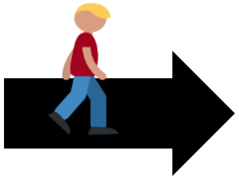
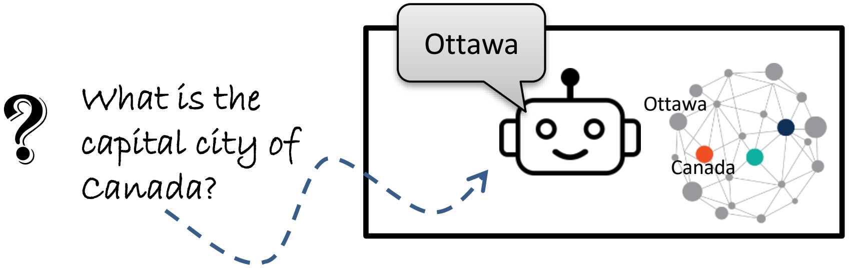
QAKG Past and Future

Towards Next-Generation
Question Answering Over Knowledge Graphs (QAKG) Systems
via Accurate Benchmarking and Large-Scale Training

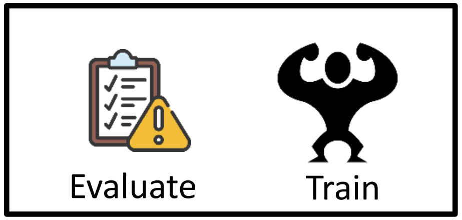
Towards Next-Generation



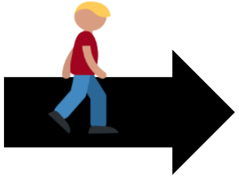
Question Answering Over Knowledge Graphs Systems



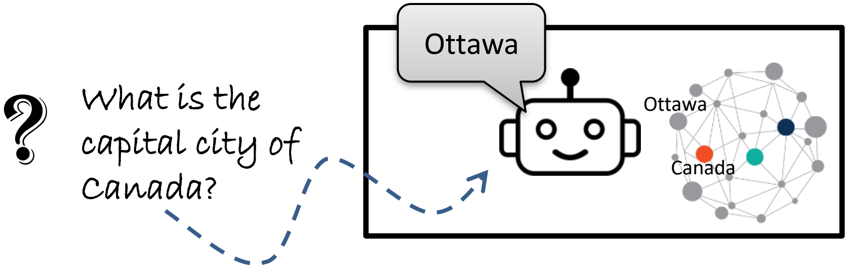
via
Accurate Benchmarking and Large-Scale Training



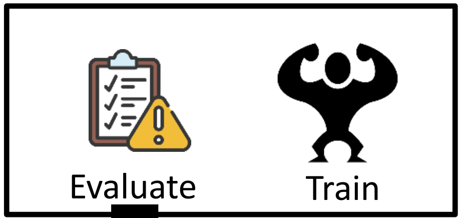
Towards Next-Generation



Question Answering Over Knowledge Graphs Systems

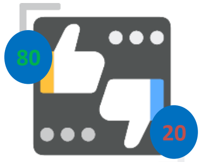


via Accurate Benchmarking and Large-Scale Training



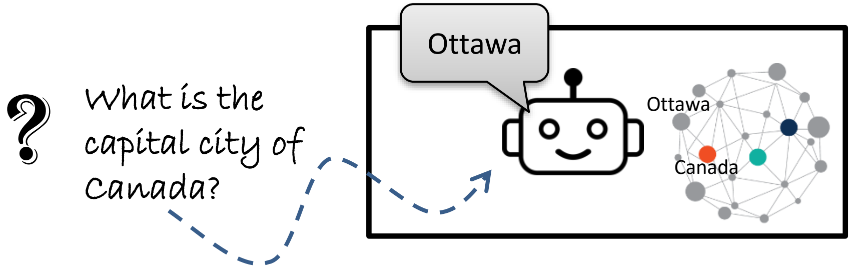
```
NLQ:
     What is the capital of Canada?
Answer:
     Ottawa
Query:
     SELECT DISTINCT ?uri WHERE
     {
        res:Canada dbo:capital  ?uri
     }
```

Benchmark



F-1 Score = 80%

QAKG Past and Future



NLQ:

What is the capital of Canada?

Answer:

Ottawa

Query:

```
SELECT DISTINCT ?uri WHERE
{
   res:Canada dbo:capital  ?uri
}
```

Benchmark

QAKG Past and Future

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

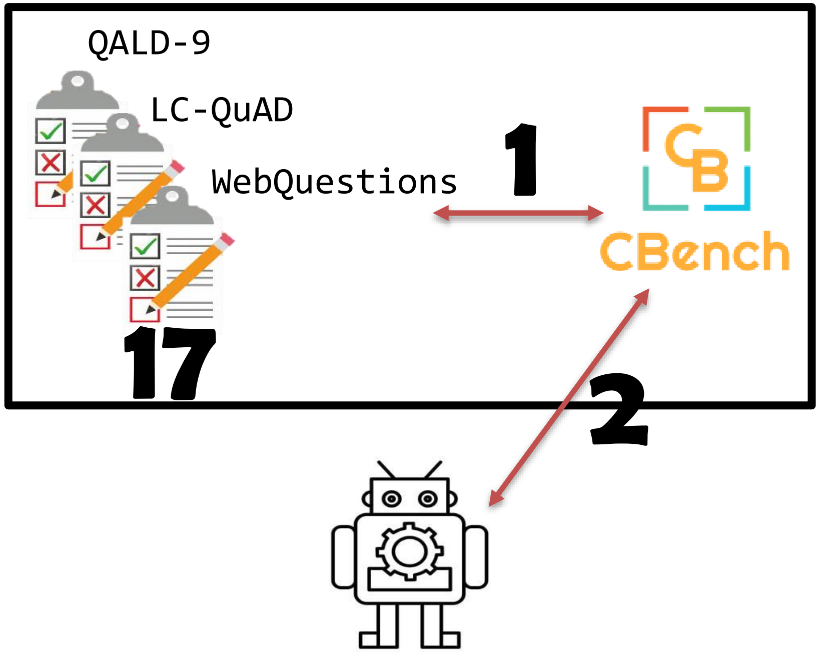
done

Step (1/3)

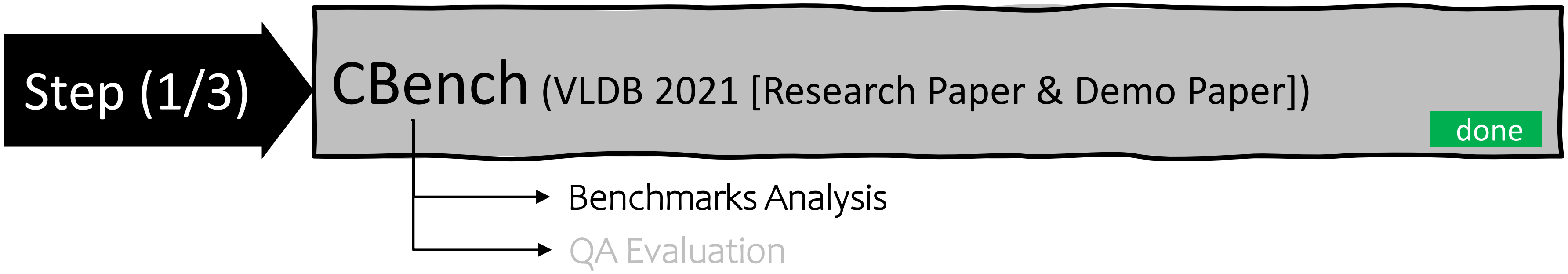
CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation



Benchmarks	#Qs	KG	Version
QALD-1 [40]	199	DB, MB	3.6
QALD-2 [14]	344	DB, MB	3.7
QALD-3 [12]	397	DB, MB	3.8
QALD-4 [41]	321	DB	3.9
QALD-5 [42]	334	DB	2014
QALD-6 [43]	431	DB, LS	10-2015
QALD-7 [46]	530	DB, WD	04-2016
QALD-8 [45]	315	DB, WD	10-2016
QALD-9 [44]	408	DB	10-2016
LC-QuAD [38]	4,998	DB	04-2016
WebQuestions [8]	5,810	FB	09-08-2015
GraphQuestions [35]	5,166	FB	06-2013
SimpleQuestions⋆† [11]	108,442	FB	FB2M, FB5M
SimpleDBpediaQA⋆† [7]	43,086	DB	10-2016
TempQuestions⋆ [26]	1,271	FB	09-08-2015
ComplexQuestions⋆ [4]	150	FB	09-08-2015
ComQA⋆ [3]	11,214	Wikipedia	-



Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

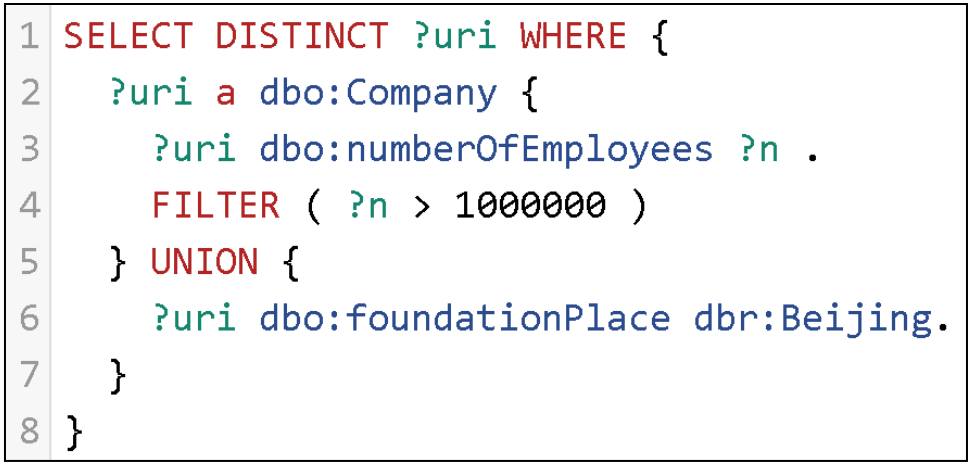
Benchmarks Analysis

QA Evaluation

nlq

Which companies
have more than 1 million employees
or
founded in Beijing?

q



```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

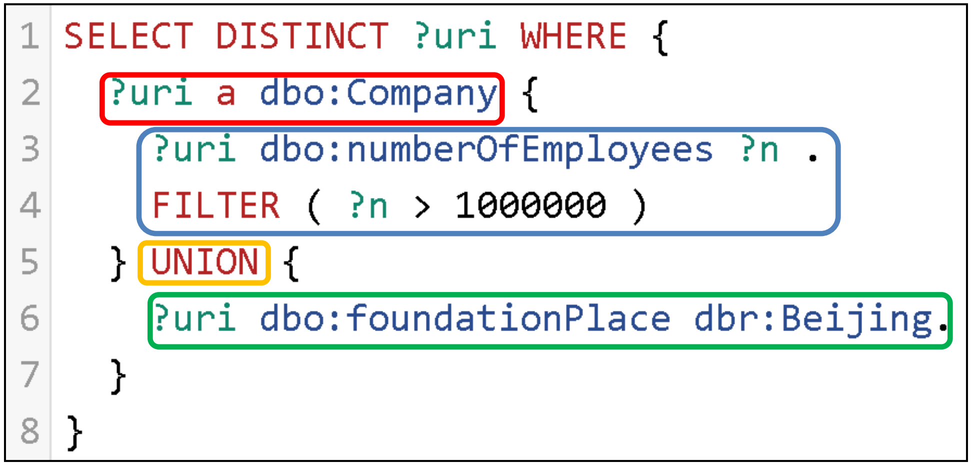
Benchmarks Analysis

QA Evaluation

nlq

Which companies
have more than 1 million employees
or
founded in Beijing?

q



```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```

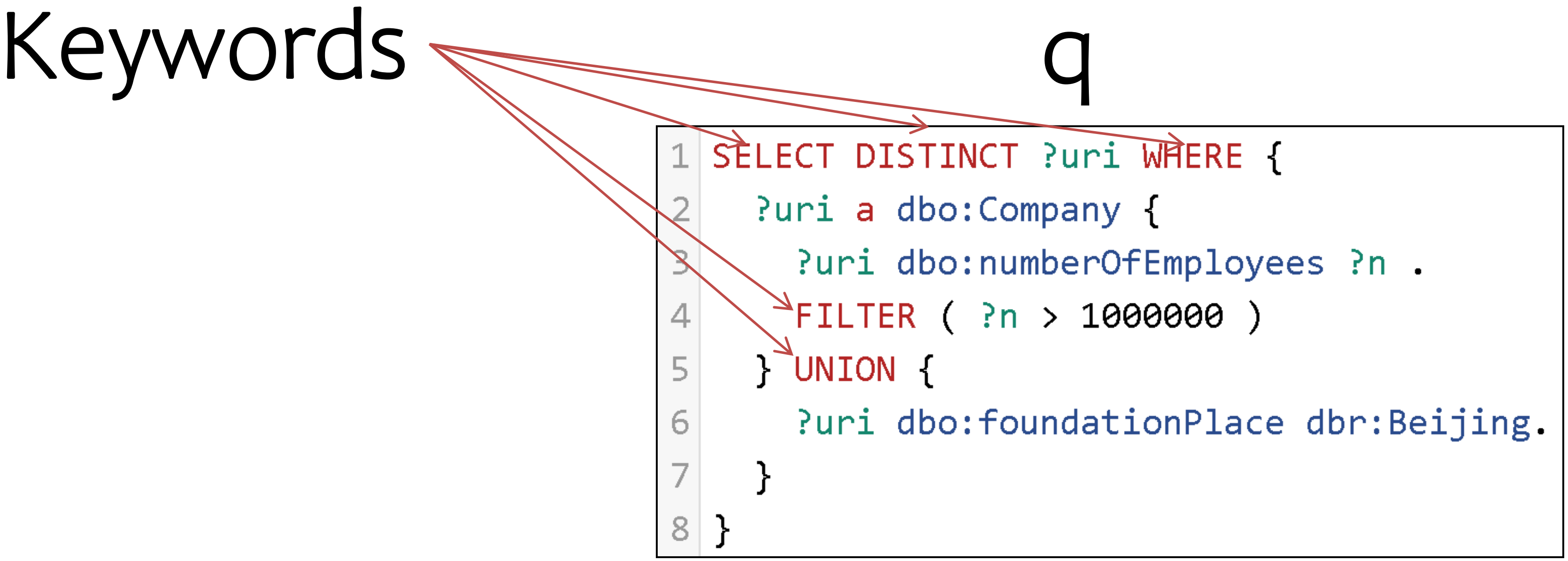
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation



