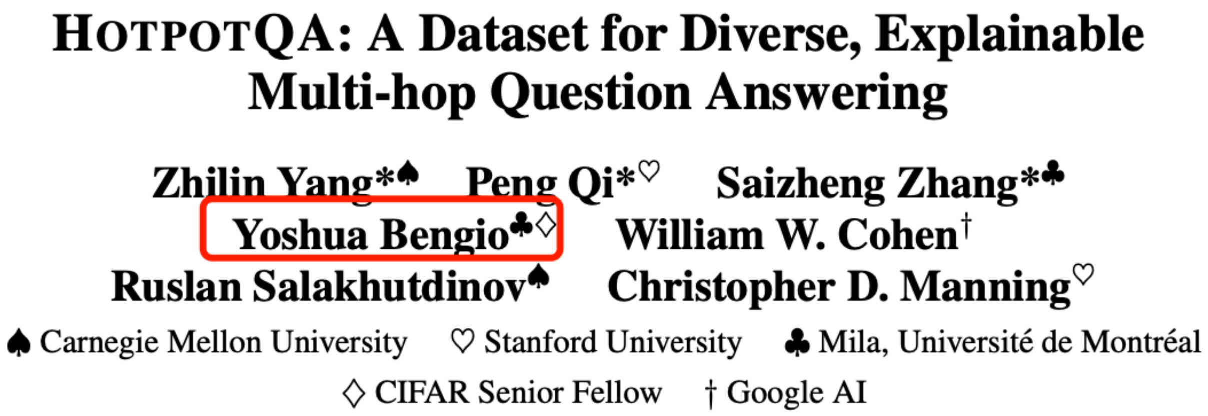
HOTPOTQA



HOTPOTQA: A Dataset for Diverse, Explainable Multi-hop Question Answering

Zhilin Yang[*♠] Peng Qi[*♡] Saizheng Zhang[*♣]
Yoshua Bengio[♣◇] William W. Cohen[†]
Ruslan Salakhutdinov[♠] Christopher D. Manning[♡]
♠ Carnegie Mellon University ♡ Stanford University ♣ Mila, Université de Montréal
◇ CIFAR Senior Fellow † Google AI

关键特点

1. 问题需要从多个支持文档中推理并找出答案；
2. 多样性：不局限于特定的知识库；
3. 提供句子级别的事实，支撑问答系统的推理和解释；
4. 新的问题类型：比较（comparison），需要问答系统对多个事实进行比较。“A和B的国籍是否相同？”（we require systems to compare two entities on some shared properties to test their understanding of both language and common concepts such as numerical magnitude.）

众包

1. 要求众包工作者从多个支持文档中提出需要在多个文档中进行推理的问题，并给出答案。
2. 确保提出的多跳问题更自然，不是针对于现有的知识库进行设计。
3. 要求众包工作者找出支撑事实。

数据格式

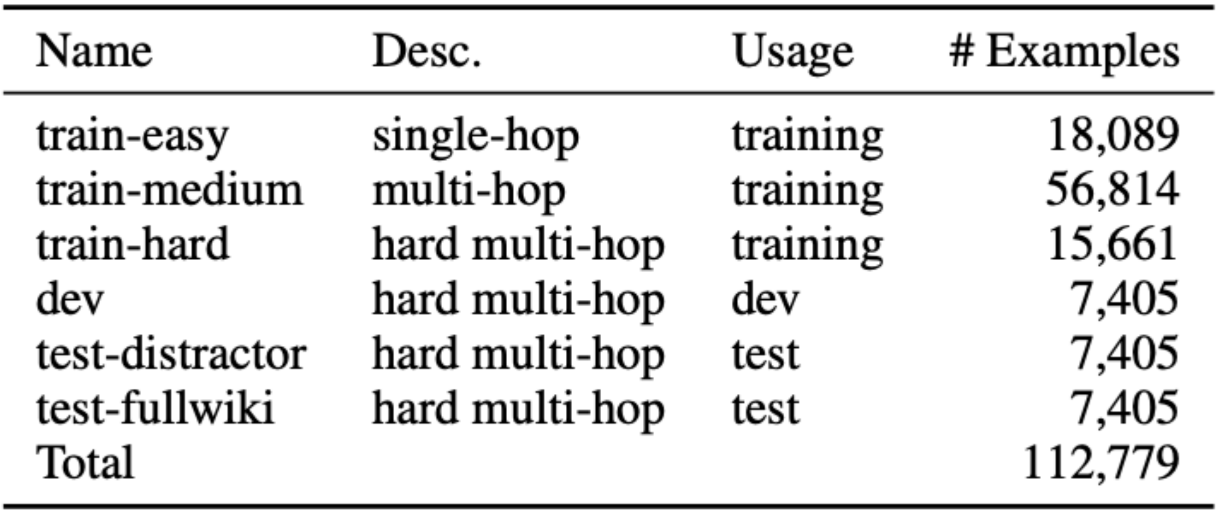
The top level structure of each JSON file is a list, where each entry represents a question-answer data point. Each data point is a dict with the following keys:

- `_id`: a unique id for this question-answer data point. This is useful for evaluation.
- `question`: a string.

