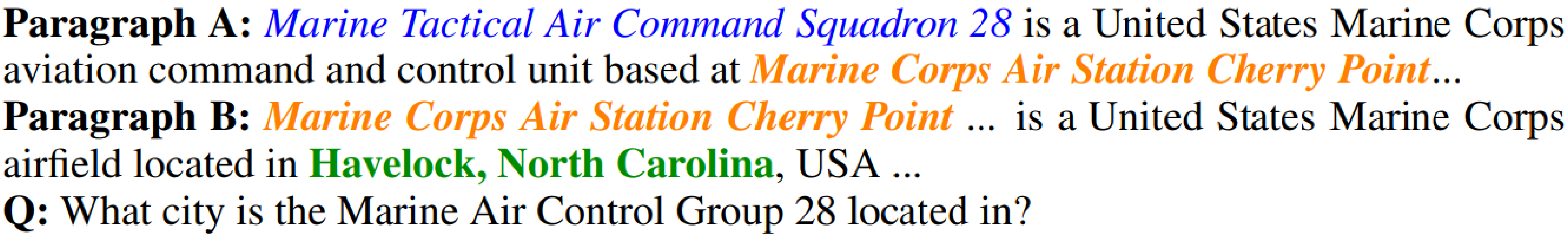
MEQA: A Benchmark for Multi-hop Event-centric Question Answering with Explanations

Ruosen Li, Zimu Wang, Son Quoc Tran, Lei Xia, Xinya Du

Department of Computer Science, University of Texas at Dallas

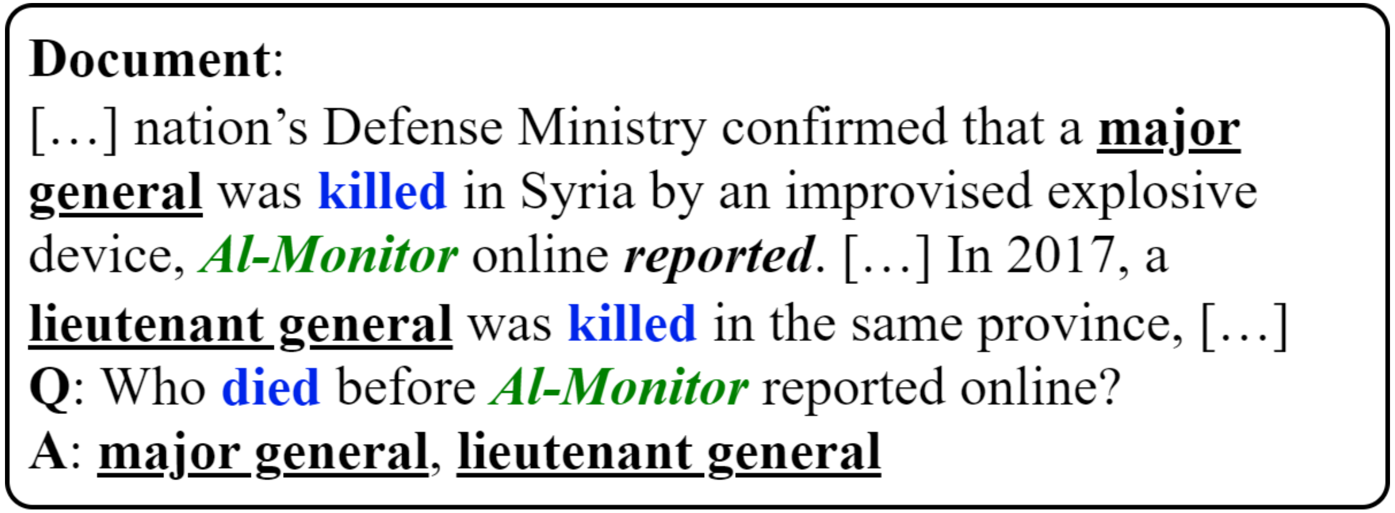
Entity-Centric Question Answering



Paragraph A: *Marine Tactical Air Command Squadron 28* is a United States Marine Corps aviation command and control unit based at ***Marine Corps Air Station Cherry Point***...
Paragraph B: ***Marine Corps Air Station Cherry Point*** ... is a United States Marine Corps airfield located in **Havelock, North Carolina**, USA ...
Q: What city is the Marine Air Control Group 28 located in?