Keywords

q

```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```

Step (1/3)

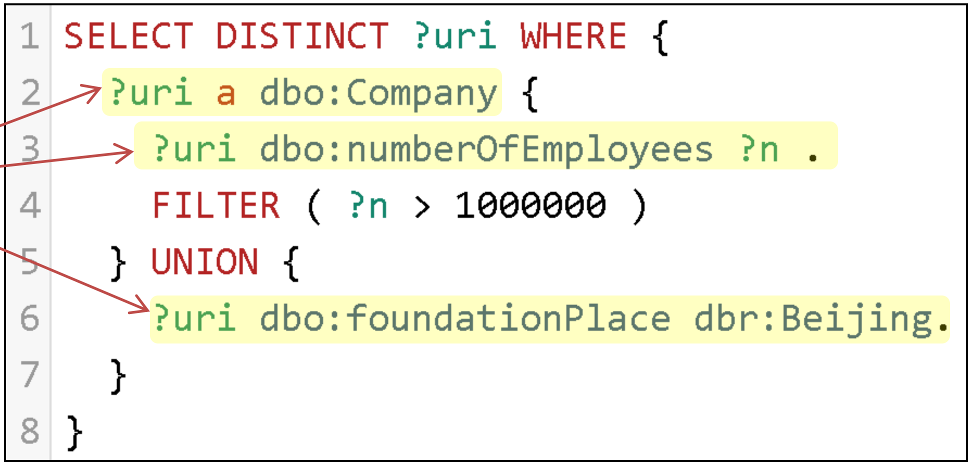
CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation

q

#Triple Patterns



```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```

QAKG Past and Future

Step (1/3)

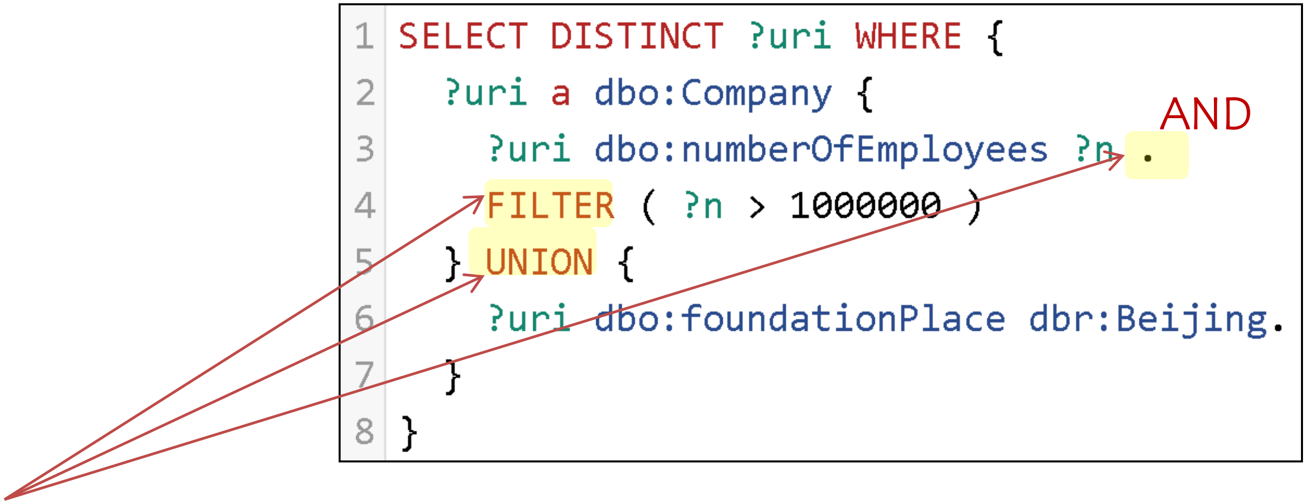
CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation

q



```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```

Operators

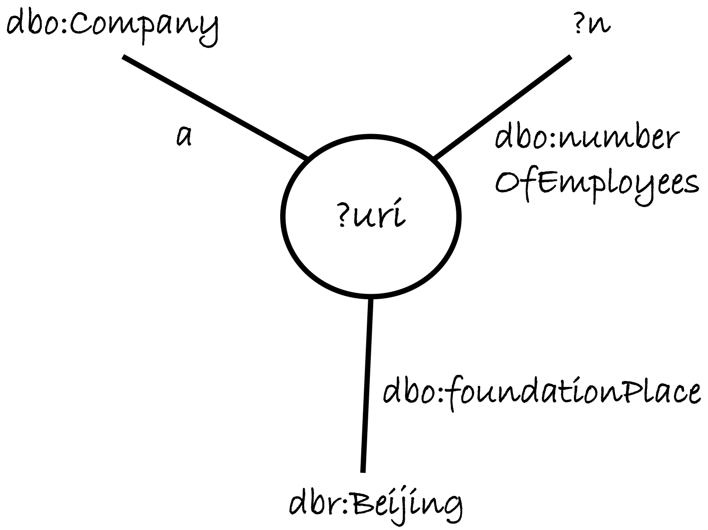
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

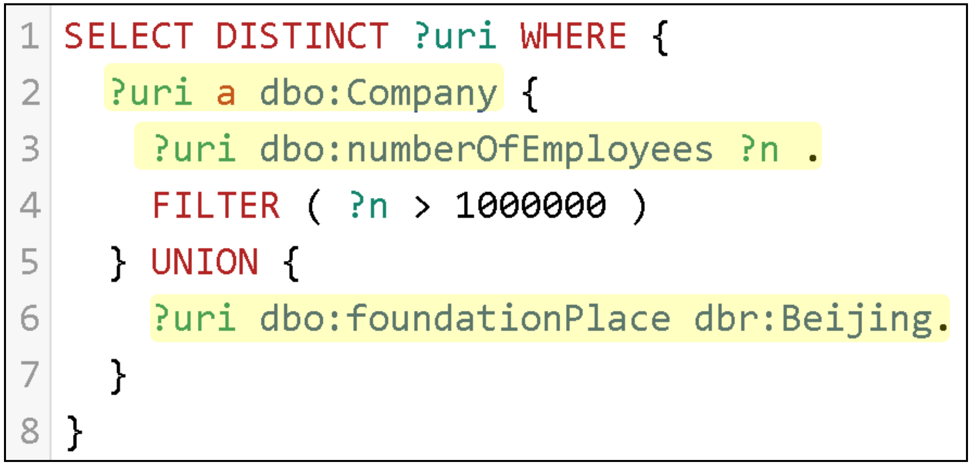
done

- Benchmarks Analysis
- QA Evaluation

Shape



q



```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```

Step (1/3)

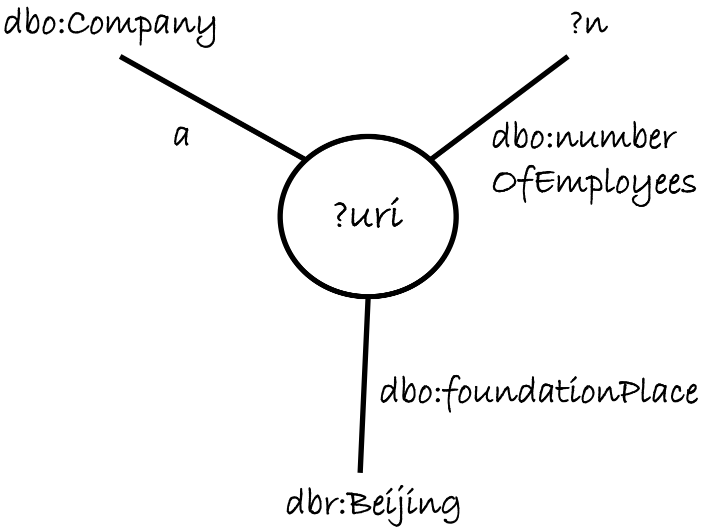
CBench (VLDB 2021 [Research Paper & Demo Paper])

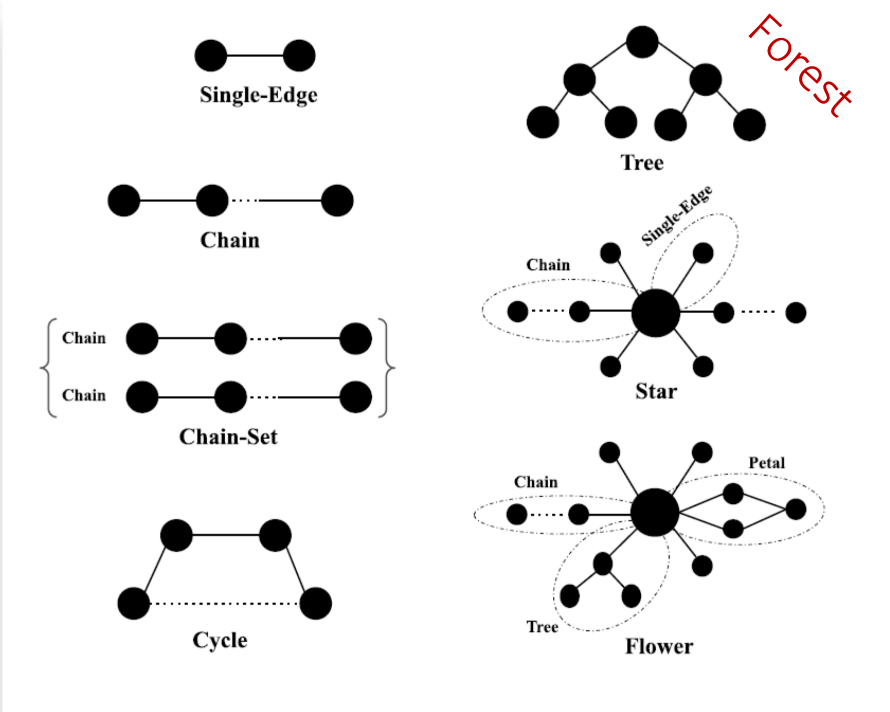
done

Benchmarks Analysis

QA Evaluation

Shape





Forest

The different shapes recognized by CBench

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation

Queries Analysis Results

Keywords

Percentage of keyword occurrences in queries for each benchmark.

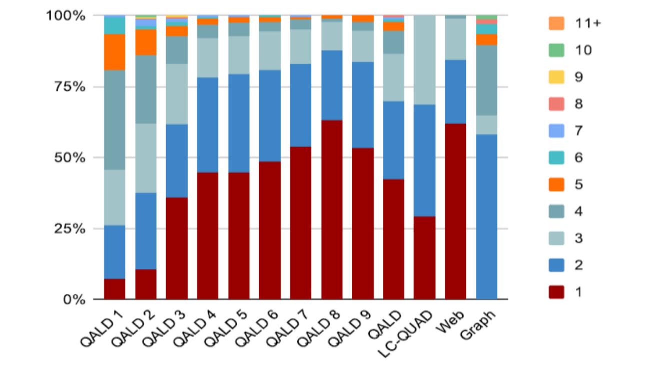
Element	QALD	LC-QuAD	Web	Graph
Select	91.63%	91.52%	100.00%	100.00%
Ask	8.37%	8.48%	0.00%	0.00%
Distinct	76.65%			
Limit	6.51%			
Offset	3.93%			
Order By	5.99%			
And	51.65%			
Filter	10.33%			
Union	6.10%			
Optional	5.37%			
Not Exists	0.21%			
Minus	0.21%			
Aggregators	5.27%			
Group By	5.27%			
Having	1.34%			

Operators

The frequency of the operators used in queries: Filter (F), And (A), Optional (O), and Union (U).

Operators	QALD	LC-QuAD	Web	Graph
none	42.25%	29.33%	0.09%	0.00%
F	0.00%	0.00%	62.19%	58.25%
A	42.87%	70.67%	0.17%	0.00%
A, F	4.65%			
CPF	**89.77%**			
O	0.00%			
O, F	2.58%			
A, O	0.10%			
A, O, F	1.45%			
CPF + O	**+4.13%**			
U	2.48%			
U, F	0.10%			
A, U	1.96%			
A, U, F	0.31%			
CPF + U	**+4.86%**			

#Triple Patterns



Percentage of queries exhibiting different number of triple patterns for each benchmark.

QAKG Past and Future

Towards Next-Generation
Question Answering Over Knowledge Graphs (QAKG) Systems
via Accurate Benchmarking and Large-Scale Training

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation

Queries Analysis Results

Keywords

Percentage of keyword occurrences in queries for each benchmark.

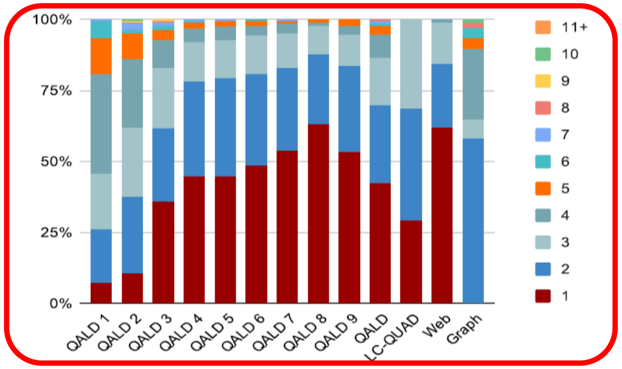
Element	QALD	LC-QuAD	Web	Graph
Select	91.63%	91.52%	100.00%	100.00%
Ask	8.37%	8.48%	0.00%	0.00%
Distinct	76.65%			
Limit	6.51%			
Offset	3.93%			
Order By	5.99%			
And	51.65%			
Filter	10.33%			
Union	6.10%			
Optional	5.37%			
Not Exists	0.21%			
Minus	0.21%			
Aggregators	5.27%			
Group By	5.27%			
Having	1.34%			

Operators

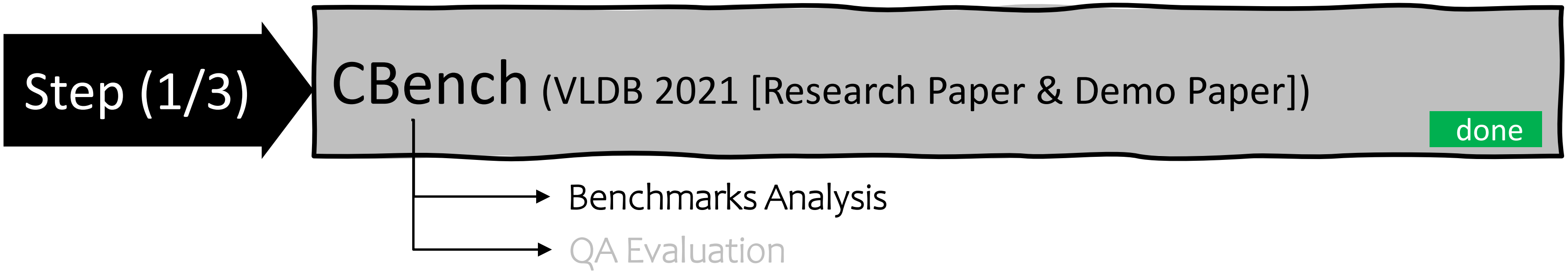
The frequency of the operators used in queries: Filter (F), And (A), Optional (O), and Union (U).

Operators	QALD	LC-QuAD	Web	Graph
none	42.25%	29.33%	0.09%	0.00%
F	0.00%	0.00%	62.19%	58.25%
A	42.87%	70.67%	0.17%	0.00%
A, F	4.65%			
CPF	**89.77%**			
O	0.00%			
O, F	2.58%			
A, O	0.10%			
A, O, F	1.45%			
CPF + O	**+4.13%**			
U	2.48%			
U, F	0.10%			
A, U	1.96%			
A, U, F	0.31%			
CPF + U	**+4.86%**			

#Triple Patterns



Percentage of queries exhibiting different number of triple patterns for each benchmark.



Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

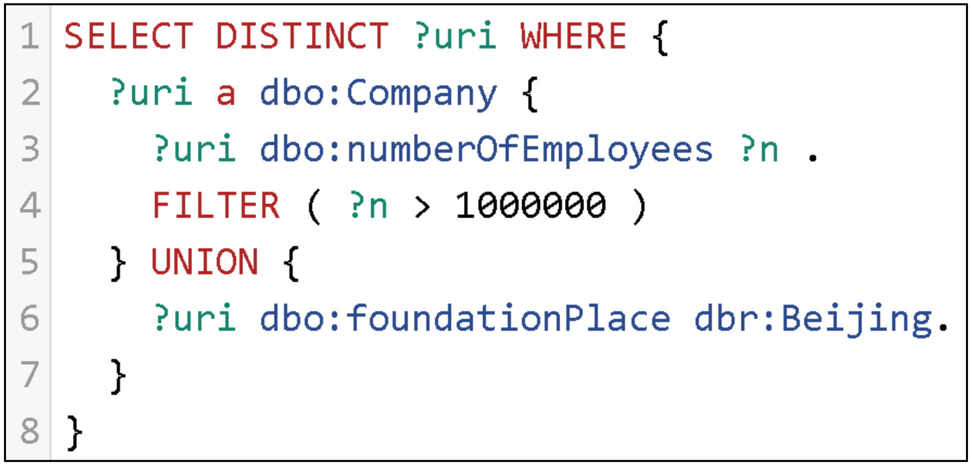
Benchmarks Analysis

QA Evaluation

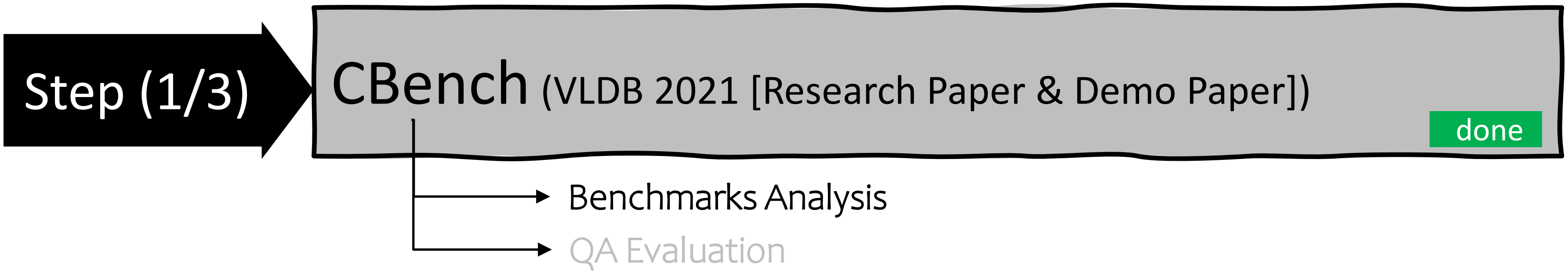
nlq

Which companies
have more than 1 million employees
or
founded in Beijing?

q



```
SELECT DISTINCT ?uri WHERE {
  ?uri a dbo:Company {
    ?uri dbo:numberOfEmployees ?n .
    FILTER ( ?n > 1000000 )
  } UNION {
    ?uri dbo:foundationPlace dbr:Beijing.
  }
}
```



Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation

nlq

Question type

<u>Which</u> companies
have more than 1 million employees
or
founded in Beijing?

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

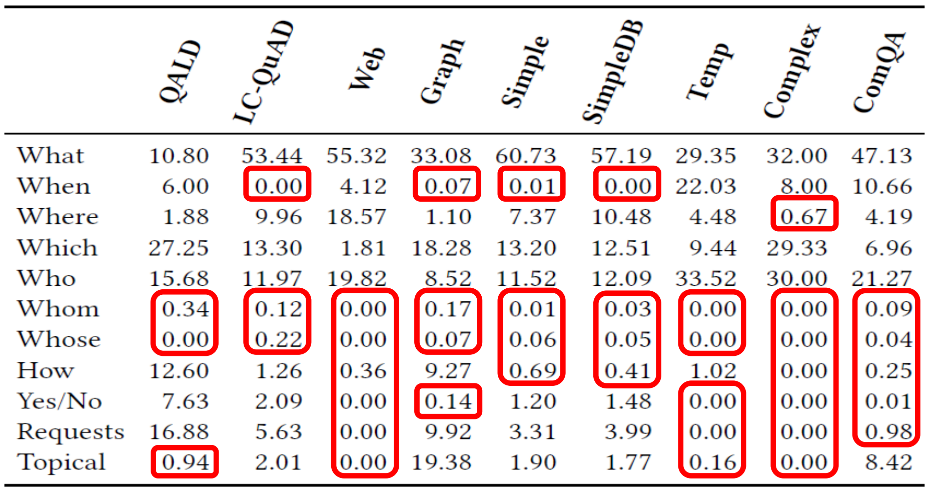
QA Evaluation

nlq

Which companies
have more than 1 million employees
or
founded in Beijing?

Questions Analysis Results

Question frequency percentages (%) by type for all benchmarks.



	QALD	LC-QuAD	Web	Graph	Simple	SimpleDB	Temp	Complex	ComQA
What	10.80	53.44	55.32	33.08	60.73	57.19	29.35	32.00	47.13
When	6.00	0.00	4.12	0.07	0.01	0.00	22.03	8.00	10.66
Where	1.88	9.96	18.57	1.10	7.37	10.48	4.48	0.67	4.19
Which	27.25	13.30	1.81	18.28	13.20	12.51	9.44	29.33	6.96
Who	15.68	11.97	19.82	8.52	11.52	12.09	33.52	30.00	21.27
Whom	0.34	0.12	0.00	0.17	0.01	0.03	0.00	0.00	0.09
Whose	0.00	0.22	0.00	0.07	0.06	0.05	0.00	0.00	0.04
How	12.60	1.26	0.36	9.27	0.69	0.41	1.02	0.00	0.25
Yes/No	7.63	2.09	0.00	0.14	1.20	1.48	0.00	0.00	0.01
Requests	16.88	5.63	0.00	9.92	3.31	3.99	0.00	0.00	0.98
Topical	0.94	2.01	0.00	19.38	1.90	1.77	0.16	0.00	8.42

□ < 1.00%

QAKG Past and Future

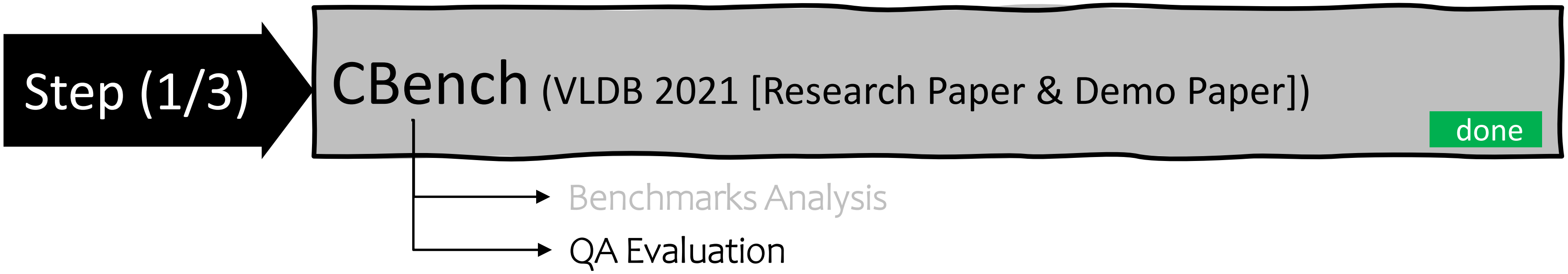
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation



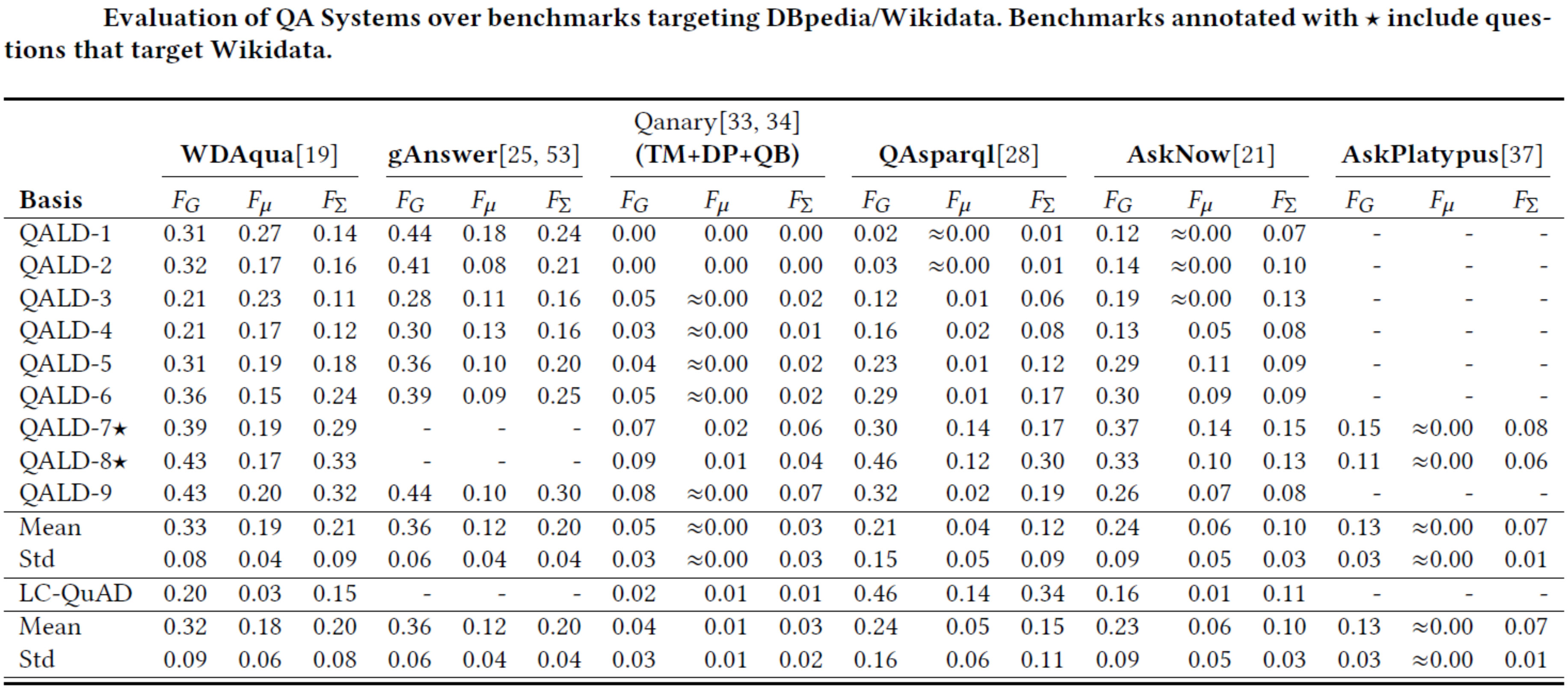
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

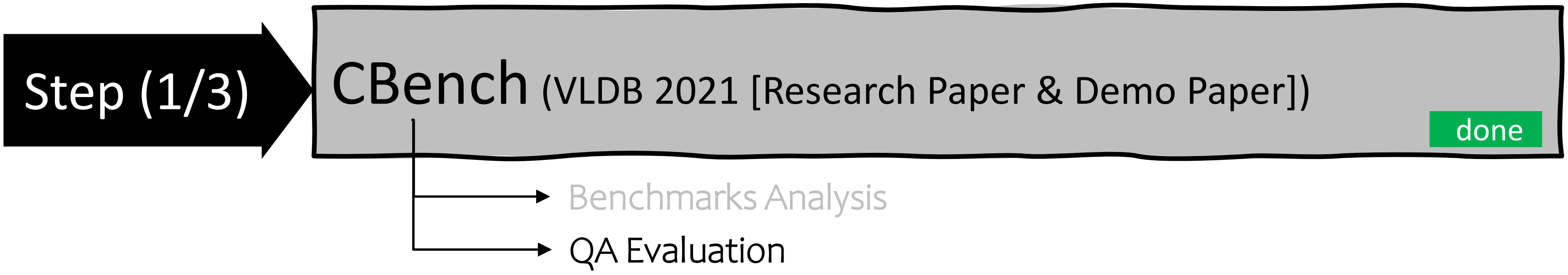
Benchmarks Analysis

QA Evaluation



Evaluation of QA Systems over benchmarks targeting DBpedia/Wikidata. Benchmarks annotated with ⋆ include questions that target Wikidata.

