CABINET:
Content RelevAnce Based NoIse ReductioN for TablE QuesTion Answering

Sohan Patnaik, Heril Changwal, Milan Aggarwal,
Sumit Bhatia, Yaman Kumar, Balaji Krishnamurthy

ICLR 2024 (Spotlight)
Poster ID: 18603
Halle B, Tue May 7
10:45 a.m. CEST – 12:45 p.m. CEST

Table Understanding using LLMs

week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

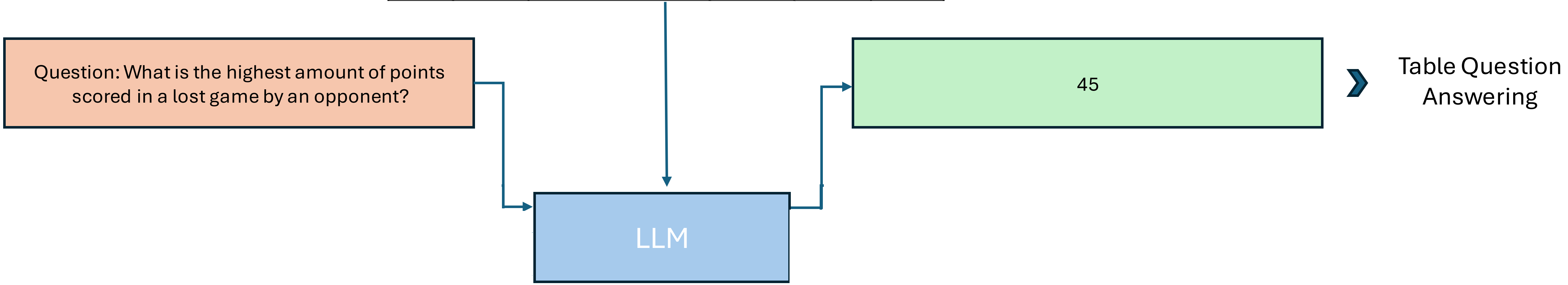


Table Understanding using LLMs

week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

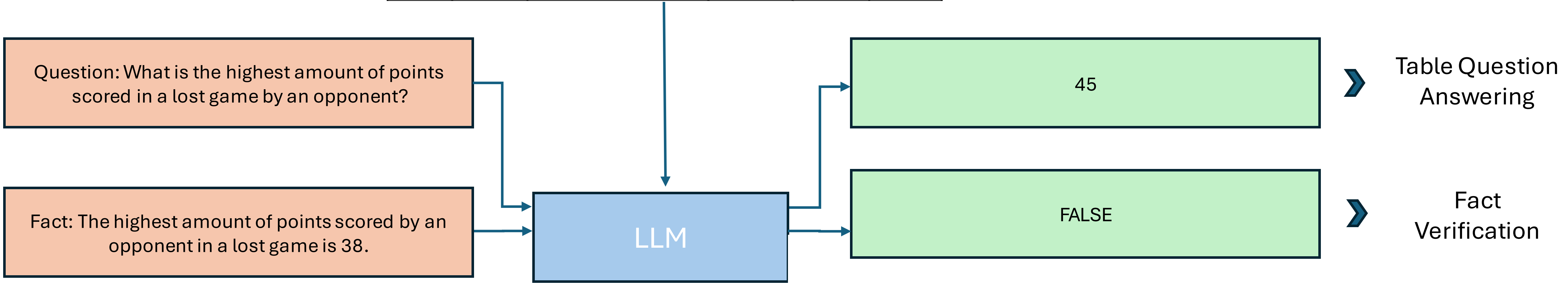
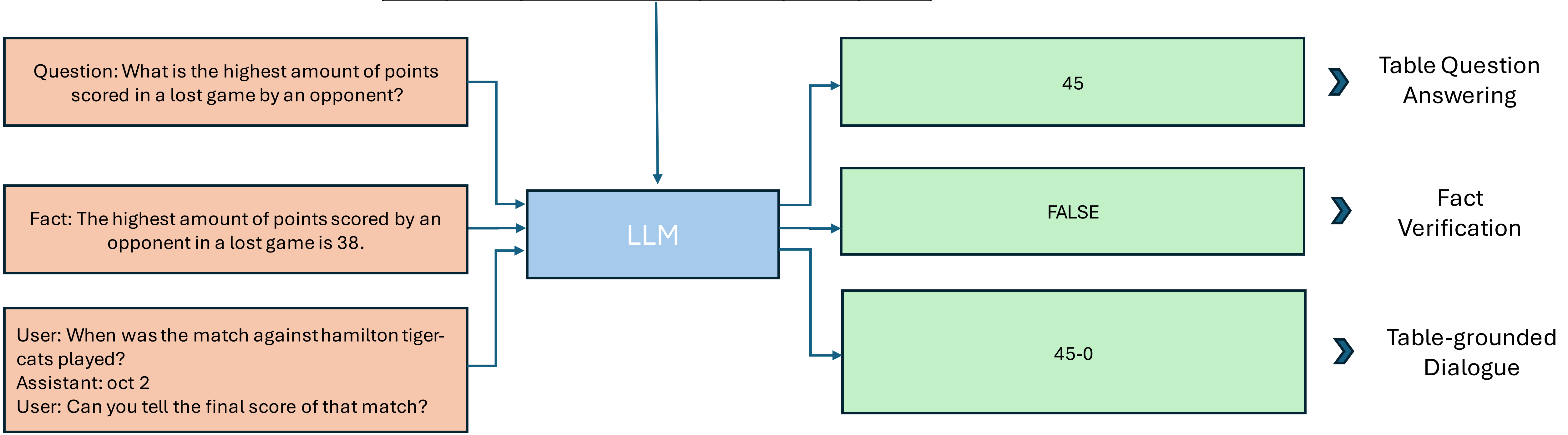