Yang, Zhilin, et al. "HotpotQA: A dataset for diverse, explainable multi-hop question answering." *arXiv preprint arXiv:1809.09600* (2018).

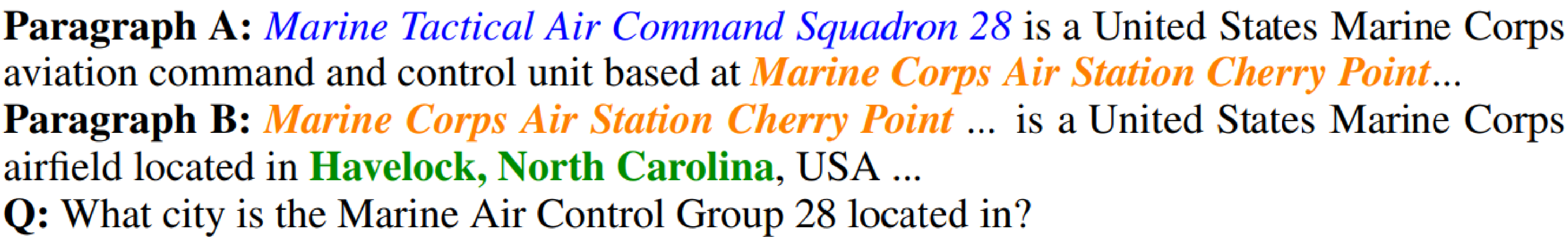
Event-Centric Question Answering



Document:
[…] nation's Defense Ministry confirmed that a **major general** was **killed** in Syria by an improvised explosive device, ***Al-Monitor*** online ***reported***. […] In 2017, a **lieutenant general** was **killed** in the same province, […]
Q: Who **died** before ***Al-Monitor*** reported online?
A: **major general**, **lieutenant general**

An example of multi-hop event-centric question in MEQA. Models should start reasoning from the ***Al-Monitor*** and first locate the ***reported*** event; then find all **events** that happened before the reported event; and finally extract **victims** in all those events, which are answers to the question.

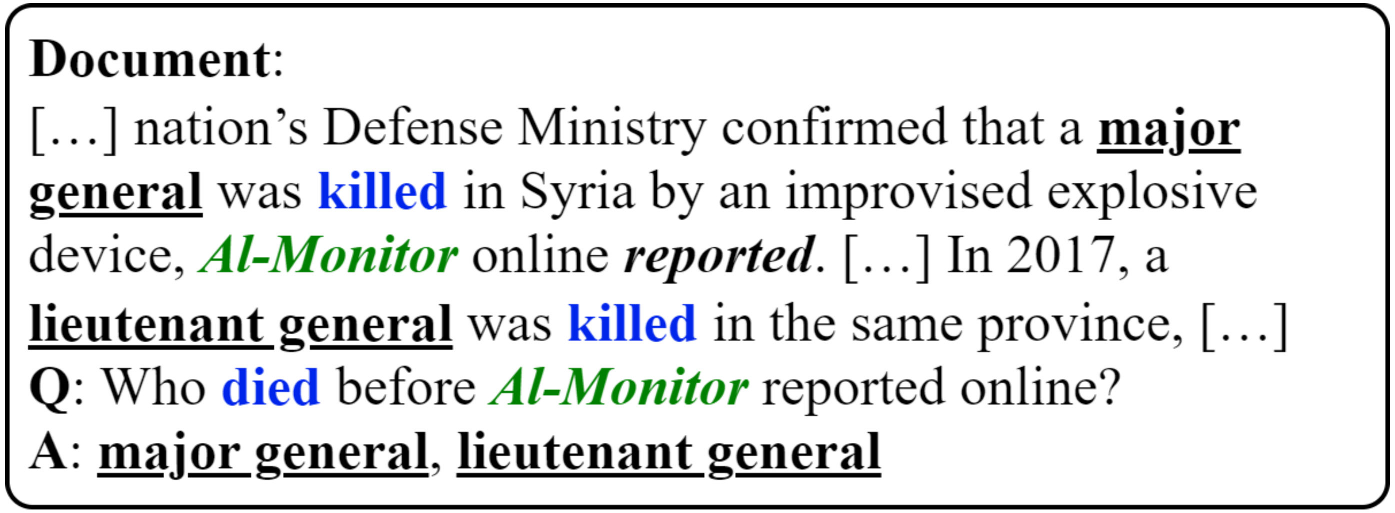
Comparison Between Entity- and Event-Centric Question Answering



Paragraph A: *Marine Tactical Air Command Squadron 28* is a United States Marine Corps aviation command and control unit based at ***Marine Corps Air Station Cherry Point***...