	WDAqua[19]			**gAnswer**[25, 53]			Qanary[33, 34] **(TM+DP+QB)**			**QAsparql**[28]			**AskNow**[21]			**AskPlatypus**[37]		
Basis	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ
QALD-1	0.31	0.27	0.14	0.44	0.18	0.24	0.00	0.00	0.00	0.02	≈0.00	0.01	0.12	≈0.00	0.07	-	-	-
QALD-2	0.32	0.17	0.16	0.41	0.08	0.21	0.00	0.00	0.00	0.03	≈0.00	0.01	0.14	≈0.00	0.10	-	-	-
QALD-3	0.21	0.23	0.11	0.28	0.11	0.16	0.05	≈0.00	0.02	0.12	0.01	0.06	0.19	≈0.00	0.13	-	-	-
QALD-4	0.21	0.17	0.12	0.30	0.13	0.16	0.03	≈0.00	0.01	0.16	0.02	0.08	0.13	0.05	0.08	-	-	-
QALD-5	0.31	0.19	0.18	0.36	0.10	0.20	0.04	≈0.00	0.02	0.23	0.01	0.12	0.29	0.11	0.09	-	-	-
QALD-6	0.36	0.15	0.24	0.39	0.09	0.25	0.05	≈0.00	0.02	0.29	0.01	0.17	0.30	0.09	0.09	-	-	-
QALD-7⋆	0.39	0.19	0.29	-	-	-	0.07	0.02	0.06	0.30	0.14	0.17	0.37	0.14	0.15	0.15	≈0.00	0.08
QALD-8⋆	0.43	0.17	0.33	-	-	-	0.09	0.01	0.04	0.46	0.12	0.30	0.33	0.10	0.13	0.11	≈0.00	0.06
QALD-9	0.43	0.20	0.32	0.44	0.10	0.30	0.08	≈0.00	0.07	0.32	0.02	0.19	0.26	0.07	0.08	-	-	-
Mean	0.33	0.19	0.21	0.36	0.12	0.20	0.05	≈0.00	0.03	0.21	0.04	0.12	0.24	0.06	0.10	0.13	≈0.00	0.07
Std	0.08	0.04	0.09	0.06	0.04	0.04	0.03	≈0.00	0.03	0.15	0.05	0.09	0.09	0.05	0.03	0.03	≈0.00	0.01
LC-QuAD	0.20	0.03	0.15	-	-	-	0.02	0.01	0.01	0.46	0.14	0.34	0.16	0.01	0.11	-	-	-
Mean	0.32	0.18	0.20	0.36	0.12	0.20	0.04	0.01	0.03	0.24	0.05	0.15	0.23	0.06	0.10	0.13	≈0.00	0.07
Std	0.09	0.06	0.08	0.06	0.04	0.04	0.03	0.01	0.02	0.16	0.06	0.11	0.09	0.05	0.03	0.03	≈0.00	0.01



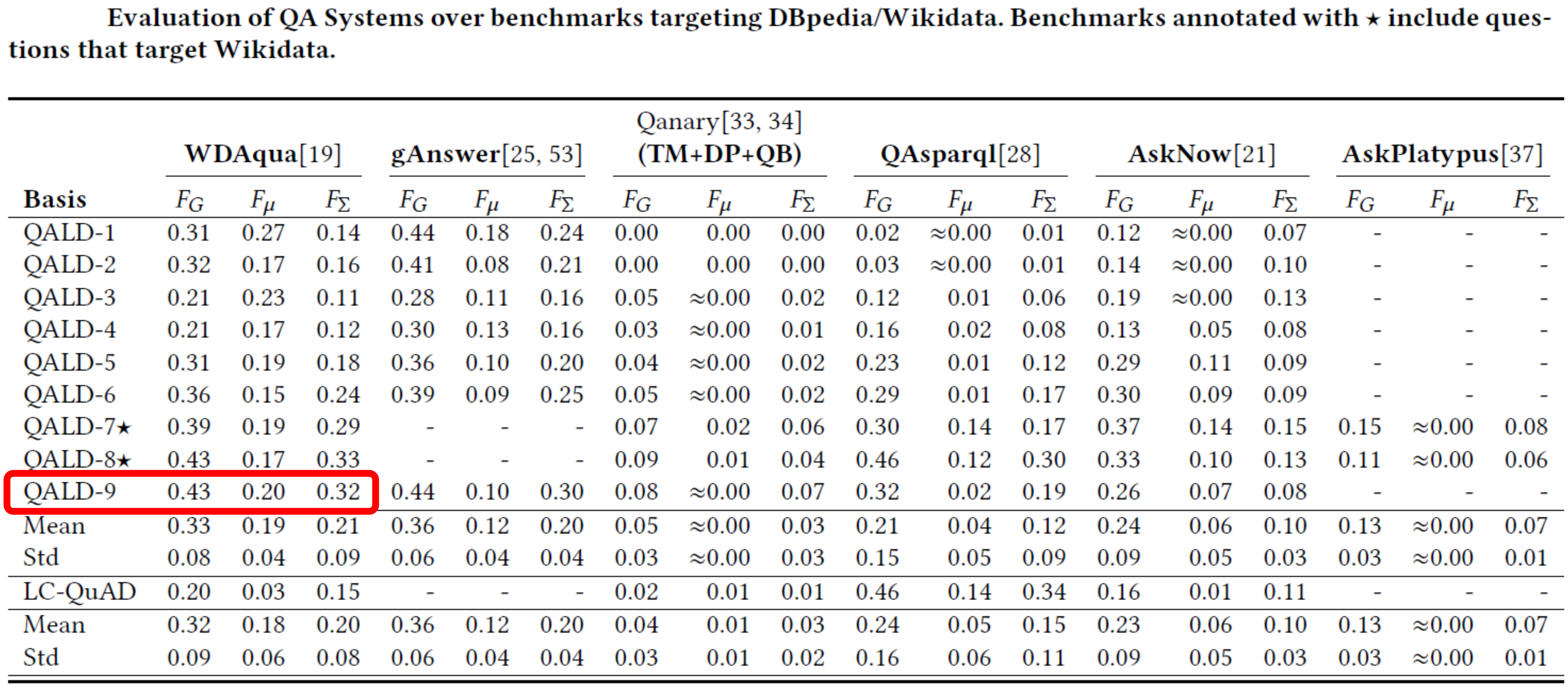
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation



Evaluation of QA Systems over benchmarks targeting DBpedia/Wikidata. Benchmarks annotated with ⋆ include questions that target Wikidata.

	WDAqua[19]			**gAnswer**[25, 53]			Qanary[33, 34] **(TM+DP+QB)**			**QAsparql**[28]			**AskNow**[21]			**AskPlatypus**[37]		
Basis	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ
QALD-1	0.31	0.27	0.14	0.44	0.18	0.24	0.00	0.00	0.00	0.02	≈0.00	0.01	0.12	≈0.00	0.07	-	-	-
QALD-2	0.32	0.17	0.16	0.41	0.08	0.21	0.00	0.00	0.00	0.03	≈0.00	0.01	0.14	≈0.00	0.10	-	-	-
QALD-3	0.21	0.23	0.11	0.28	0.11	0.16	0.05	≈0.00	0.02	0.12	0.01	0.06	0.19	≈0.00	0.13	-	-	-
QALD-4	0.21	0.17	0.12	0.30	0.13	0.16	0.03	≈0.00	0.01	0.16	0.02	0.08	0.13	0.05	0.08	-	-	-
QALD-5	0.31	0.19	0.18	0.36	0.10	0.20	0.04	≈0.00	0.02	0.23	0.01	0.12	0.29	0.11	0.09	-	-	-
QALD-6	0.36	0.15	0.24	0.39	0.09	0.25	0.05	≈0.00	0.02	0.29	0.01	0.17	0.30	0.09	0.09	-	-	-
QALD-7⋆	0.39	0.19	0.29	-	-	-	0.07	0.02	0.06	0.30	0.14	0.17	0.37	0.14	0.15	0.15	≈0.00	0.08
QALD-8⋆	0.43	0.17	0.33	-	-	-	0.09	0.01	0.04	0.46	0.12	0.30	0.33	0.10	0.13	0.11	≈0.00	0.06
QALD-9	0.43	0.20	0.32	0.44	0.10	0.30	0.08	≈0.00	0.07	0.32	0.02	0.19	0.26	0.07	0.08	-	-	-
Mean	0.33	0.19	0.21	0.36	0.12	0.20	0.05	≈0.00	0.03	0.21	0.04	0.12	0.24	0.06	0.10	0.13	≈0.00	0.07
Std	0.08	0.04	0.09	0.06	0.04	0.04	0.03	≈0.00	0.03	0.15	0.05	0.09	0.09	0.05	0.03	0.03	≈0.00	0.01
LC-QuAD	0.20	0.03	0.15	-	-	-	0.02	0.01	0.01	0.46	0.14	0.34	0.16	0.01	0.11	-	-	-
Mean	0.32	0.18	0.20	0.36	0.12	0.20	0.04	0.01	0.03	0.24	0.05	0.15	0.23	0.06	0.10	0.13	≈0.00	0.07
Std	0.09	0.06	0.08	0.06	0.04	0.04	0.03	0.01	0.02	0.16	0.06	0.11	0.09	0.05	0.03	0.03	≈0.00	0.01

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation

Evaluation of QA Systems over benchmarks targeting DBpedia/Wikidata. Benchmarks annotated with ⋆ include questions that target Wikidata.

	WDAqua[19]			**gAnswer**[25, 53]			Qanary[33, 34] **(TM+DP+QB)**			**QAsparql**[28]			**AskNow**[21]			**AskPlatypus**[37]		
Basis	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ
QALD-1	0.31	0.27	0.14	0.44	0.18	0.24	0.00	0.00	0.00	0.02	≈0.00	0.01	0.12	≈0.00	0.07	-	-	-
QALD-2	0.32	0.17	0.16	0.41	0.08	0.21	0.00	0.00	0.00	0.03	≈0.00	0.01	0.14	≈0.00	0.10	-	-	-
QALD-3	0.21	0.23	0.11	0.28	0.11	0.16	0.05	≈0.00	0.02	0.12	0.01	0.06	0.19	≈0.00	0.13	-	-	-
QALD-4	0.21	0.17	0.12	0.30	0.13	0.16	0.03	≈0.00	0.01	0.16	0.02	0.08	0.13	0.05	0.08	-	-	-
QALD-5	0.31	0.19	0.18	0.36	0.10	0.20	0.04	≈0.00	0.02	0.23	0.01	0.12	0.29	0.11	0.09	-	-	-
QALD-6	0.36	0.15	0.24	0.39	0.09	0.25	0.05	≈0.00	0.02	0.29	0.01	0.17	0.30	0.09	0.09	-	-	-
QALD-7⋆	0.39	0.19	0.29	-	-	-	0.07	0.02	0.06	0.30	0.14	0.17	0.37	0.14	0.15	0.15	≈0.00	0.08
QALD-8⋆	0.43	0.17	0.33	-	-	-	0.09	0.01	0.04	0.46	0.12	0.30	0.33	0.10	0.13	0.11	≈0.00	0.06
QALD-9	0.43	0.20	0.32	0.44	0.10	0.30	0.08	≈0.00	0.07	0.32	0.02	0.19	0.26	0.07	0.08	-	-	-
Mean	0.33	0.19	0.21	0.36	0.12	0.20	0.05	≈0.00	0.03	0.21	0.04	0.12	0.24	0.06	0.10	0.13	≈0.00	0.07
Std	0.08	0.04	0.09	0.06	0.04	0.04	0.03	≈0.00	0.03	0.15	0.05	0.09	0.09	0.05	0.03	0.03	≈0.00	0.01
LC-QuAD	0.20	0.03	0.15	-	-	-	0.02	0.01	0.01	0.46	0.14	0.34	0.16	0.01	0.11	-	-	-
Mean	0.32	0.18	0.20	0.36	0.12	0.20	0.04	0.01	0.03	0.24	0.05	0.15	0.23	0.06	0.10	0.13	≈0.00	0.07
Std	0.09	0.06	0.08	0.06	0.04	0.04	0.03	0.01	0.02	0.16	0.06	0.11	0.09	0.05	0.03	0.03	≈0.00	0.01

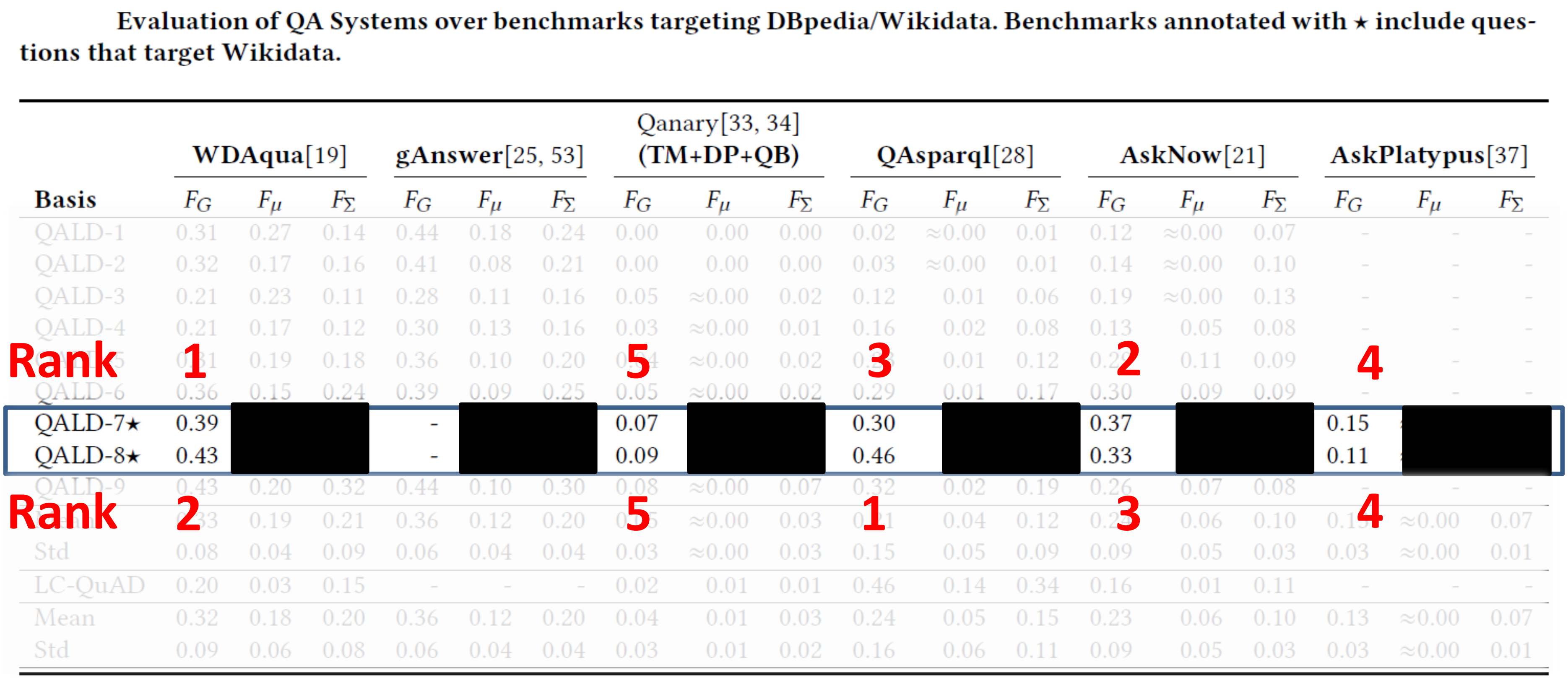
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation



Evaluation of QA Systems over benchmarks targeting DBpedia/Wikidata. Benchmarks annotated with ⋆ include questions that target Wikidata.