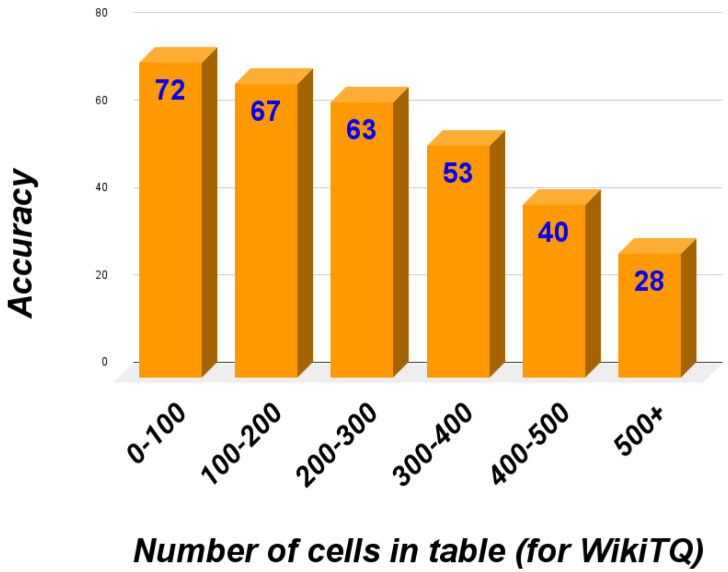
Table Understanding using LLMs

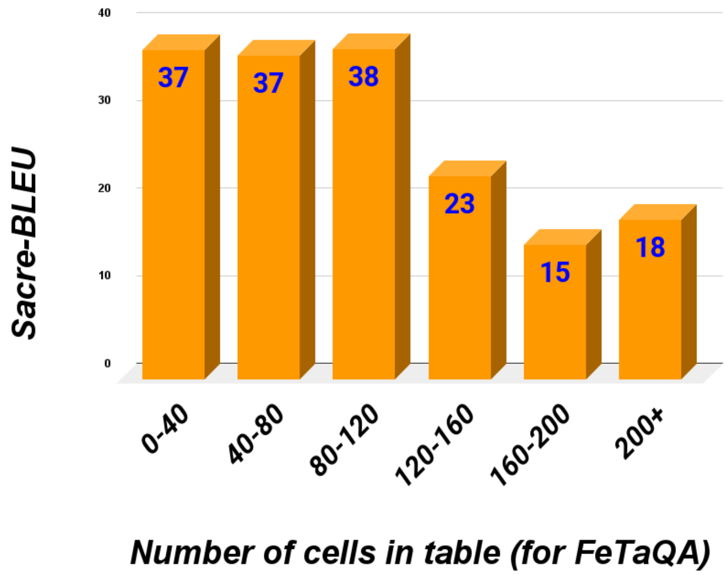
week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12



LLMs suffer drop in performance with increase in Table size

- Transformer-based LLMs trained on natural language text often struggle to comprehend the structure and compositionality of tabular data.
- LLMs have been adapted for tables through joint learning and pre-training on table semantic parsing, and synthesizing table-based template-based questions.
- The presence of irrelevant tabular data acts as noise or distracting information, leading to suboptimal performance by SoTA methods **while performing question-answering**, especially in large tables.





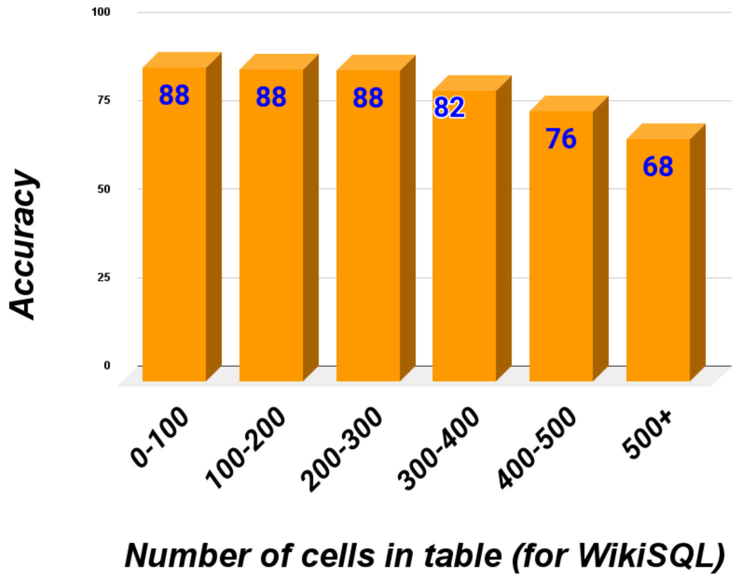


Table Decomposition based Prior Art

Question:
What is the highest amount of points scored in a lost game by an opponent?

week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

Table

Table Decomposition based Prior Art

Question:
What is the highest amount of points scored in a lost game by an opponent?

week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

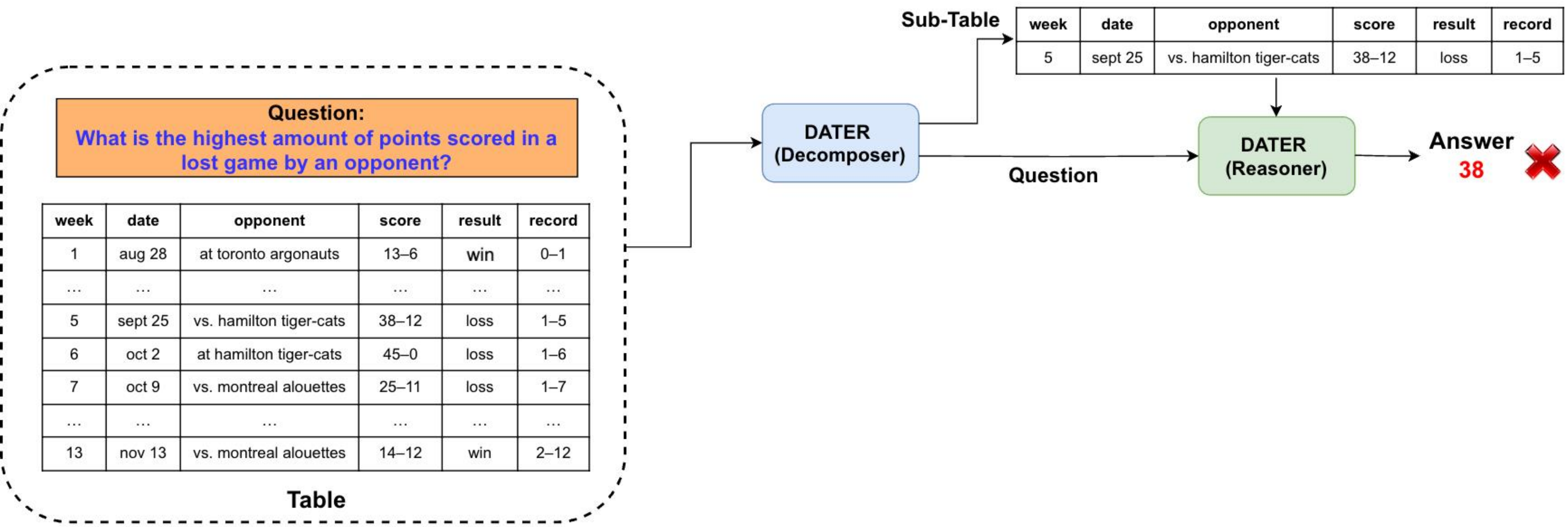
Table

DATER (Decomposer)