Paragraph B: ***Marine Corps Air Station Cherry Point*** ... is a United States Marine Corps airfield located in **Havelock, North Carolina**, USA ...

Q: What city is the Marine Air Control Group 28 located in?



Document:

[…] nation's Defense Ministry confirmed that a **major general** was **killed** in Syria by an improvised explosive device, ***Al-Monitor*** online ***reported***. […] In 2017, a **lieutenant general** was **killed** in the same province, […]

Q: Who **died** before ***Al-Monitor*** reported online?

A: **major general**, **lieutenant general**

Event-Centric Question Answering Requires:

1. **More Complex Relations**
2. **More Complex Reasoning Process**

Reasoning Types

- Event Relation
- Entity Bridging
- Event Listing and Counting
- Event Comparison
- Unanswerable

Event Relation

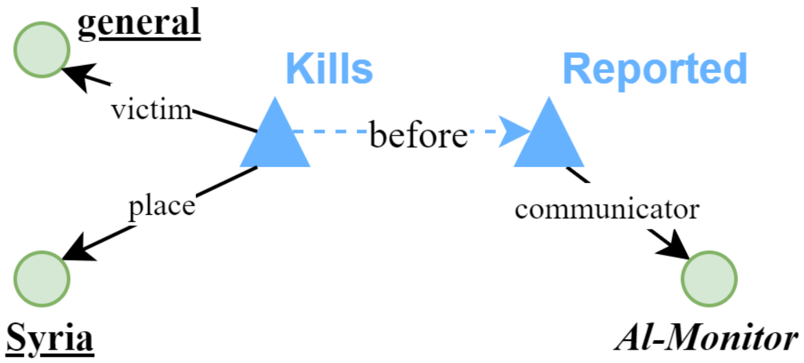
Document:

[…] nation’s Defense Ministry confirmed that a major **general** was killed in Syria by an improvised explosive device, ***Al-Monitor*** online reported. […] ***Al-Monitor*** is a news website launched by Jamal Daniel (born in **Syria**) [...]

Questions and Answers:

Q1: Who died before ***Al-Monitor*** announced it online?
A: **general**
Q2: Where was the founder of ***Al-Monitor*** born?
A: **Syria**

Graph:



Explanations:

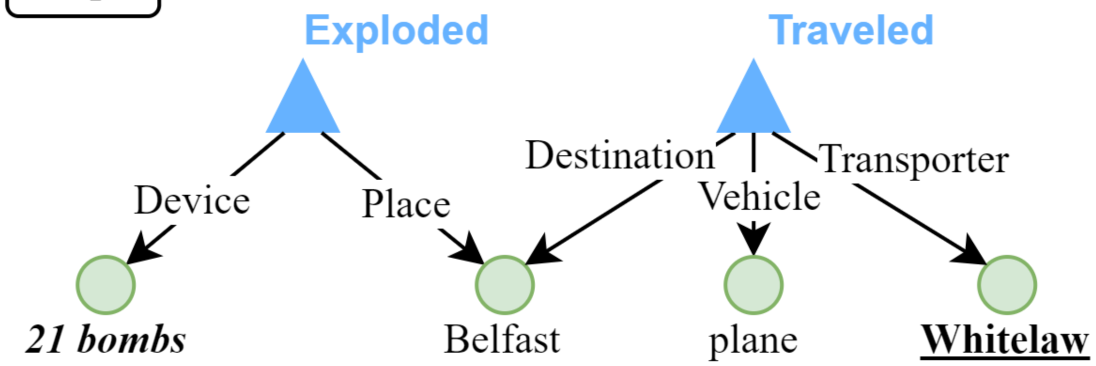
Q1-1: What happened before ***Al-Monitor*** announced the death online?
A1-1: **Kills**
Q1-2: Who is the victim in the **Kills** event?
A1-2: **General**
Q2-1: Who was the founder of ***Al-Monitor***?
A2-1: **Jamal**
Q2-2: Where was **Jamal** born?
A2-2: **Syria**