	WDAqua[19]			gAnswer[25, 53]			Qanary[33, 34] (TM+DP+QB)			QAsparql[28]			AskNow[21]			AskPlatypus[37]		
Basis	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ
QALD-1	0.31	0.27	0.14	0.44	0.18	0.24	0.00	0.00	0.00	0.02	≈0.00	0.01	0.12	≈0.00	0.07	-	-	-
QALD-2	0.32	0.17	0.16	0.41	0.08	0.21	0.00	0.00	0.00	0.03	≈0.00	0.01	0.14	≈0.00	0.10	-	-	-
QALD-3	0.21	0.23	0.11	0.28	0.11	0.16	0.05	≈0.00	0.02	0.12	0.01	0.06	0.19	≈0.00	0.13	-	-	-
QALD-4	0.21	0.17	0.12	0.30	0.13	0.16	0.03	≈0.00	0.01	0.16	0.02	0.08	0.13	0.05	0.08	-	-	-
QALD-5	[illegible]	0.19	0.18	0.36	0.10	0.20	[illegible]	≈0.00	0.02	[illegible]	0.01	0.12	[illegible]	0.11	0.09	-	-	-
QALD-6	0.36	0.15	0.24	0.39	0.09	0.25	0.05	≈0.00	0.02	0.29	0.01	0.17	0.30	0.09	0.09	-	-	-
QALD-7⋆	0.39	[redacted]	[redacted]	-	[redacted]	[redacted]	0.07	[redacted]	[redacted]	0.30	[redacted]	[redacted]	0.37	[redacted]	[redacted]	0.15	[redacted]	[redacted]
QALD-8⋆	0.43	[redacted]	[redacted]	-	[redacted]	[redacted]	0.09	[redacted]	[redacted]	0.46	[redacted]	[redacted]	0.33	[redacted]	[redacted]	0.11	[redacted]	[redacted]
QALD-9	0.43	0.20	0.32	0.44	0.10	0.30	0.08	≈0.00	0.07	0.32	0.02	0.19	0.26	0.07	0.08	-	-	-
Mean	[illegible]	0.19	0.21	0.36	0.12	0.20	[illegible]	≈0.00	0.03	[illegible]	0.04	0.12	[illegible]	0.06	0.10	[illegible]	≈0.00	0.07
Std	0.08	0.04	0.09	0.06	0.04	0.04	0.03	≈0.00	0.03	0.15	0.05	0.09	0.09	0.05	0.03	0.03	≈0.00	0.01
LC-QuAD	0.20	0.03	0.15	-	-	-	0.02	0.01	0.01	0.46	0.14	0.34	0.16	0.01	0.11	-	-	-
Mean	0.32	0.18	0.20	0.36	0.12	0.20	0.04	0.01	0.03	0.24	0.05	0.15	0.23	0.06	0.10	0.13	≈0.00	0.07
Std	0.09	0.06	0.08	0.06	0.04	0.04	0.03	0.01	0.02	0.16	0.06	0.11	0.09	0.05	0.03	0.03	≈0.00	0.01

Rank 1 5 3 2 4

Rank 2 5 1 3 4

Step (1/3)

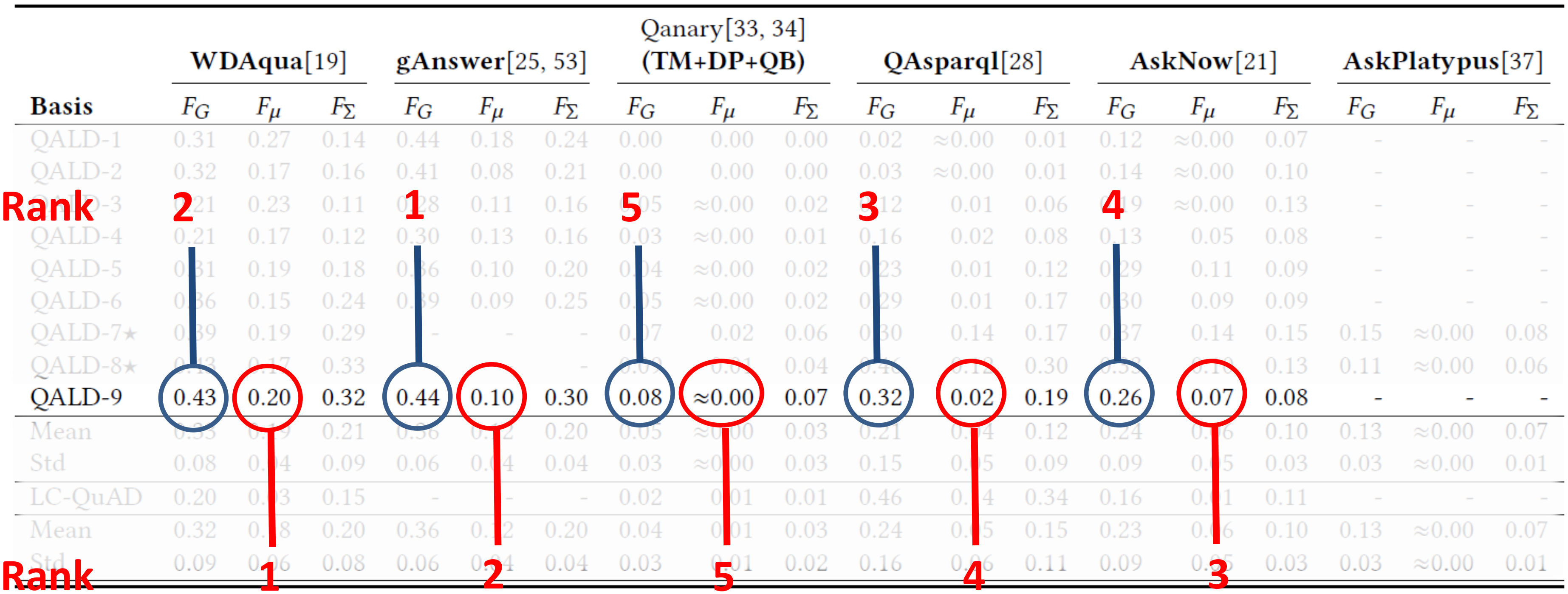
CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation

Evaluation of QA Systems over benchmarks targeting DBpedia/Wikidata. Benchmarks annotated with ★ include questions that target Wikidata.



	WDAqua[19]			gAnswer[25, 53]			Qanary[33, 34] (TM+DP+QB)			QAsparql[28]			AskNow[21]			AskPlatypus[37]		
Basis	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ	F_G	F_μ	F_Σ
QALD-1	0.31	0.27	0.14	0.44	0.18	0.24	0.00	0.00	0.00	0.02	≈0.00	0.01	0.12	≈0.00	0.07	-	-	-
QALD-2	0.32	0.17	0.16	0.41	0.08	0.21	0.00	0.00	0.00	0.03	≈0.00	0.01	0.14	≈0.00	0.10	-	-	-
QALD-3	0.21	0.23	0.11	0.28	0.11	0.16	0.05	≈0.00	0.02	0.12	0.01	0.06	0.19	≈0.00	0.13	-	-	-
QALD-4	0.21	0.17	0.12	0.30	0.13	0.16	0.03	≈0.00	0.01	0.16	0.02	0.08	0.13	0.05	0.08	-	-	-
QALD-5	0.31	0.19	0.18	0.36	0.10	0.20	0.04	≈0.00	0.02	0.23	0.01	0.12	0.29	0.11	0.09	-	-	-
QALD-6	0.36	0.15	0.24	0.39	0.09	0.25	0.05	≈0.00	0.02	0.29	0.01	0.17	0.30	0.09	0.09	-	-	-
QALD-7★	0.39	0.19	0.29	-	-	-	0.07	0.02	0.06	0.30	0.14	0.17	0.37	0.14	0.15	0.15	≈0.00	0.08
QALD-8★	0.43	0.17	0.33	[illegible]	[illegible]	-	0.09	0.01	0.04	0.46	0.12	0.30	0.33	0.10	0.13	0.11	≈0.00	0.06
QALD-9	0.43	0.20	0.32	0.44	0.10	0.30	0.08	≈0.00	0.07	0.32	0.02	0.19	0.26	0.07	0.08	-	-	-
Mean	0.33	0.19	0.21	0.36	0.12	0.20	0.05	≈0.00	0.03	0.21	0.04	0.12	0.24	0.06	0.10	0.13	≈0.00	0.07
Std	0.08	0.04	0.09	0.06	0.04	0.04	0.03	≈0.00	0.03	0.15	0.05	0.09	0.09	0.05	0.03	0.03	≈0.00	0.01
LC-QuAD	0.20	0.03	0.15	-	-	-	0.02	0.01	0.01	0.46	0.14	0.34	0.16	0.01	0.11	-	-	-
Mean	0.32	0.18	0.20	0.36	0.12	0.20	0.04	0.01	0.03	0.24	0.05	0.15	0.23	0.06	0.10	0.13	≈0.00	0.07
Std	0.09	0.06	0.08	0.06	0.04	0.04	0.03	0.01	0.02	0.16	0.06	0.11	0.09	0.05	0.03	0.03	≈0.00	0.01

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Conclusion

There are **high degree of variations** between available benchmarks.

The variation affects the measured **Quality Score** of the QA systems.

We need a **comprehensive benchmark**.

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

CBench
Source code

SCAN ME

QAKG Past and Future

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark**.

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark.**

Manually generating comprehensive benchmarks?

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark.**

Manually generating comprehensive benchmarks?

KG always updatable

There are many KGs.

There are many features to cover.

I cannot

I cannot

I cannot

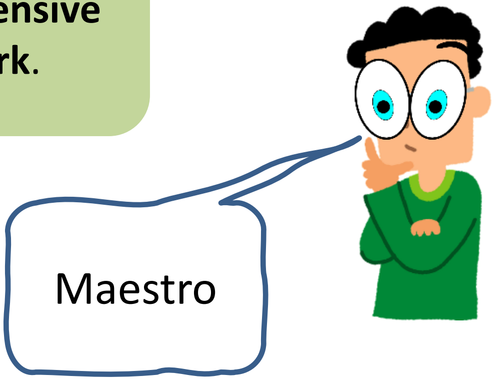
I cannot

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark.**



~~Manually~~
Automatically
generating
comprehensive
benchmarks?

Step (2/3)

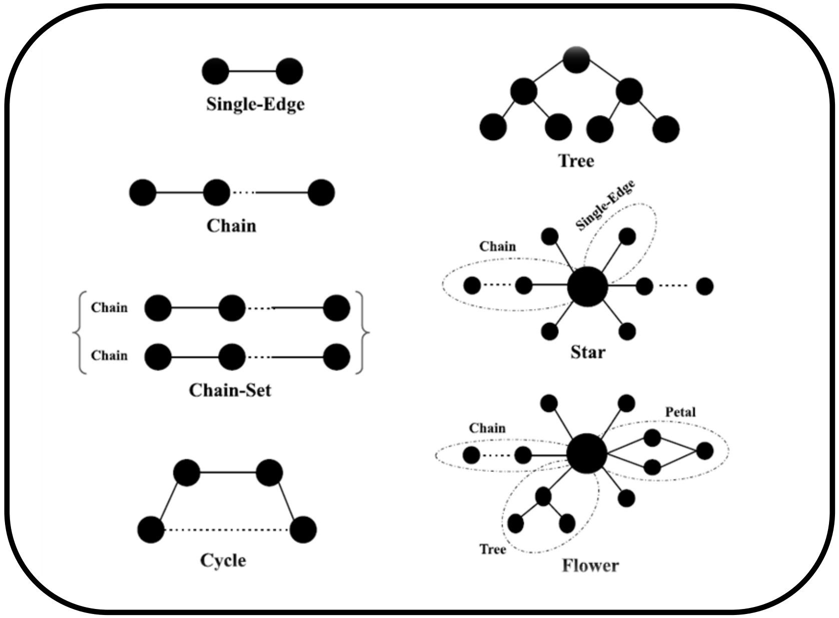
Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

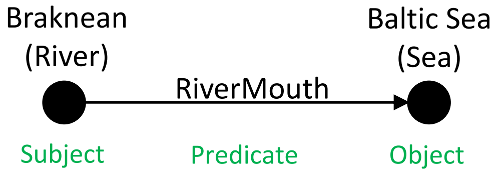
We need a **comprehensive benchmark**.

Maestro is based on two main ideas

There is a limited set of query shapes in KGs



The predicate can be represented by 4 different ways



Predicate Representations

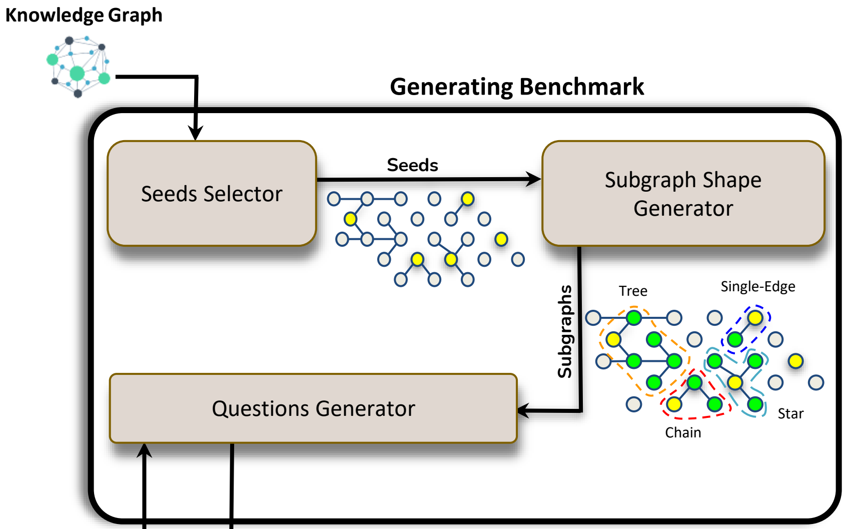
- $(\overrightarrow{P}_{NP})$:- **Braknean** is the tributary of **Baltic Sea**.
- $(\overrightarrow{P}_{VP})$:- **Braknean** flows into **Baltic Sea**.
- $(\overleftarrow{P}_{NP})$:- **Baltic Sea** is the river mouth of **Braknean**.
- $(\overleftarrow{P}_{VP})$:- **Baltic Sea** receives flow from **Braknean**.

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark**.



Predicates Lexicons

Predicate	Context	$\overrightarrow{P}_{NP}$	$\overrightarrow{P}_{VP}$	$\overleftarrow{P}_{NP}$	$\overleftarrow{P}_{VP}$
riverMouth	(River, Sea)	tributary	flows into	river mouth	receives flow from
…	…	…	…	…	…

Benchmark

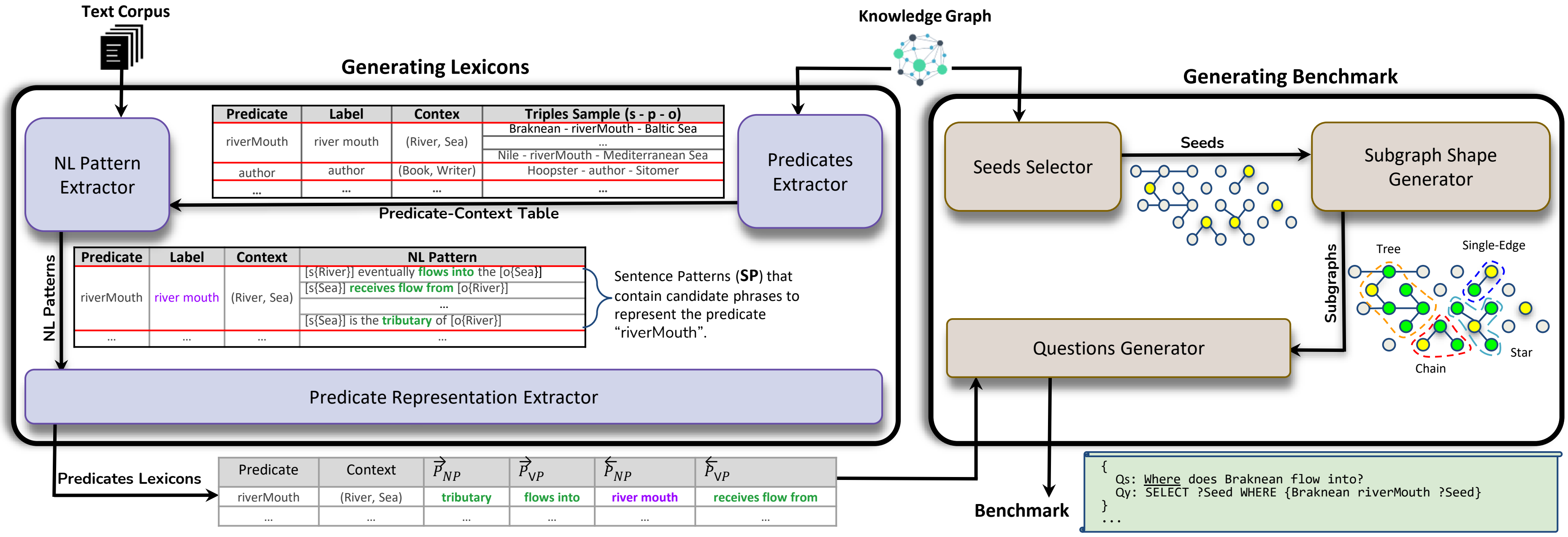
```
{
  Qs: Where does Braknean flow into?
  Qy: SELECT ?Seed WHERE {Braknean riverMouth ?Seed}
}
...
```

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark.**



Predicate	Label	Contex	Triples Sample (s - p - o)
riverMouth	river mouth	(River, Sea)	Braknean - riverMouth - Baltic Sea … Nile - riverMouth - Mediterranean Sea
author	author	(Book, Writer)	Hoopster - author - Sitomer
…	…	…	…

Predicate	Label	Context	NL Pattern
riverMouth	river mouth	(River, Sea)	[s{River}] eventually **flows into** the [o{Sea}] [s{Sea}] **receives flow from** [o{River}] … [s{Sea}] is the **tributary** of [o{River}]
…	…	…	…

Sentence Patterns (**SP**) that contain candidate phrases to represent the predicate "riverMouth".