Sub-Table

week	date	opponent	score	result	record
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5

Table Decomposition based Prior Art



week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

week	date	opponent	score	result	record
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5

Table Decomposition based Prior Art vs CABINET

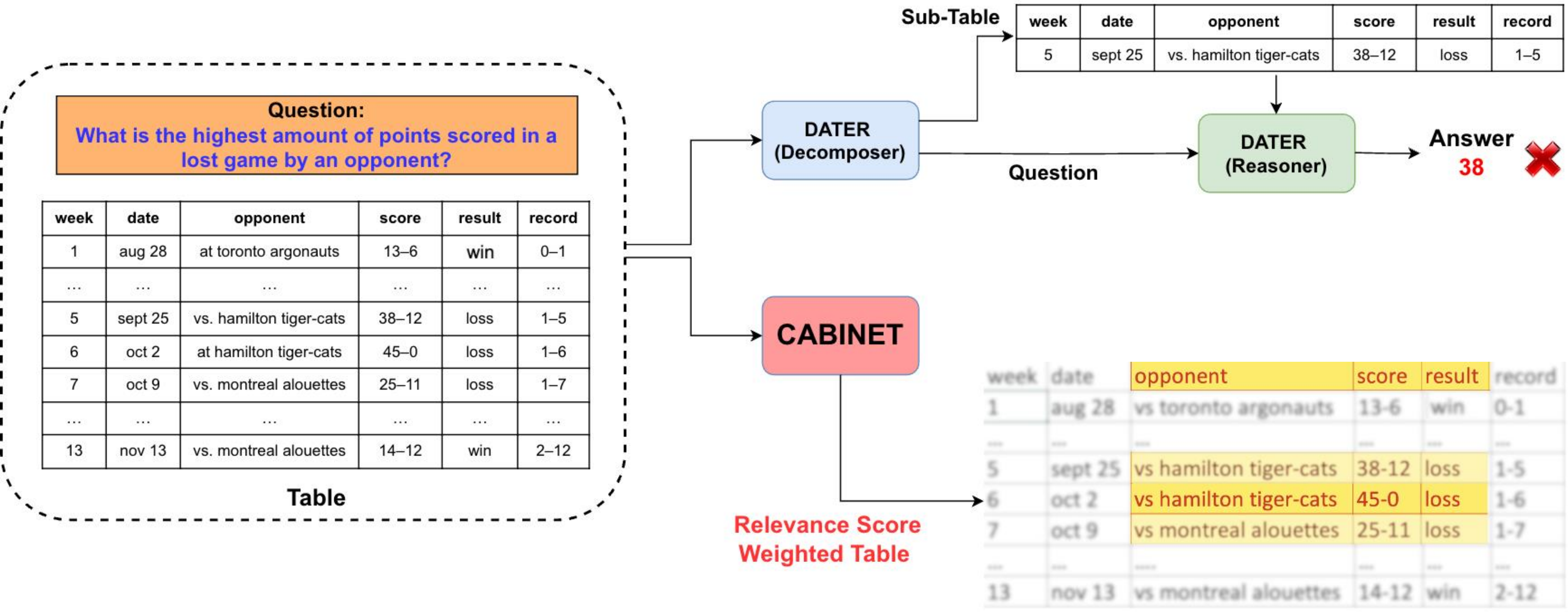
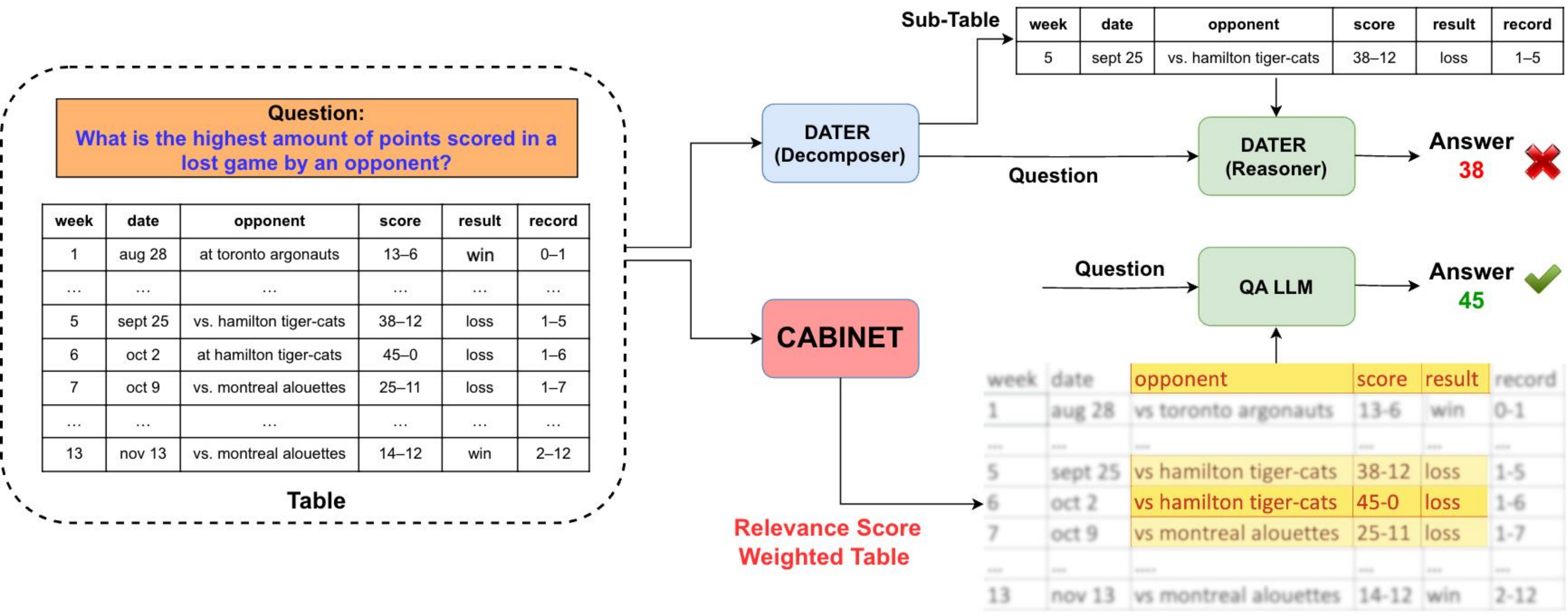