Entity Bridging

Document:

Early today, ***21 bombs*** had **exploded** in Belfast […] Mr. **Whitelaw** immediately **traveled** back to Belfast by plane […]

Graph:



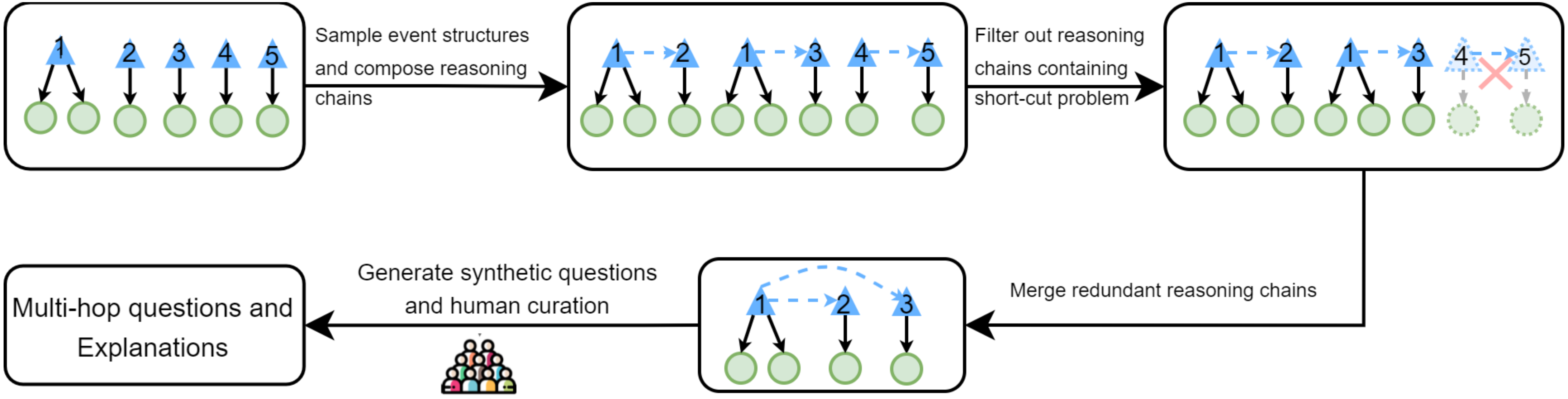
Question and Answer:

Q: Who traveled by plane to the place where ***21 bombs*** exploded.
A: **Whitelaw**

Explanations:

Q1: Where was ***21 bombs*** exploded?
A1: **Belfast**
Q2: Who traveled to the place by plane?
A2: **Whitelaw**