Predicate	Context	$\overrightarrow{P}_{NP}$	$\overrightarrow{P}_{VP}$	$\overleftarrow{P}_{NP}$	$\overleftarrow{P}_{VP}$
riverMouth	(River, Sea)	**tributary**	**flows into**	**river mouth**	**receives flow from**
…	…	…	…	…	…

```
{
  Qs: Where does Braknean flow into?
  Qy: SELECT ?Seed WHERE {Braknean riverMouth ?Seed}
}
...
```

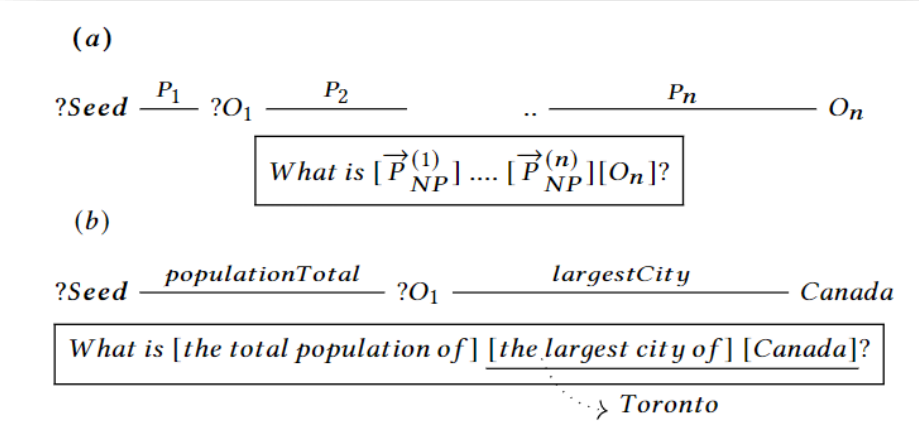
Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark**.

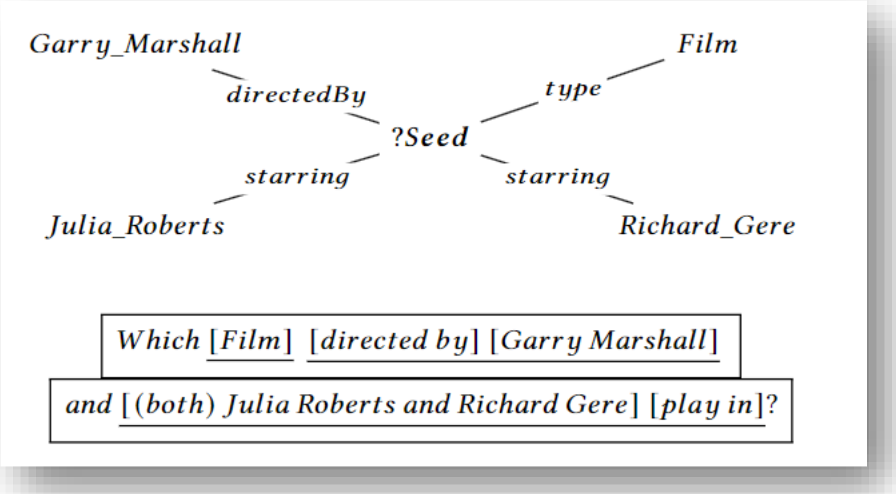
Question Example



What is $[\overrightarrow{P}^{(1)}_{NP}]$ .... $[\overrightarrow{P}^{(n)}_{NP}][O_n]$?

What is [the total population of] [the largest city of] [Canada]?

Chain Question



Which [Film] [directed by] [Garry Marshall]

and [(both) Julia Roberts and Richard Gere] [play in]?

Star Question

Step (2/3)

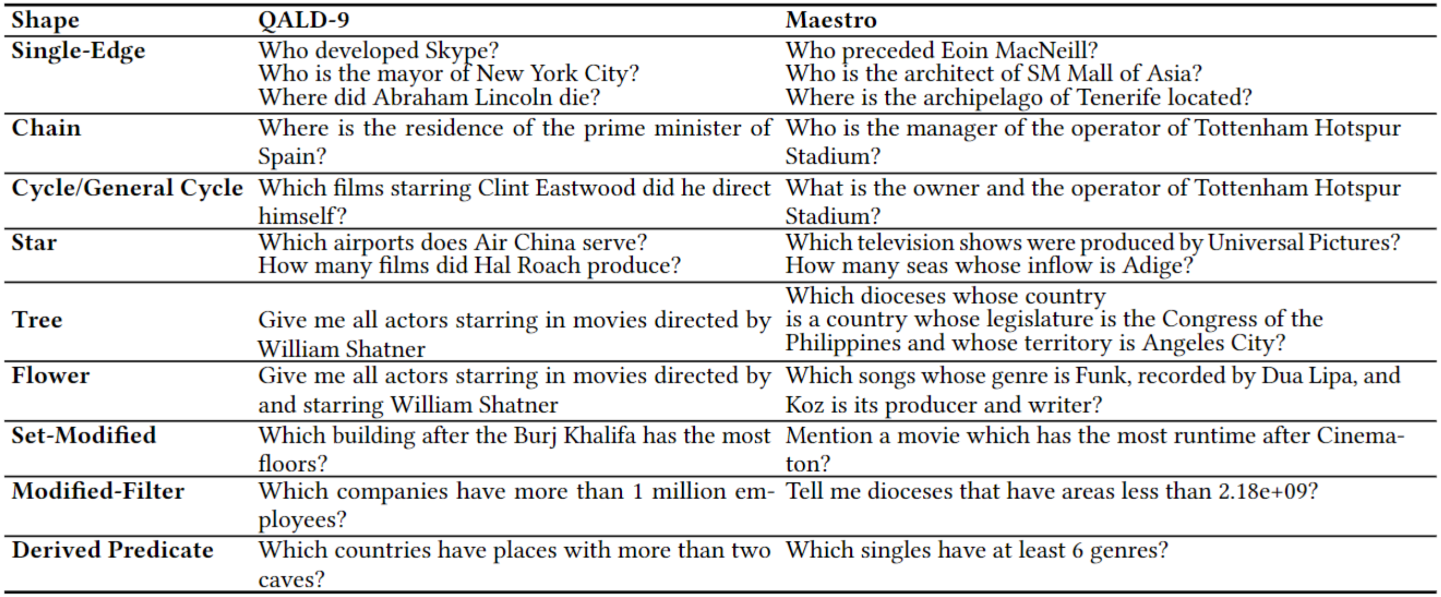
Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

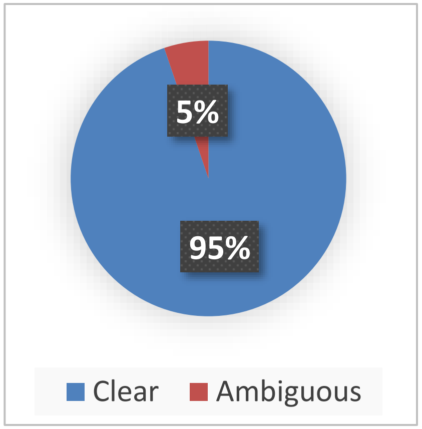
We need a **comprehensive benchmark.**

Evaluation: Correctness

Questions generated by Maestro compared to the QALD-9 questions that follow the same subgraph shape.



Shape	QALD-9	Maestro
Single-Edge	Who developed Skype? Who is the mayor of New York City? Where did Abraham Lincoln die?	Who preceded Eoin MacNeill? Who is the architect of SM Mall of Asia? Where is the archipelago of Tenerife located?
Chain	Where is the residence of the prime minister of Spain?	Who is the manager of the operator of Tottenham Hotspur Stadium?
Cycle/General Cycle	Which films starring Clint Eastwood did he direct himself?	What is the owner and the operator of Tottenham Hotspur Stadium?
Star	Which airports does Air China serve? How many films did Hal Roach produce?	Which television shows were produced by Universal Pictures? How many seas whose inflow is Adige?
Tree	Give me all actors starring in movies directed by William Shatner	Which dioceses whose country is a country whose legislature is the Congress of the Philippines and whose territory is Angeles City?
Flower	Give me all actors starring in movies directed by and starring William Shatner	Which songs whose genre is Funk, recorded by Dua Lipa, and Koz is its producer and writer?
Set-Modified	Which building after the Burj Khalifa has the most floors?	Mention a movie which has the most runtime after Cinematon?
Modified-Filter	Which companies have more than 1 million employees?	Tell me dioceses that have areas less than 2.18e+09?
Derived Predicate	Which countries have places with more than two caves?	Which singles have at least 6 genres?



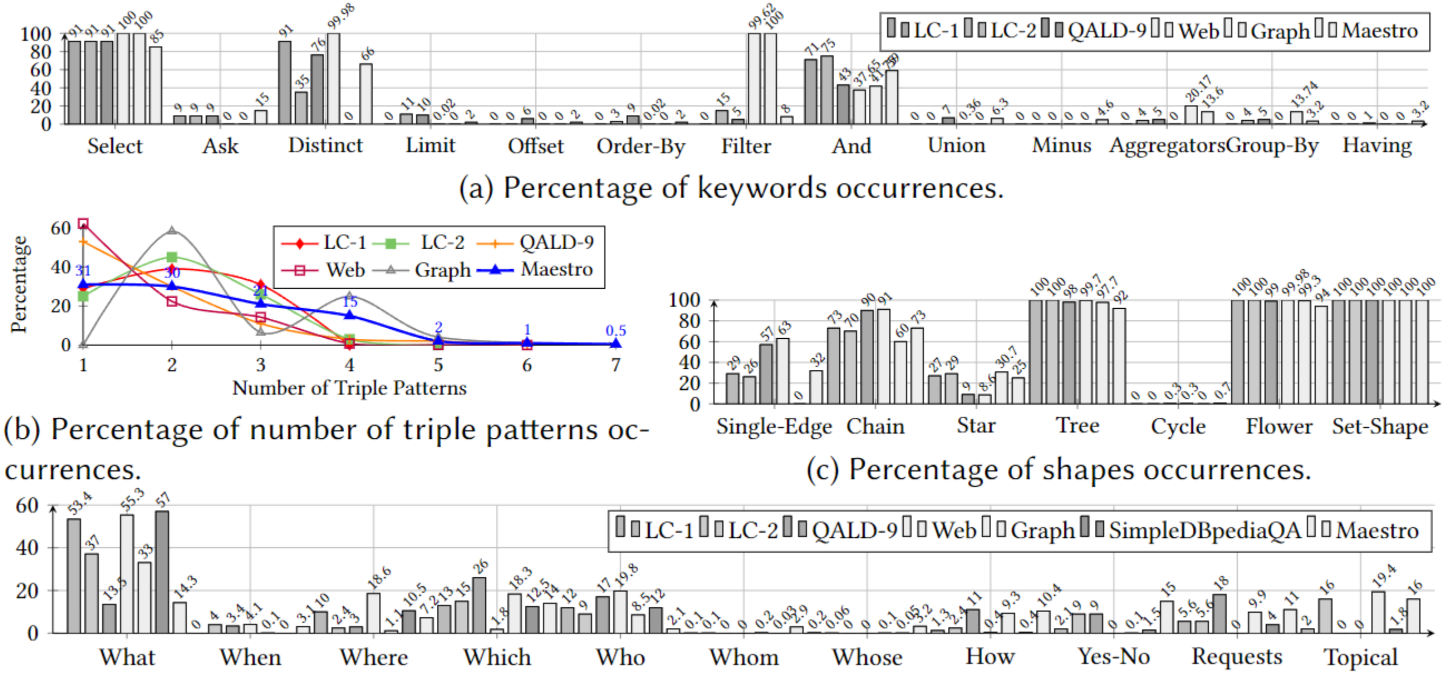
Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark**.

Evaluation: Comprehensiveness



(a) Percentage of keywords occurrences.

(b) Percentage of number of triple patterns occurrences.

(c) Percentage of shapes occurrences.

(d) Percentage of the occurrences of question types.

The coverage of the query properties of Maestro's generated benchmark vs. other benchmarks in the literature.

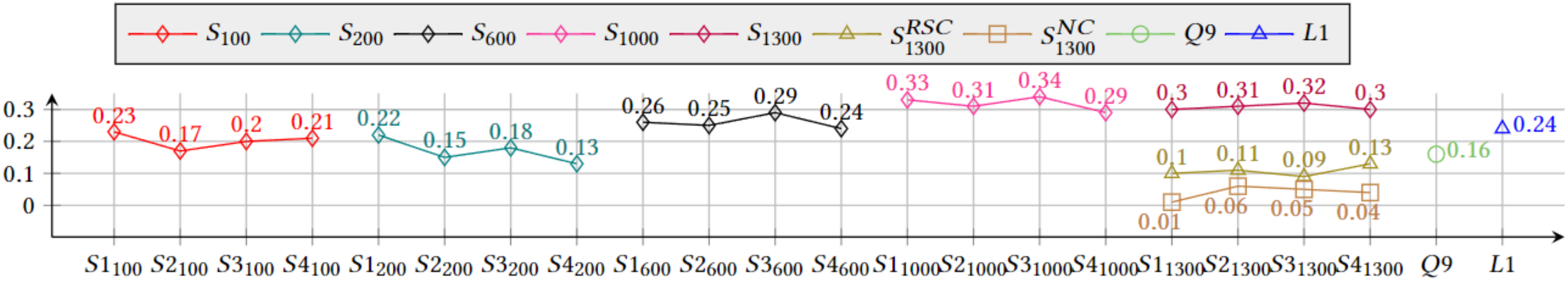
Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

We need a **comprehensive benchmark**.