Table Decomposition based Prior Art vs CABINET



week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

week	date	opponent	score	result	record
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5

week	date	opponent	score	result	record
1	aug 28	vs toronto argonauts	13-6	win	0-1
...	...	...	...	...	...
5	sept 25	vs hamilton tiger-cats	38-12	loss	1-5
6	oct 2	vs hamilton tiger-cats	45-0	loss	1-6
7	oct 9	vs montreal alouettes	25-11	loss	1-7
...	...	.....	...	...	...
13	nov 13	vs montreal alouettes	14-12	win	2-12

CABINET – Architecture of Proposed Method

Question + Flattened Table

1

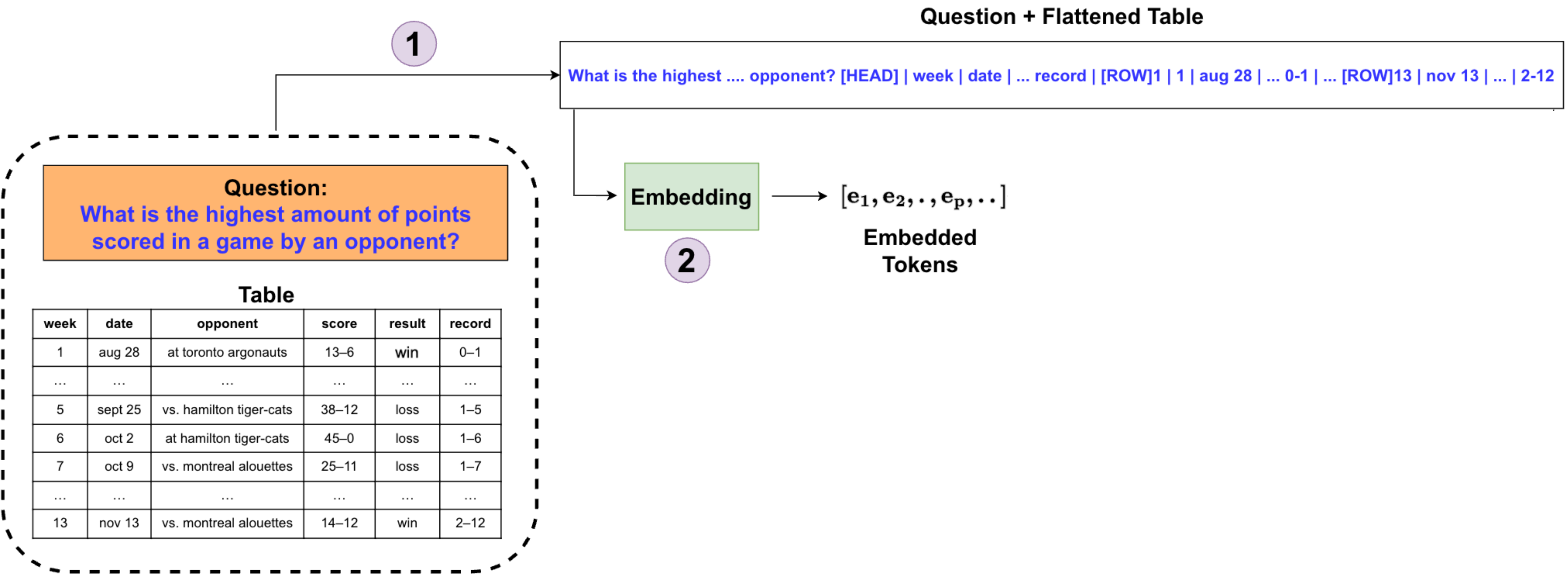
What is the highest .... opponent? [HEAD] | week | date | ... record | [ROW]1 | 1 | aug 28 | ... 0-1 | ... [ROW]13 | nov 13 | ... | 2-12

Question:
What is the highest amount of points scored in a game by an opponent?

Table

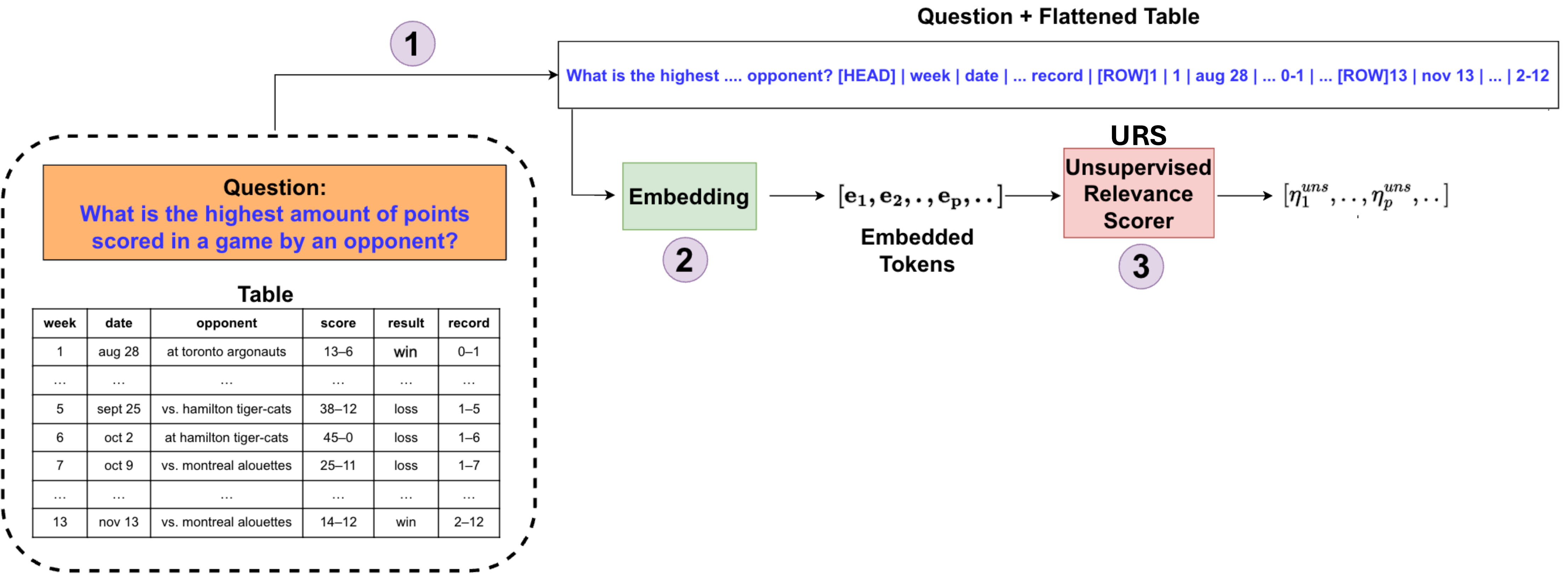
week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