MEQA Construction Pipeline



Sample event structures and compose reasoning chains

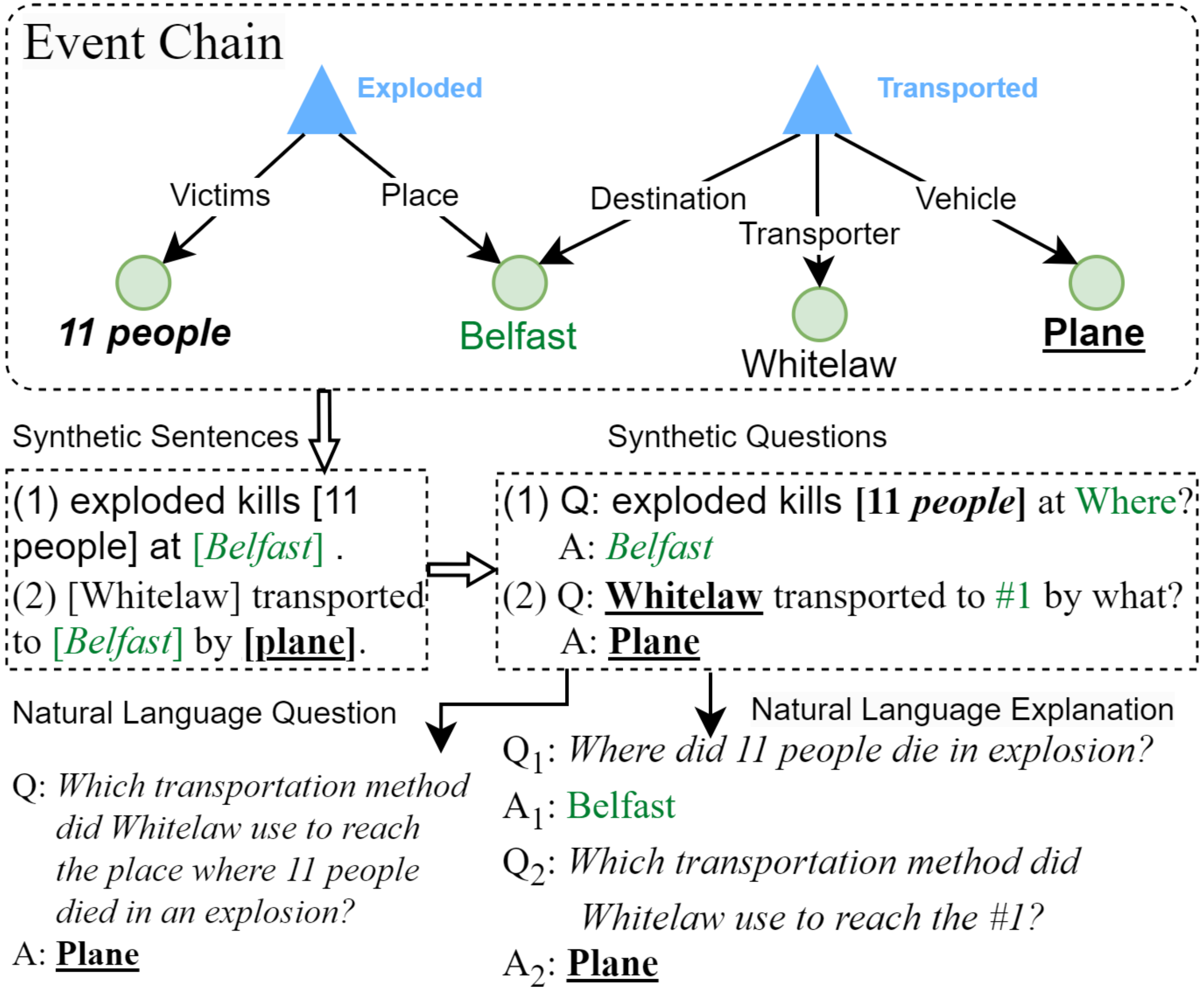
Filter out reasoning chains containing short-cut problem

Merge redundant reasoning chains

Generate synthetic questions and human curation

Multi-hop questions and Explanations

Generating Synthetic Questions



Synthetic Sentences

(1) exploded kills [11 people] at *[Belfast]* .
(2) [Whitelaw] transported to *[Belfast]* by **[plane]**.

Synthetic Questions

(1) Q: exploded kills **[11 *people*]** at Where?
A: *Belfast*
(2) Q: **Whitelaw** transported to #1 by what?
A: **Plane**

Natural Language Question

Q: *Which transportation method did Whitelaw use to reach the place where 11 people died in an explosion?*
A: **Plane**

Natural Language Explanation

Q_1: *Where did 11 people die in explosion?*
A_1: Belfast
Q_2: *Which transportation method did Whitelaw use to reach the #1?*
A_2: **Plane**