Evaluation: Consistency



QA evaluation on multiple benchmarks generated by Maestro with different numbers of questions and comparing them to QALD-9 (Q) and LCQuAD-1 (L1) (test files only).

QAKG Past and Future

Step (2/3)

Maestro (VLDB 2022 [Demo Paper] & ACM SIGMOD 2023 [Research Paper])

done

Maestro
Source code

SCAN ME

QAKG Past and Future

Step (3/3)

Dataset

InProgress

We need a **very large dataset**

Step (3/3)

Dataset

InProgress

We need a **very large dataset**

Main idea

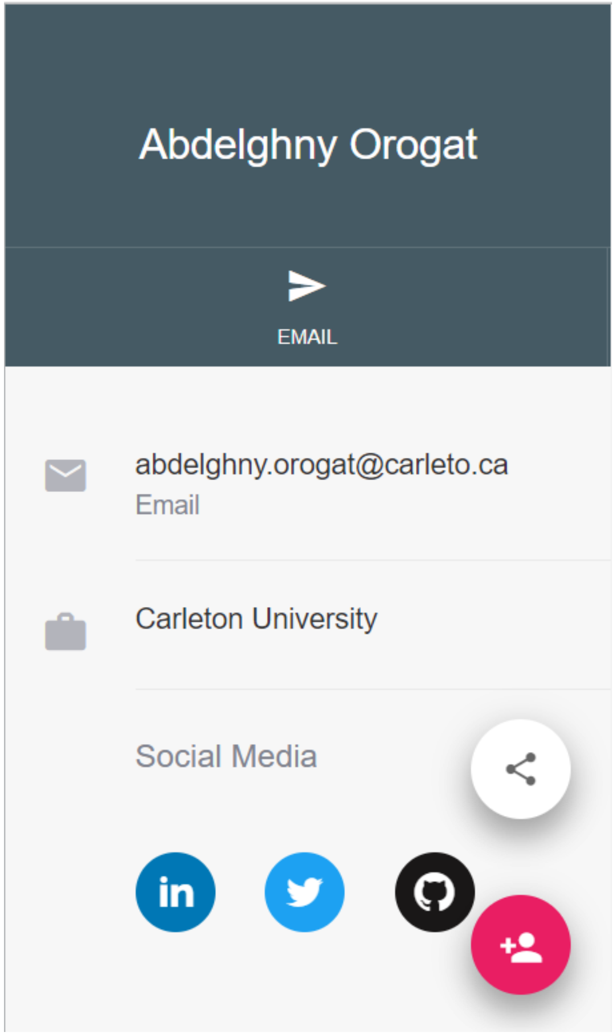
Annotate the questions while constructing them using Maestro

Question

Give me the institution of the scientist ’Lane P. Hughston’ ?

Annotated Question

<qt>Give me</qt> <p>the institution of</p> <o>the scientist ’Lane P. Hughston’</o>?



Abdelghny Orogat

abdelghny.orogat@carleto.ca
Email

Carleton University

Social Media

Thank You

Appendix

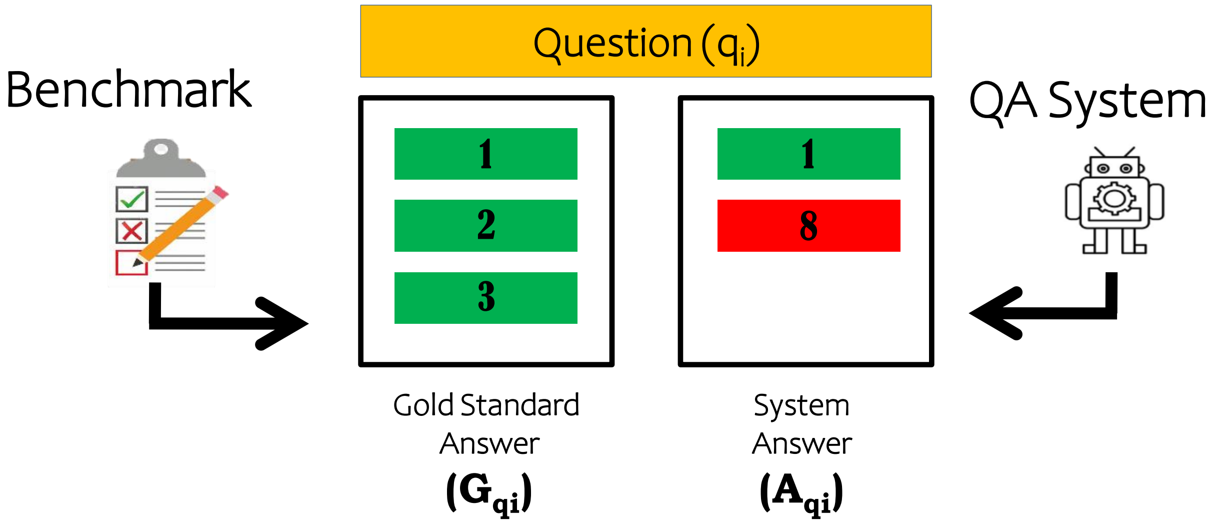
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation

Evaluation Metrics



Question (q_i)

Benchmark

QA System

Gold Standard Answer (G_{qi})

System Answer (A_{qi})

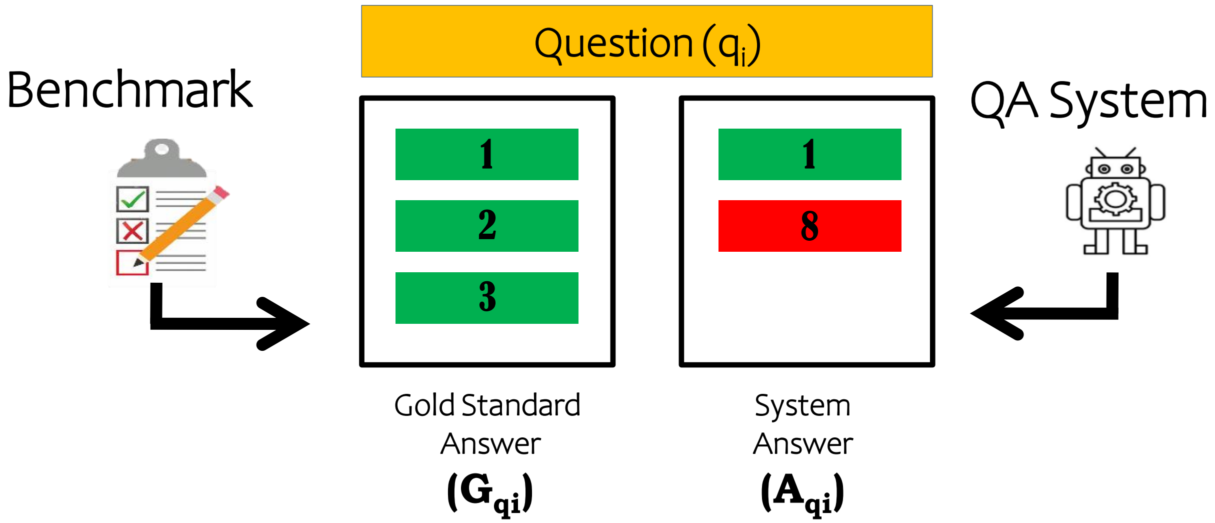
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation

Evaluation Metrics



Benchmark

Question (q_i)

QA System

Gold Standard Answer (G_{qi})

System Answer (A_{qi})

$$R_{q_i} = \frac{|G_{q_i} \cap A_{q_i}|}{|G_{q_i}|} = \frac{\left|\begin{array}{c}1\end{array}\right|}{\left|\begin{array}{c}1\\2\\3\end{array}\right|} = 0.33$$

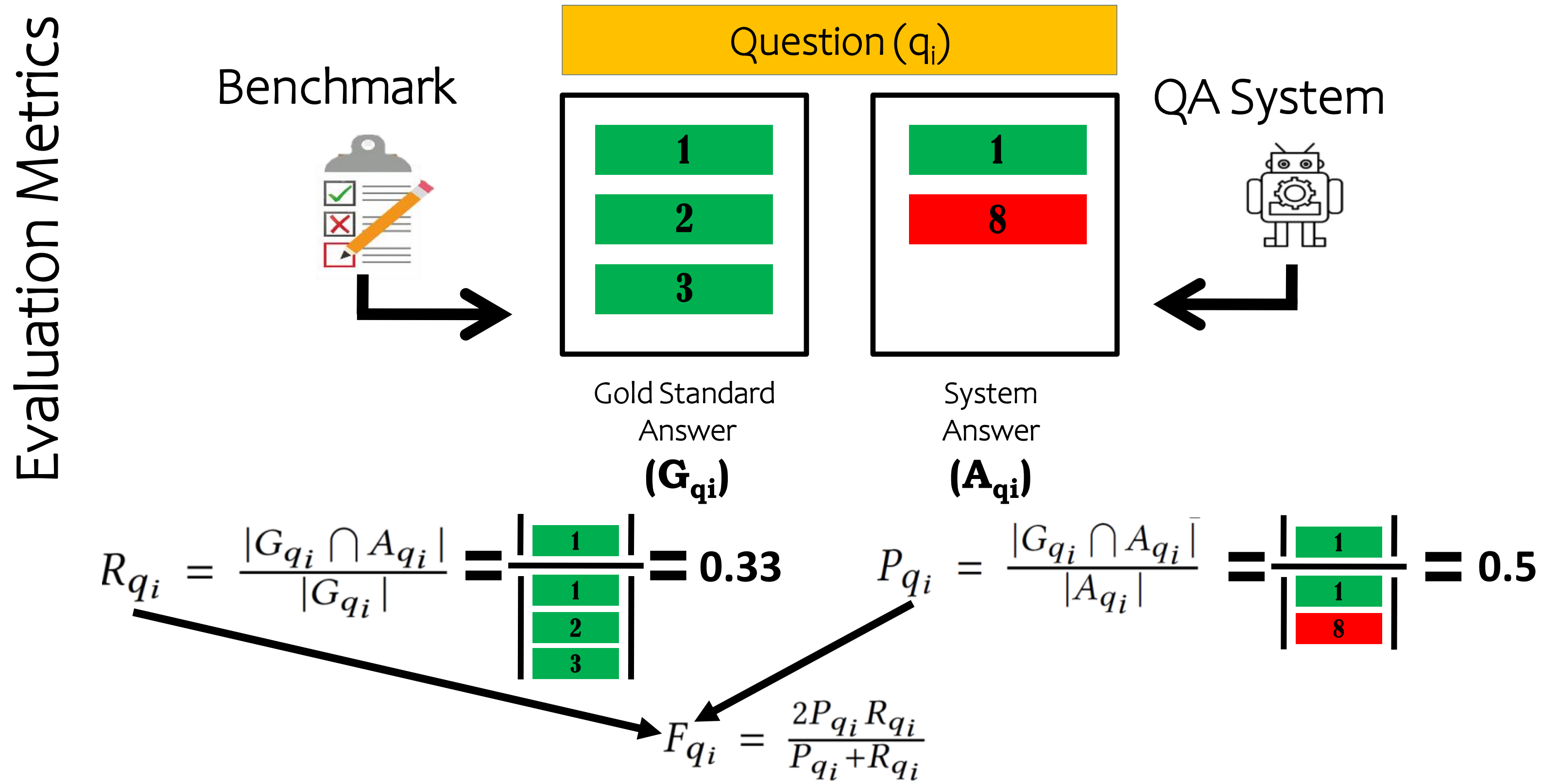
$$P_{q_i} = \frac{|G_{q_i} \cap A_{q_i}|}{|A_{q_i}|} = \frac{\left|\begin{array}{c}1\end{array}\right|}{\left|\begin{array}{c}1\\8\end{array}\right|} = 0.5$$

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation



$$R_{q_i} = \frac{|G_{q_i} \cap A_{q_i}|}{|G_{q_i}|} = \frac{|\,1\,|}{|\,1\ 2\ 3\,|} = 0.33$$

$$P_{q_i} = \frac{|G_{q_i} \cap A_{q_i}|}{|A_{q_i}|} = \frac{|\,1\,|}{|\,1\ 8\,|} = 0.5$$

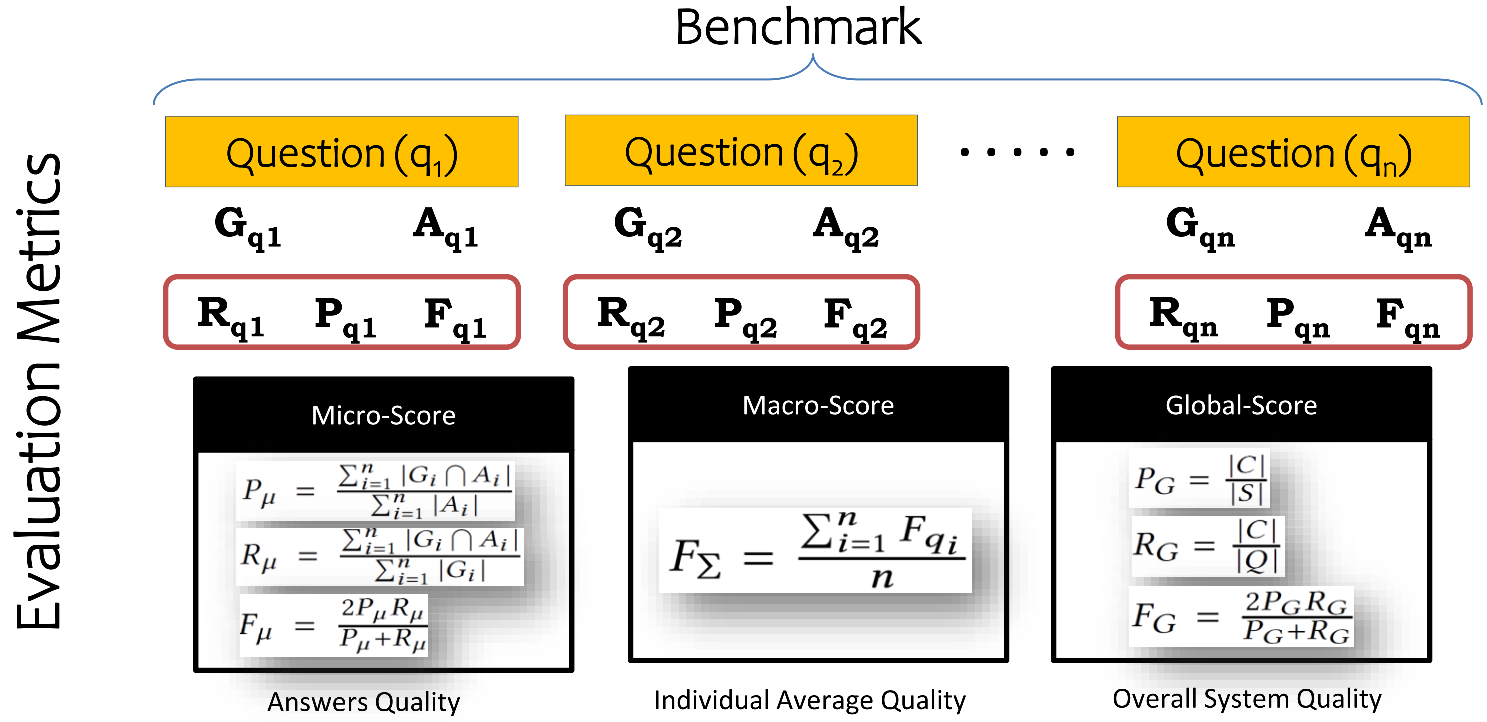
$$F_{q_i} = \frac{2P_{q_i}R_{q_i}}{P_{q_i}+R_{q_i}}$$