CABINET – Architecture of Proposed Method



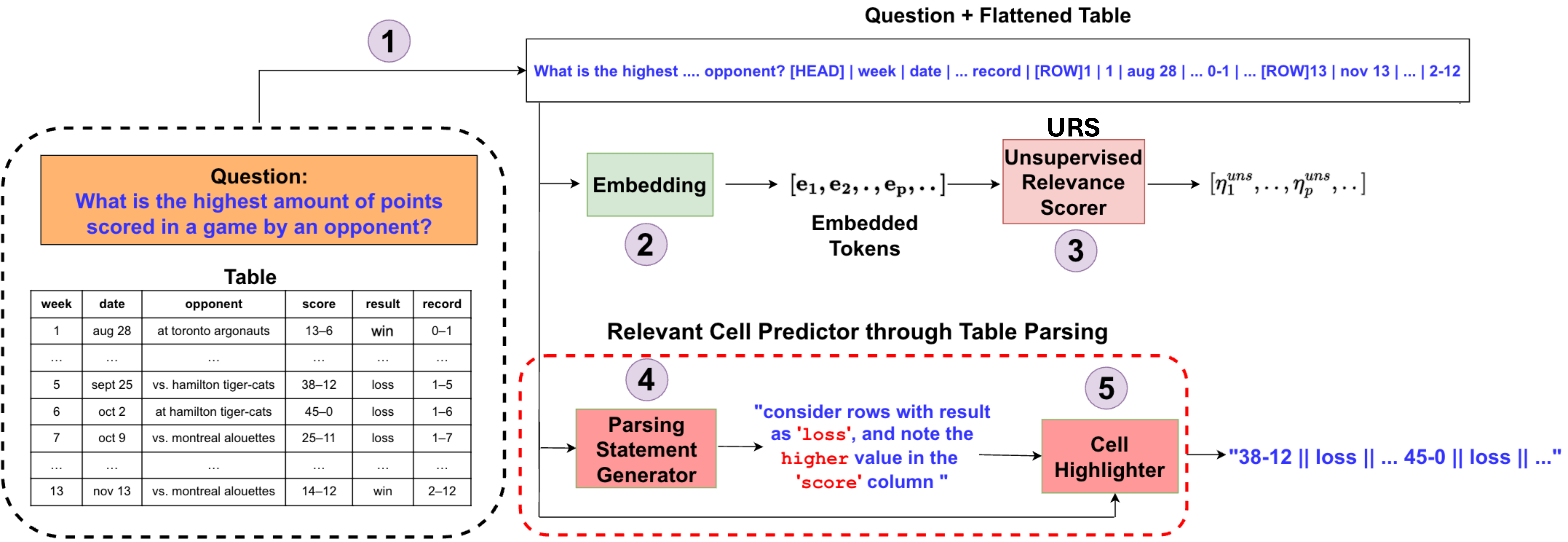
week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