Data Difficulty Evaluation

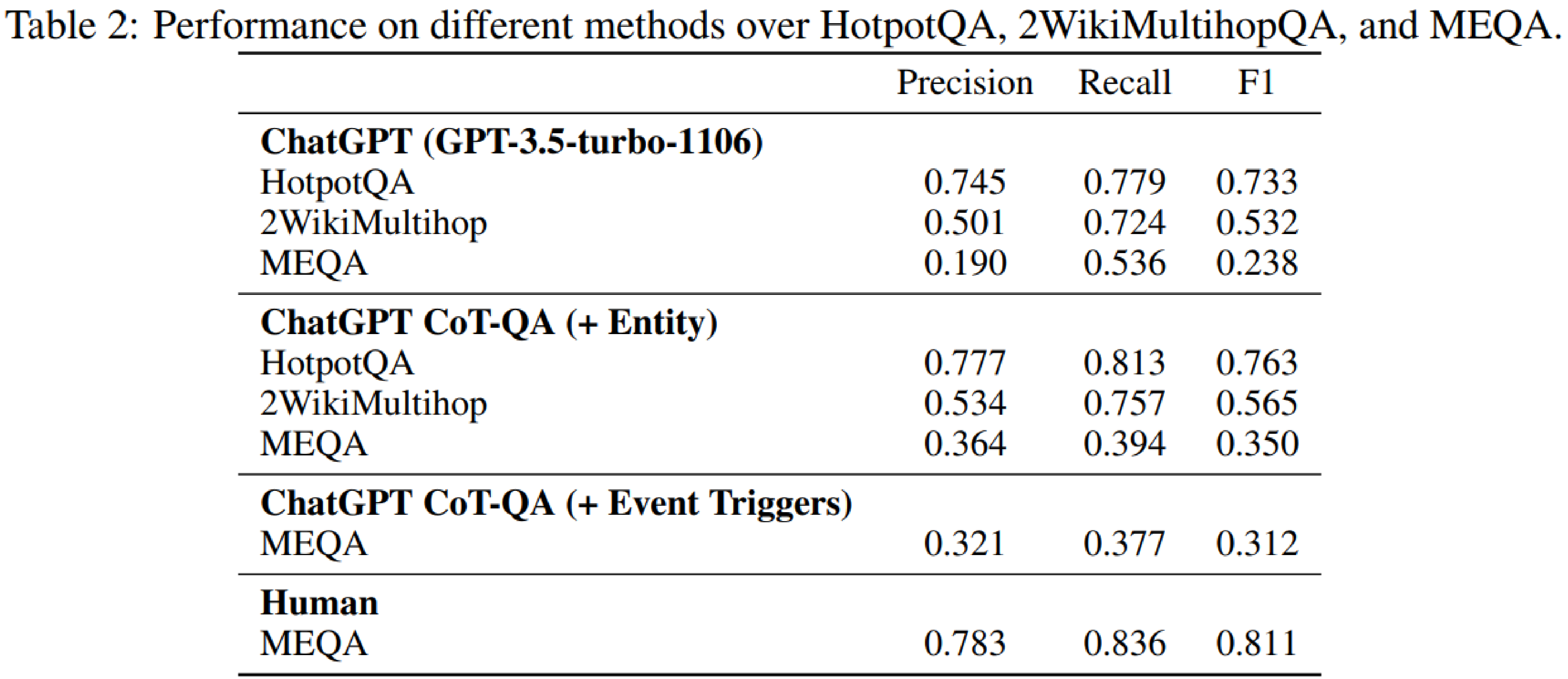


Table 2: Performance on different methods over HotpotQA, 2WikiMultihopQA, and MEQA.

	Precision	Recall	F1
ChatGPT (GPT-3.5-turbo-1106)			
HotpotQA	0.745	0.779	0.733
2WikiMultihop	0.501	0.724	0.532
MEQA	0.190	0.536	0.238
ChatGPT CoT-QA (+ Entity)			
HotpotQA	0.777	0.813	0.763
2WikiMultihop	0.534	0.757	0.565
MEQA	0.364	0.394	0.350
ChatGPT CoT-QA (+ Event Triggers)			
MEQA	0.321	0.377	0.312
Human			
MEQA	0.783	0.836	0.811

