 0 {8}
      question_id : 3642
      question : What is the genre of the tv series High Seas?
      domain : tvseries
    ▼ entities [1]
    ▼ 0 {2}
          id  : Q59591953
          label : High Seas
▼ answers [1]
  ▼ 0 {2}
        id  : Q186424
        label : detective fiction
  answer_src : kb
  answer_text : detective fiction
```


Answer:

- Text
- Wikidata ID
- Answer source
  (used by crowdworker)

Take-aways

★ New **CompMix** dataset for **heterogeneous QA**

★ 9,410 questions

★ Questions generated by humans

★ Integrating **heterogeneous sources inherently required**
(KB, text, tables, infoboxes)

★ Existing methods **merely** answer **50%** of questions **correctly**

Download CompMix at: https://qa.mpi-inf.mpg.de/compmix

Leaderboard

CompMix Leaderboard

Model	P@1	MRR	Hit@5
HQA-GPT-4 * Lehmann et al. '24	**0.655**	-	-
GPT-3 (text-davinci-003) Brown et al. '20	0.502	-	-
EXPLAIGNN Christmann et al. '23	0.442	**0.518**	**0.617**
UniK-QA Oguz et al. '22	0.440	0.467	0.494
CONVINSE Christmann et al. '22	0.407	0.437	0.483

Thank you!