CABINET – Architecture of Proposed Method

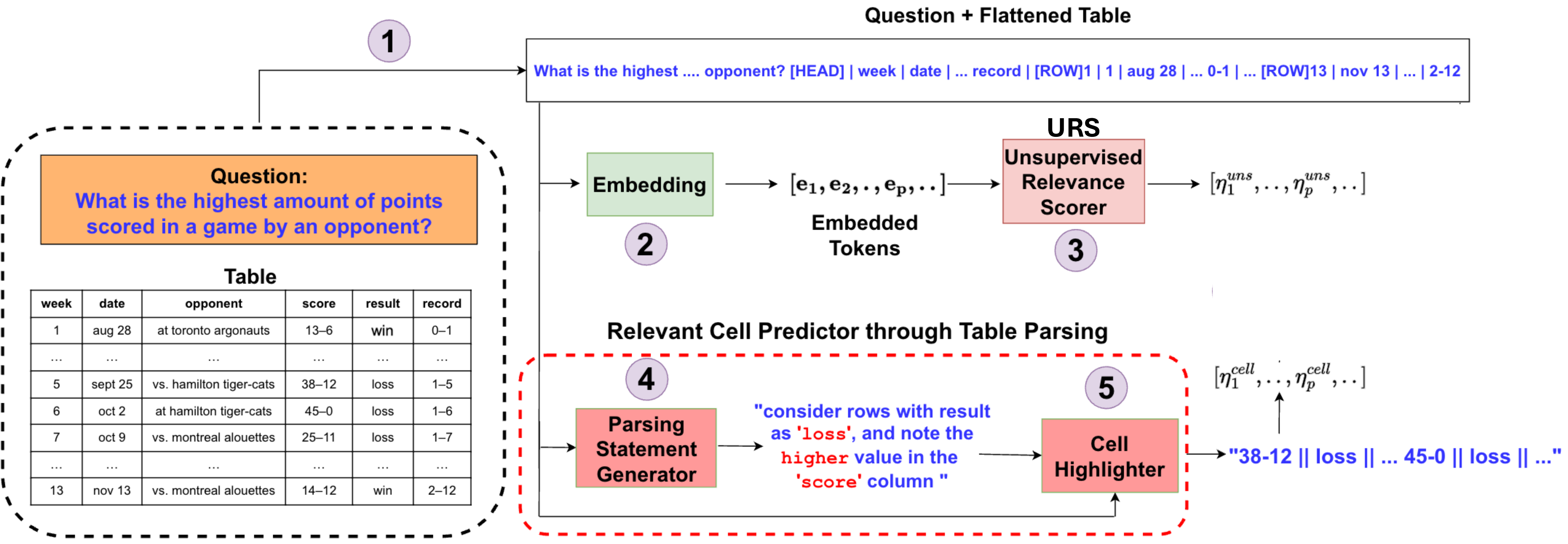


week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

CABINET – Architecture of Proposed Method

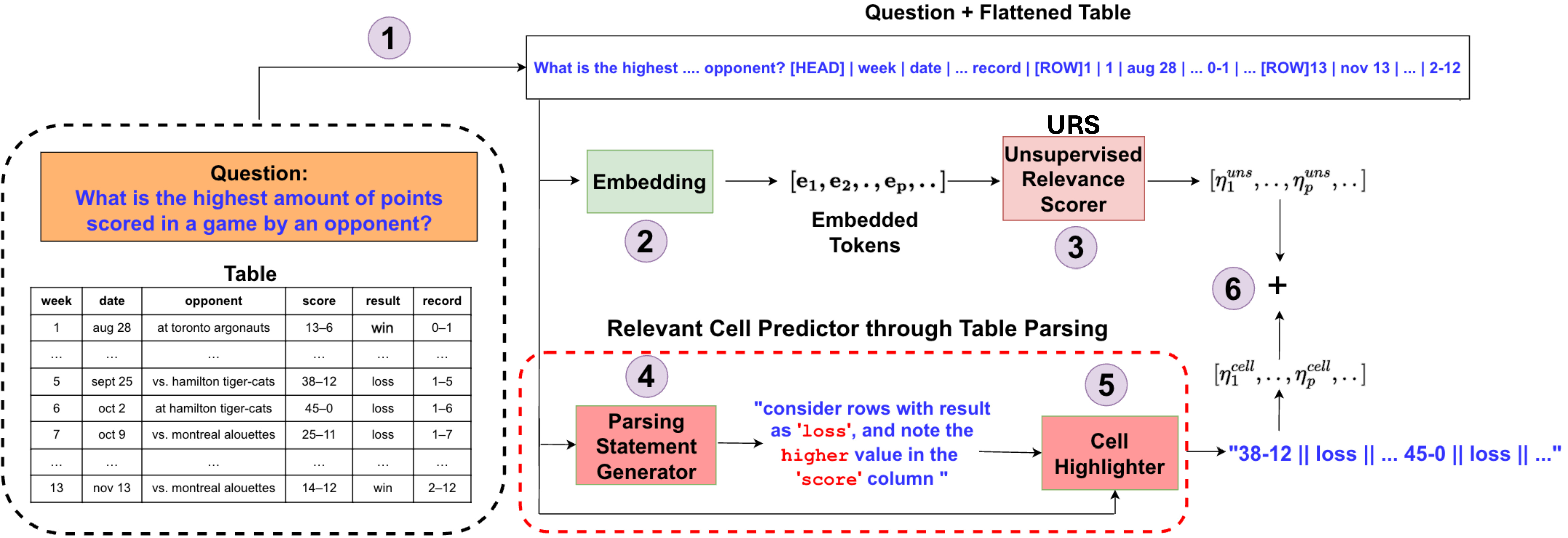


CABINET – Architecture of Proposed Method

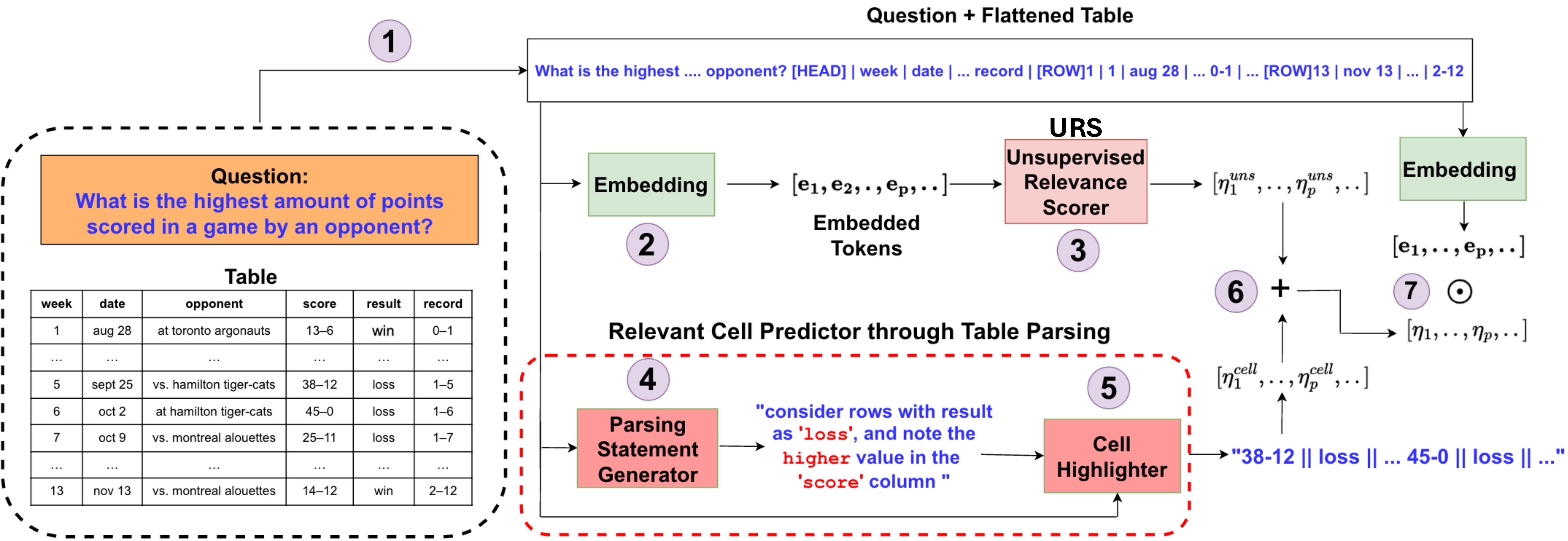


week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
...	...	...	...	...	...
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
...	...	...	...	...	...
13	nov 13	vs. montreal alouettes	14–12	win	2–12