Experiment Results

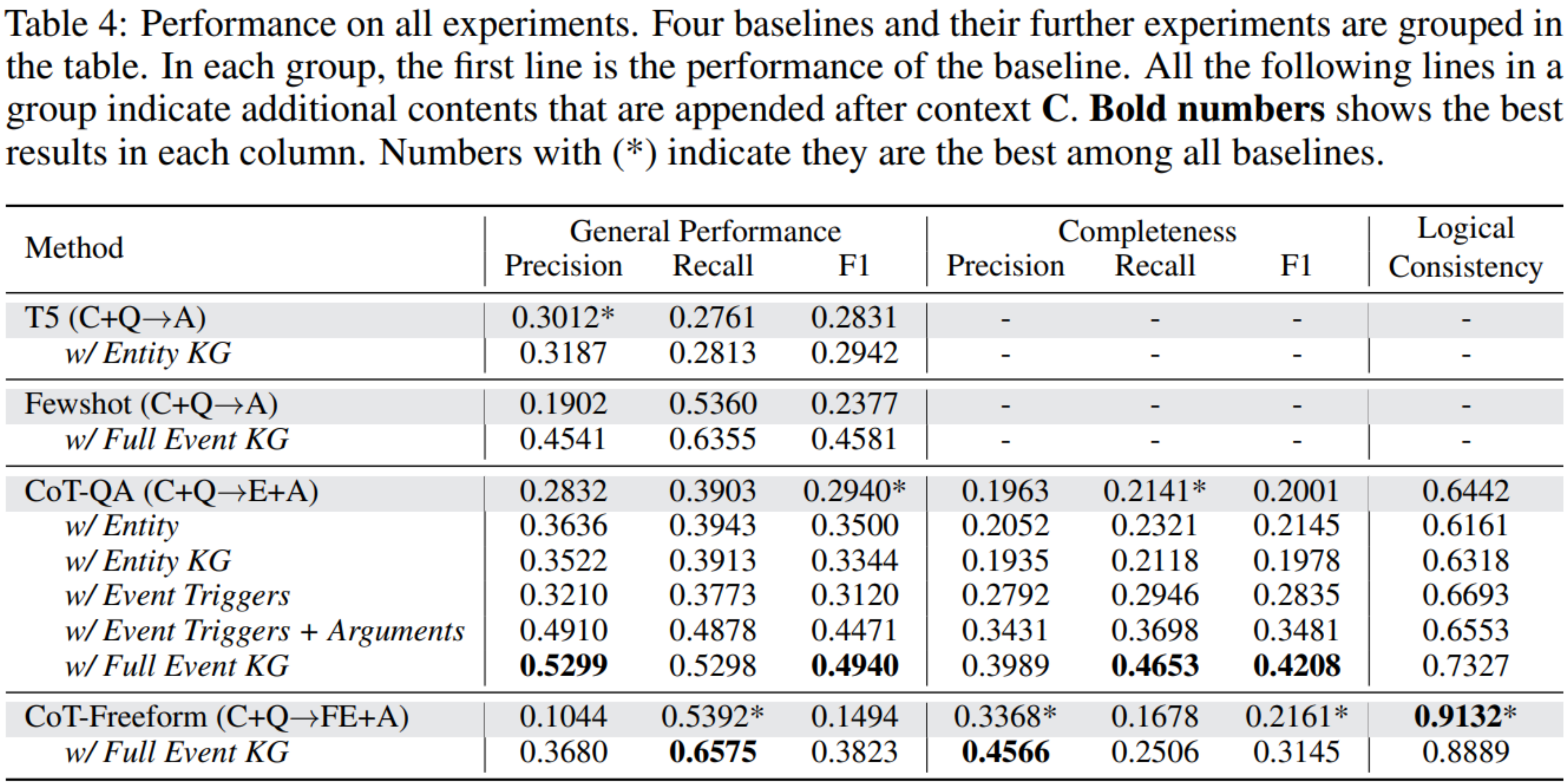


Table 4: Performance on all experiments. Four baselines and their further experiments are grouped in the table. In each group, the first line is the performance of the baseline. All the following lines in a group indicate additional contents that are appended after context **C**. **Bold numbers** shows the best results in each column. Numbers with (*) indicate they are the best among all baselines.

Method	General Performance			Completeness			Logical
	Precision	Recall	F1	Precision	Recall	F1	Consistency
T5 (C+Q→A)	0.3012*	0.2761	0.2831	-	-	-	-
w/ Entity KG	0.3187	0.2813	0.2942	-	-	-	-
Fewshot (C+Q→A)	0.1902	0.5360	0.2377	-	-	-	-
w/ Full Event KG	0.4541	0.6355	0.4581	-	-	-	-
CoT-QA (C+Q→E+A)	0.2832	0.3903	0.2940*	0.1963	0.2141*	0.2001	0.6442
w/ Entity	0.3636	0.3943	0.3500	0.2052	0.2321	0.2145	0.6161
w/ Entity KG	0.3522	0.3913	0.3344	0.1935	0.2118	0.1978	0.6318
w/ Event Triggers	0.3210	0.3773	0.3120	0.2792	0.2946	0.2835	0.6693
w/ Event Triggers + Arguments	0.4910	0.4878	0.4471	0.3431	0.3698	0.3481	0.6553
w/ Full Event KG	**0.5299**	0.5298	**0.4940**	0.3989	**0.4653**	**0.4208**	0.7327
CoT-Freeform (C+Q→FE+A)	0.1044	0.5392*	0.1494	0.3368*	0.1678	0.2161*	**0.9132***
w/ Full Event KG	0.3680	**0.6575**	0.3823	**0.4566**	0.2506	0.3145	0.8889