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation



Micro-Score

$$P_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |A_i|}$$

$$R_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |G_i|}$$

$$F_\mu = \frac{2P_\mu R_\mu}{P_\mu + R_\mu}$$

Answers Quality

Macro-Score

$$F_\Sigma = \frac{\sum_{i=1}^{n} F_{q_i}}{n}$$

Individual Average Quality

Global-Score

$$P_G = \frac{|C|}{|S|}$$

$$R_G = \frac{|C|}{|Q|}$$

$$F_G = \frac{2P_G R_G}{P_G + R_G}$$

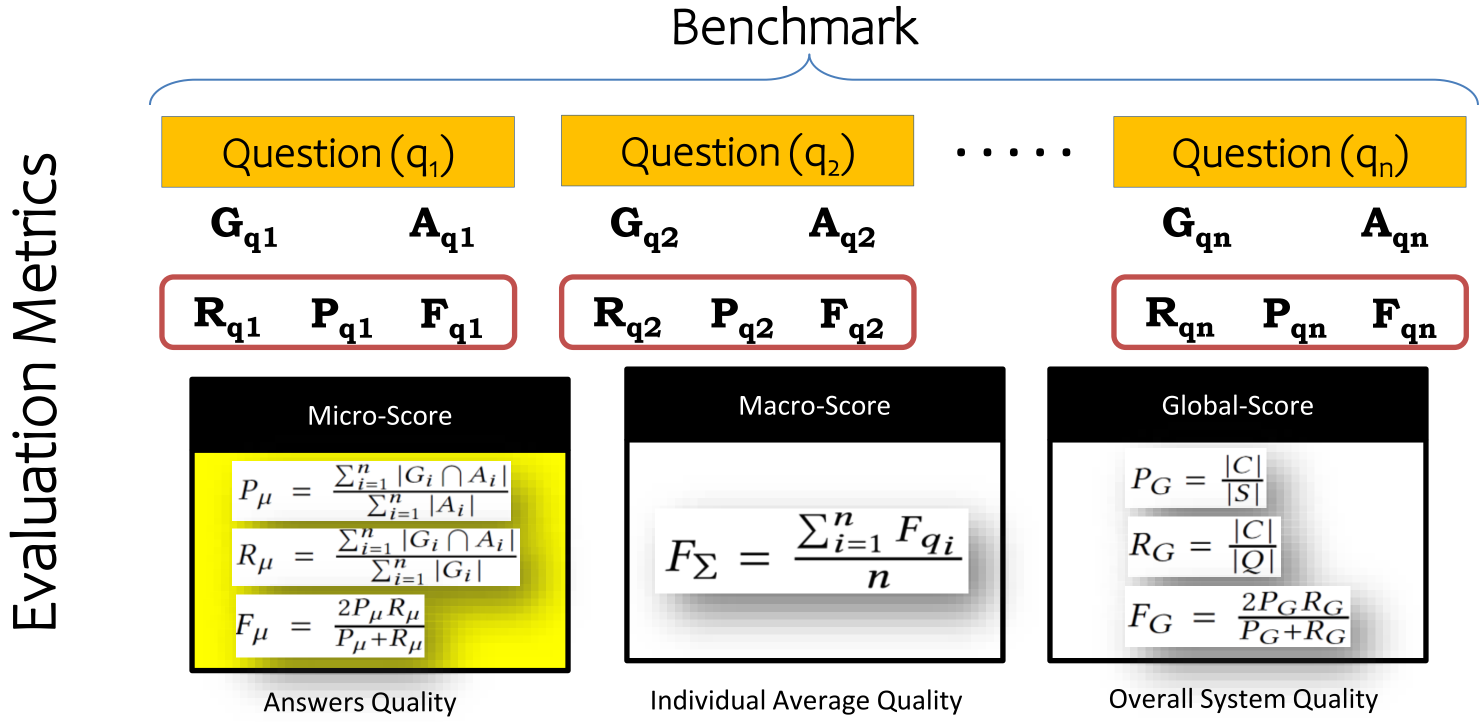
Overall System Quality

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation



Benchmark

Evaluation Metrics

Question (q_1) · Question (q_2) · ...... · Question (q_n)

$\mathbf{G_{q1}}$ $\mathbf{A_{q1}}$ · $\mathbf{G_{q2}}$ $\mathbf{A_{q2}}$ · $\mathbf{G_{qn}}$ $\mathbf{A_{qn}}$

$\mathbf{R_{q1}}$ $\mathbf{P_{q1}}$ $\mathbf{F_{q1}}$ · $\mathbf{R_{q2}}$ $\mathbf{P_{q2}}$ $\mathbf{F_{q2}}$ · $\mathbf{R_{qn}}$ $\mathbf{P_{qn}}$ $\mathbf{F_{qn}}$

Micro-Score

$$P_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |A_i|}$$

$$R_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |G_i|}$$

$$F_\mu = \frac{2P_\mu R_\mu}{P_\mu + R_\mu}$$

Answers Quality

Macro-Score

$$F_\Sigma = \frac{\sum_{i=1}^{n} F_{q_i}}{n}$$

Individual Average Quality

Global-Score

$$P_G = \frac{|C|}{|S|}$$

$$R_G = \frac{|C|}{|Q|}$$

$$F_G = \frac{2P_G R_G}{P_G + R_G}$$

Overall System Quality

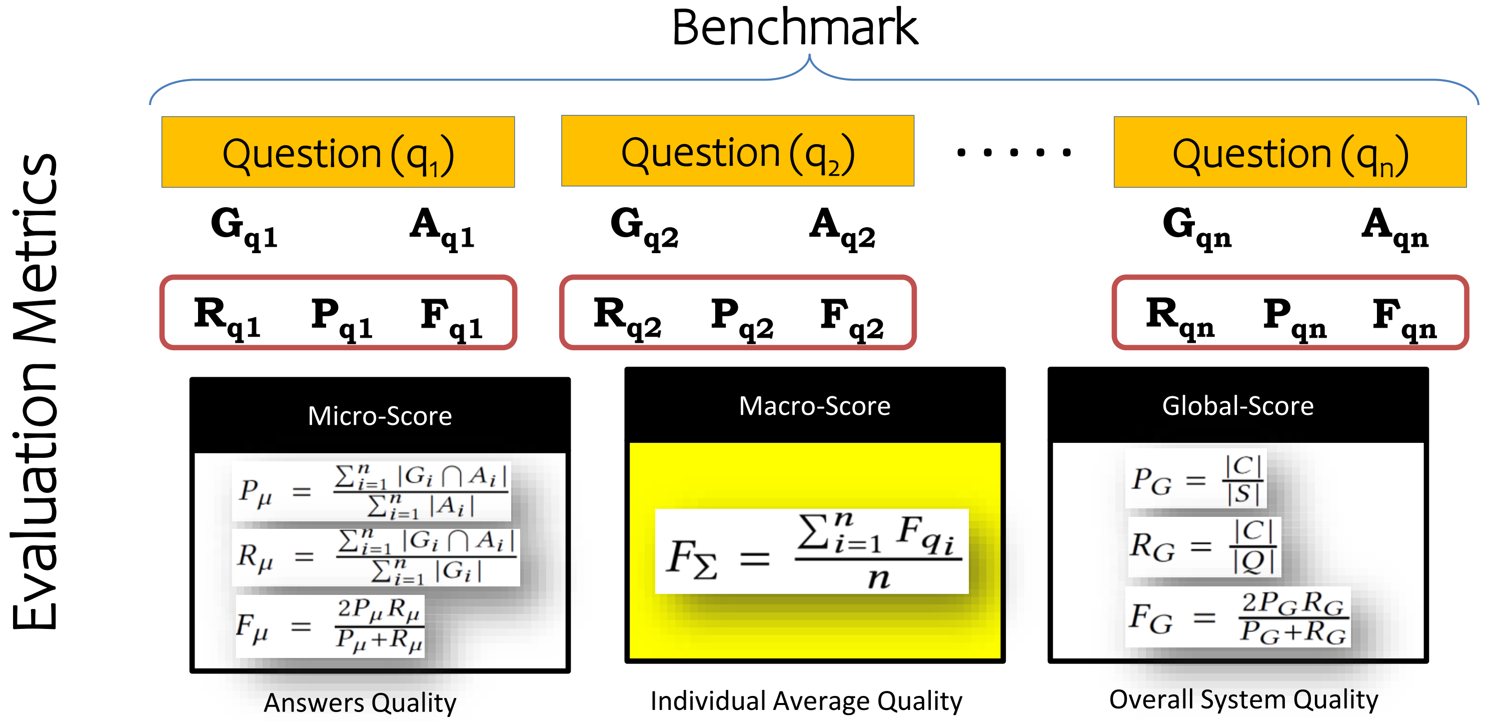
QAKG Past and Future

Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

- Benchmarks Analysis
- QA Evaluation



Benchmark

Evaluation Metrics

Question (q_1)	Question (q_2)	······	Question (q_n)
$\mathbf{G_{q1}}$ $\mathbf{A_{q1}}$	$\mathbf{G_{q2}}$ $\mathbf{A_{q2}}$		$\mathbf{G_{qn}}$ $\mathbf{A_{qn}}$
$\mathbf{R_{q1}}$ $\mathbf{P_{q1}}$ $\mathbf{F_{q1}}$	$\mathbf{R_{q2}}$ $\mathbf{P_{q2}}$ $\mathbf{F_{q2}}$		$\mathbf{R_{qn}}$ $\mathbf{P_{qn}}$ $\mathbf{F_{qn}}$

Micro-Score

$$P_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |A_i|}$$

$$R_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |G_i|}$$

$$F_\mu = \frac{2P_\mu R_\mu}{P_\mu + R_\mu}$$

Answers Quality

Macro-Score

$$F_\Sigma = \frac{\sum_{i=1}^{n} F_{q_i}}{n}$$

Individual Average Quality

Global-Score

$$P_G = \frac{|C|}{|S|}$$

$$R_G = \frac{|C|}{|Q|}$$

$$F_G = \frac{2P_G R_G}{P_G + R_G}$$

Overall System Quality

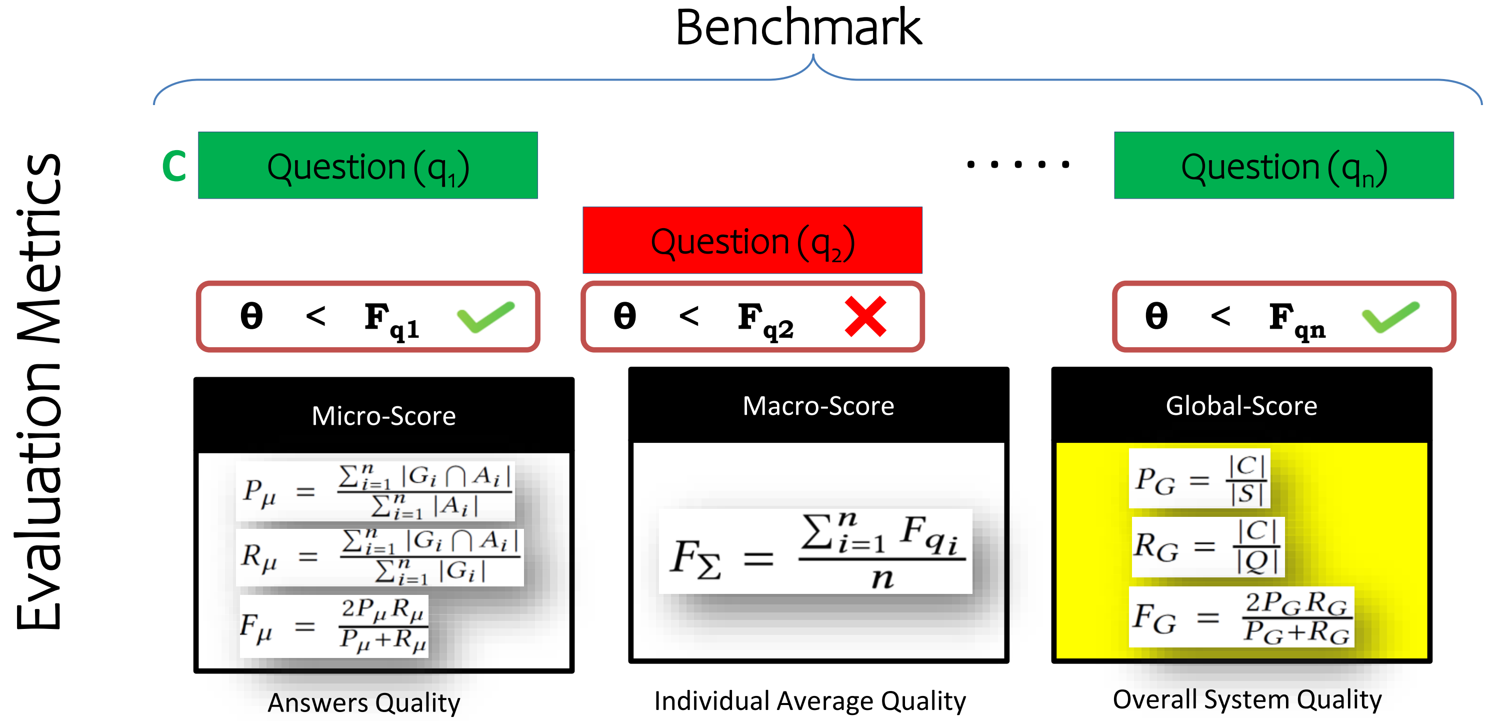
Step (1/3)

CBench (VLDB 2021 [Research Paper & Demo Paper])

done

Benchmarks Analysis

QA Evaluation



Benchmark

Evaluation Metrics

C

Question (q_1)

Question (q_2)

. . . . .

Question (q_n)

$\theta < F_{q1}$ ✓

$\theta < F_{q2}$ ✗

$\theta < F_{qn}$ ✓

Micro-Score

$$P_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |A_i|}$$

$$R_\mu = \frac{\sum_{i=1}^{n} |G_i \cap A_i|}{\sum_{i=1}^{n} |G_i|}$$

$$F_\mu = \frac{2 P_\mu R_\mu}{P_\mu + R_\mu}$$

Answers Quality

Macro-Score

$$F_\Sigma = \frac{\sum_{i=1}^{n} F_{qi}}{n}$$

Individual Average Quality

Global-Score

$$P_G = \frac{|C|}{|S|}$$

$$R_G = \frac{|C|}{|Q|}$$

$$F_G = \frac{2 P_G R_G}{P_G + R_G}$$

Overall System Quality