CABINET – Architecture of Proposed Method

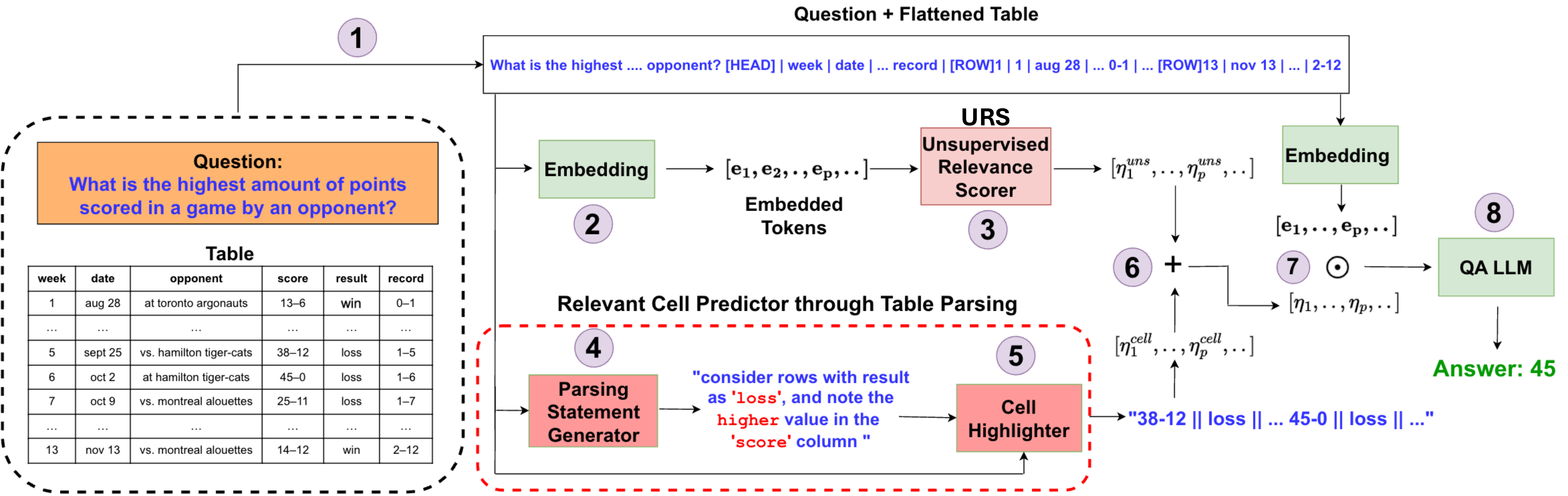


CABINET – Architecture of Proposed Method



week	date	opponent	score	result	record
1	aug 28	at toronto argonauts	13–6	win	0–1
…	…	…	…	…	…
5	sept 25	vs. hamilton tiger-cats	38–12	loss	1–5
6	oct 2	at hamilton tiger-cats	45–0	loss	1–6
7	oct 9	vs. montreal alouettes	25–11	loss	1–7
…	…	…	…	…	…
13	nov 13	vs. montreal alouettes	14–12	win	2–12

CABINET – Architecture of Proposed Method



CABINET outperforms several SoTA methods while being orders of magnitude smaller in # parameters

Method	Acc.	# params
Fine-tuning Table-specific LLMs		
TAPAS (Herzig et al., 2020)	48.8	345 M
TaBERT (Yin et al., 2020)	52.3	345 M
MATE (Eisenschlos et al., 2021)	51.5	340 M
GraPPa (Yu et al., 2021)	52.7	355 M
DoT (Krichene et al., 2021)	54.0	299 M
TableFormer (Yang et al., 2022)	52.6	345 M
TAPEX (Liu et al., 2022)	55.5	405 M
ReasTAP (Zhao et al., 2022)	58.6	406 M
TaCube (Zhou et al., 2022)	60.8	406 M
OmniTab (Jiang et al., 2022)	62.7	406 M
Fine-tuning text-based LLMs		
T5-3b (Xie et al., 2022))	49.3	2.9 B
FlanT5-xl (Chung et al., 2022a)	64.4	2.9 B
Few/zero shot Prompting of LLMs		
Codex (Ye et al., 2023)	47.6	175 B
Codex-COT (Chen, 2023)	48.8	175 B
Binder (Cheng et al., 2023)	64.6	175 B
LEVER (Ni et al., 2023)	65.8	175 B
DATER (Ye et al., 2023)	65.9	175 B
ChatGPT (Jiang et al., 2023)	43.3	175 B
StructGPT (Jiang et al., 2023)	48.4	175 B
CABINET (Ours)	**69.1**	560 M

Table 1: Comparison of CABINET with different baselines on WikiTQ. CABINET achieves significantly better accuracy.

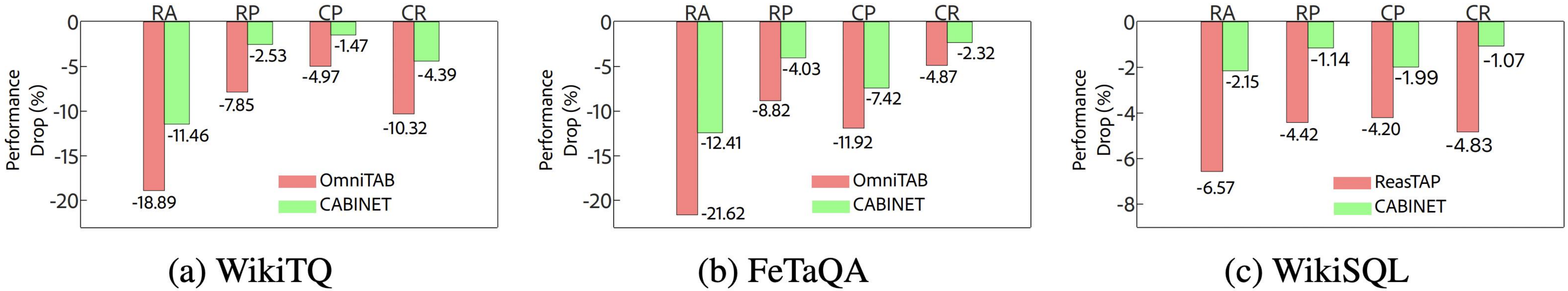
Method	S-BLEU	# params
Fine-tuning Table-specific LLMs		
PeaQA (Pal et al., 2022)	33.5	406 M
TAPEX (Liu et al., 2022)	34.7	406 M
OmniTab (Jiang et al., 2022)	34.9	406 M
Fine-tuning text-based LLMs		
T5-small (Nan et al., 2022)	21.6	60 M
T5-base (Nan et al., 2022)	28.1	222 M
T5-large (Nan et al., 2022)	30.5	738 M
T5-3b (Xie et al., 2022)	33.4	2.9 B
FlanT5-xl	36.2	2.9 B
Few/zero shot Prompting of LLMs		
Codex-COT (Chen, 2023)	27.0	175 B
Codex (Ye et al., 2023)	27.9	175 B
DATER (Ye et al., 2023)	30.9	175 B
CABINET (Ours)	**40.5**	560 M

Table 2: Comparison with different categories of baselines on FeTaQA. CABINET achieves significantly better Sacre-BLEU (S-BLEU).