Experiment Results

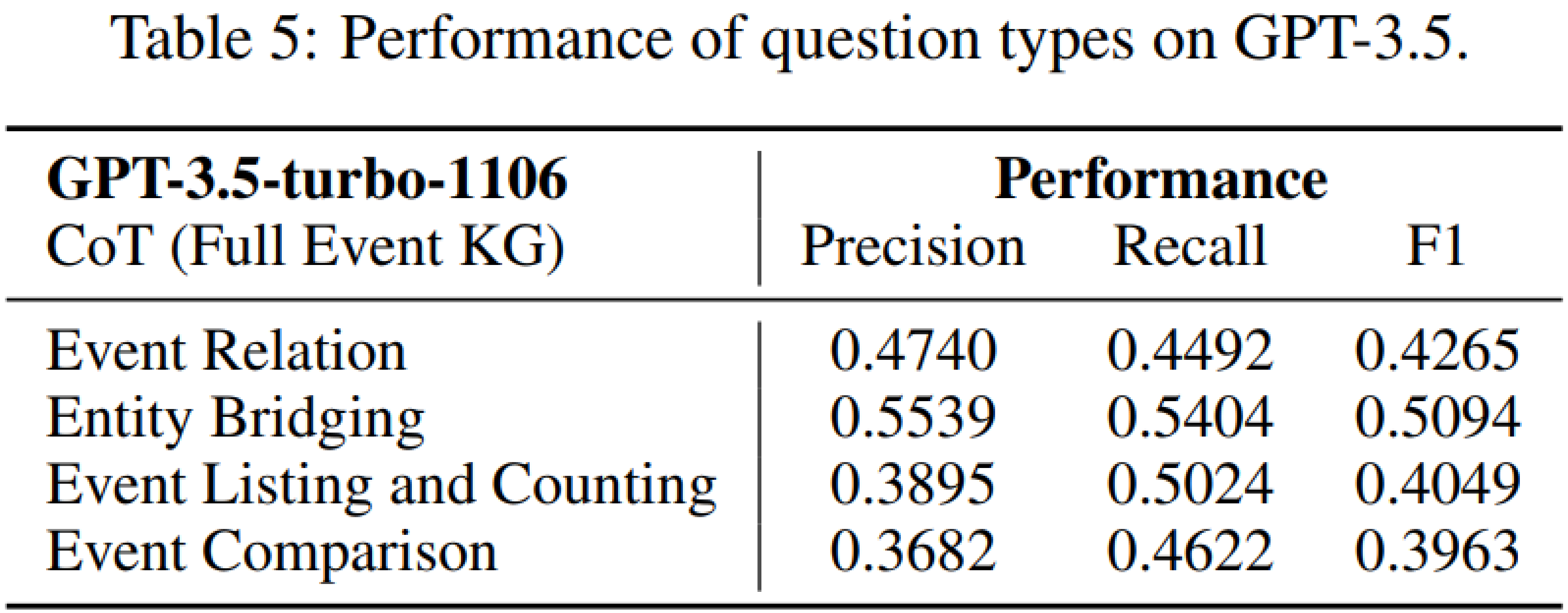


Table 5: Performance of question types on GPT-3.5.

GPT-3.5-turbo-1106		**Performance**	
CoT (Full Event KG)	Precision	Recall	F1
Event Relation	0.4740	0.4492	0.4265
Entity Bridging	0.5539	0.5404	0.5094
Event Listing and Counting	0.3895	0.5024	0.4049
Event Comparison	0.3682	0.4622	0.3963

Thank you
for hearing!