Method	Acc.	# params
Fine-tuning Table-specific LLMs		
TAPAS (Herzig et al., 2020)	86.4	345 M
GraPPa (Yu et al., 2021)	84.7	355 M
DoT (Krichene et al., 2021)	85.5	299 M
TAPEX (Liu et al., 2022)	86.4	406 M
OmniTab (Jiang et al., 2022)	87.9	406 M
UTP (Chen et al., 2023b)	88.1	345 M
ReasTAP (Zhao et al., 2022)	88.8	406 M
Fine-tuning text-based LLMs		
T5-3b (Xie et al., 2022)	85.9	2.9 B
FlanT5-xl	87.8	2.9 B
Few/zero shot Prompting of LLMs		
ChatGPT (Jiang et al., 2023)	51.6	175 B
StructGPT (Jiang et al., 2023)	54.4	175 B
CABINET (Ours)	**89.5**	560 M

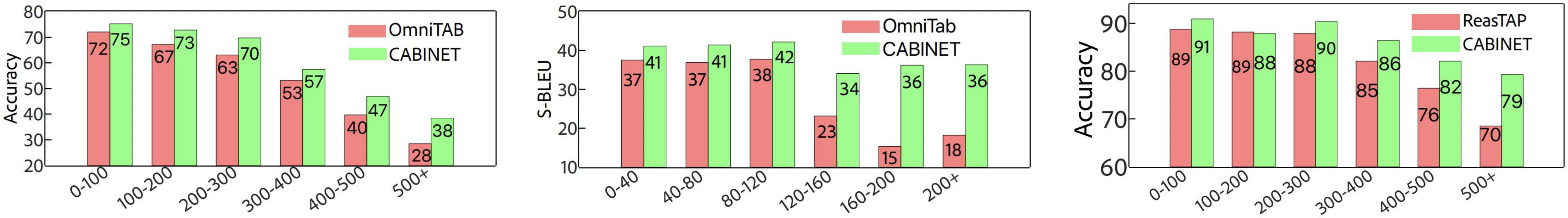
Table 3: Comparison with different categories of baselines on WikiSQL. CABINET achieves better Accuracy (Acc.).

CABINET is more robust against noisy information and table perturbations



Relative performance drop (%) with perturbations (RA - Row Addition, RP - Row Permutation, CP - Column Permutation, CR - Cell Replacement). We compare CABINET (green) with OmniTab (red) on WikiTQ and FeTaQA ; and against ReasTAP (red) on WikiSQL. CABINET is more robust to addition of noise to table and shuffling of row and column ordering.

CABINET is more performant while handling high volume of data in large tables

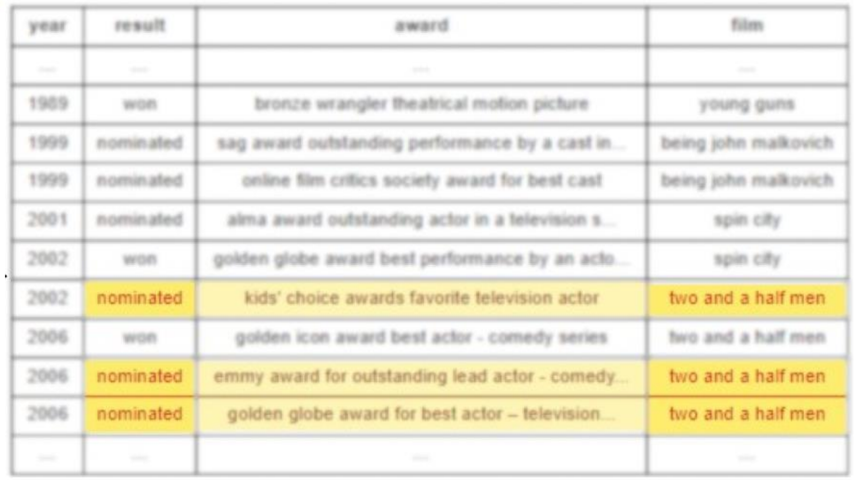


(a) WikiTQ (b) FeTaQA (c) WikiSQL

Variation in performance with table size (# cells). We compare CABINET (green) with OmniTab (red) on WikiTQ (left) and FeTaQA (middle), and against ReasTAP (red) for WikiSQL (right). It can be seen that CABINET performs much better than the baselines on larger tables.

Clustering Relevance Score through URS and Sparsifying Relevance Scores is Helpful

Question:
What is the total number of awards (nominations and wins) for two and a half men?



year	result	award	film
...	...	...	...
1989	won	bronze wrangler theatrical motion picture	young guns
1999	nominated	sag award outstanding performance by a cast in...	being john malkovich
1999	nominated	online film critics society award for best cast	being john malkovich
2001	nominated	alma award outstanding actor in a television s...	spin city
2002	won	golden globe award best performance by an acto...	spin city
2002	nominated	kids' choice awards favorite television actor	two and a half men
2006	won	golden icon award best actor - comedy series	two and a half men
2006	nominated	emmy award for outstanding lead actor - comedy...	two and a half men
2006	nominated	golden globe award for best actor – television...	two and a half men
...	...	...	...

$\mathcal{L}_{clu}$	$\mathcal{L}_{sep}$	$\mathcal{L}_{sparse}$	**WikiTQ**	**FeTaQA**	**WikiSQL**
✗	✗	✗	60.8	35.1	86.2
✗	✗	✓	60.9	35.1	86.3
✓	✗	✗	62.7	35.0	88.9
✓	✗	✓	61.0	35.0	**89.5**
✓	✓	✗	61.0	35.1	89.1
✓	✓	✓	**65.6**	**35.8**	89.3

Table 4: Effect of applying clustering ($\mathcal{L}_{clu}$), centroid separation ($\mathcal{L}_{sep}$) and relevance score sparsification loss ($\mathcal{L}_{sparse}$). Clustering table tokens by enforcing sparsity in relevance scores and distance between cluster centroids improves performance.

Conclusion and Future Work

- CABINET effectively addresses the challenge of question-answering over tables by identifying and weighing relevant portions of tabular data, enhancing performance by mitigating noise.
- Outperforming existing methods on three challenging benchmarks, CABINET establishes new state-of-the-art results, surpassing table-specific models and larger scale GPT-3-based approaches.
- Empirical evidence demonstrates CABINET's robustness to noise and its ability to generalize well to larger tables, indicating its efficacy in overcoming structural biases and enhancing performance in table-based question-answering tasks.
- As future work, it can be explored if CABINET can be used to improve performance while dealing with multiple tables and additional information present in the form of passages.

Adobe