- `answer`: a string. The test set does not have this key.
- `supporting_facts`: a list. Each entry in the list is a list with two elements `[title, sent_id]`, where `title` denotes the title of the paragraph, and `sent_id` denotes the supporting fact's id (0-based) in this paragraph. The test set does not have this key.
- `context`: a list. Each entry is a paragraph, which is represented as a list with two elements `[title, sentences]` and `sentences` is a list of strings.

There are other keys that are not used in our code, but might be used for other purposes (note that these keys are not present in the test sets, and your model should not rely on these two keys for making preditions on the test sets):

- `type`: either `comparison` or `bridge`, indicating the question type. (See our paper for more details).
- `level`: one of `easy`, `medium`, and `hard`. (See our paper for more details).

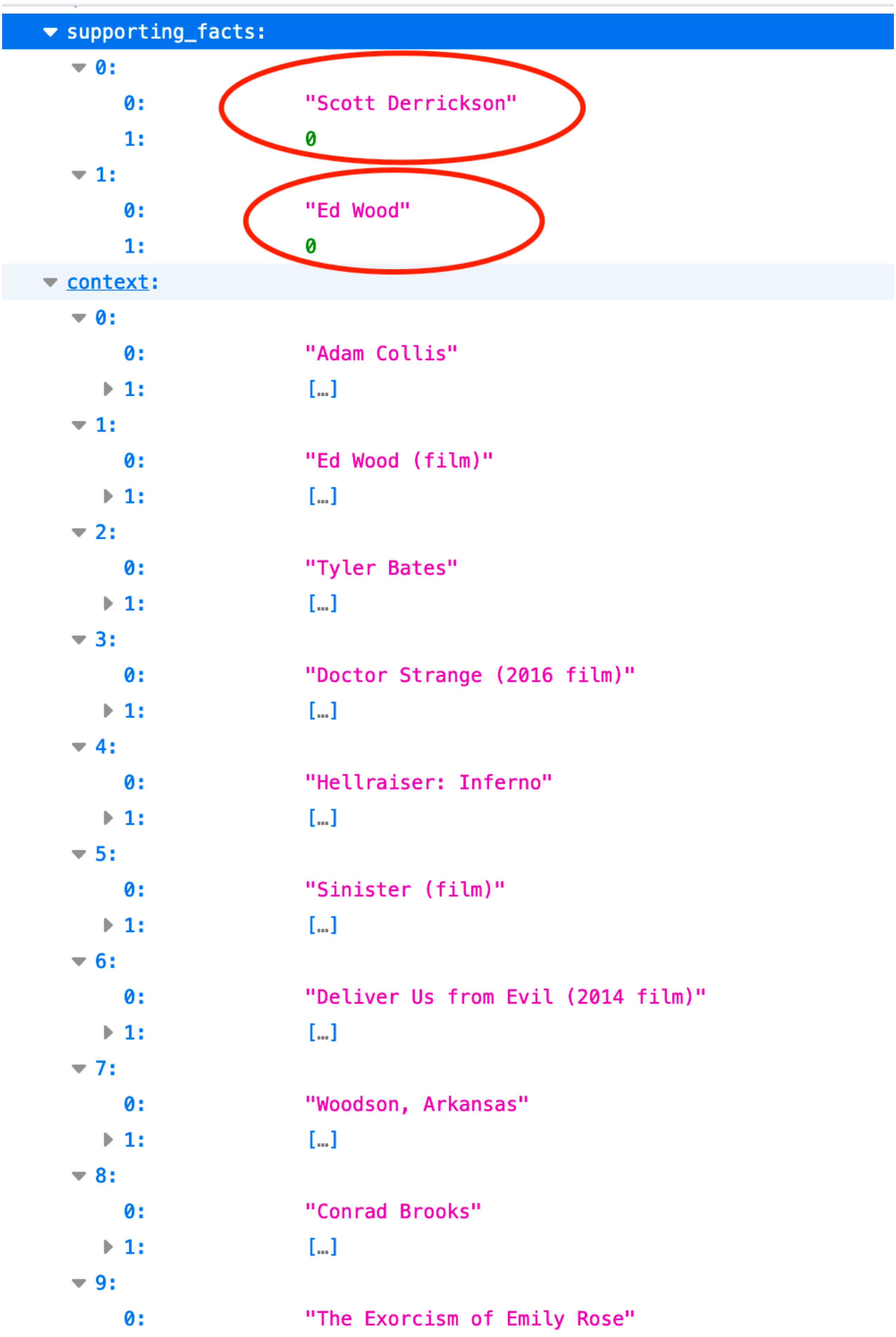


Name	Desc.	Usage	# Examples
train-easy	single-hop	training	18,089
train-medium	multi-hop	training	56,814
train-hard	hard multi-hop	training	15,661
dev	hard multi-hop	dev	7,405
test-distractor	hard multi-hop	test	7,405
test-fullwiki	hard multi-hop	test	7,405
Total			112,779

- 单跳问题：18089；多跳问题：94690；
- 单跳问题作为`train-easy`；
- 多跳问题中的baseline模型损失最小的前60%作为`train-medium`；
- 剩余40%的多跳问题**随机划分**为四个部分：(1)`train-hard` (2)`dev` (3)`test-distractor` (4) `test-fullwiki`；大约2:1:1:1;

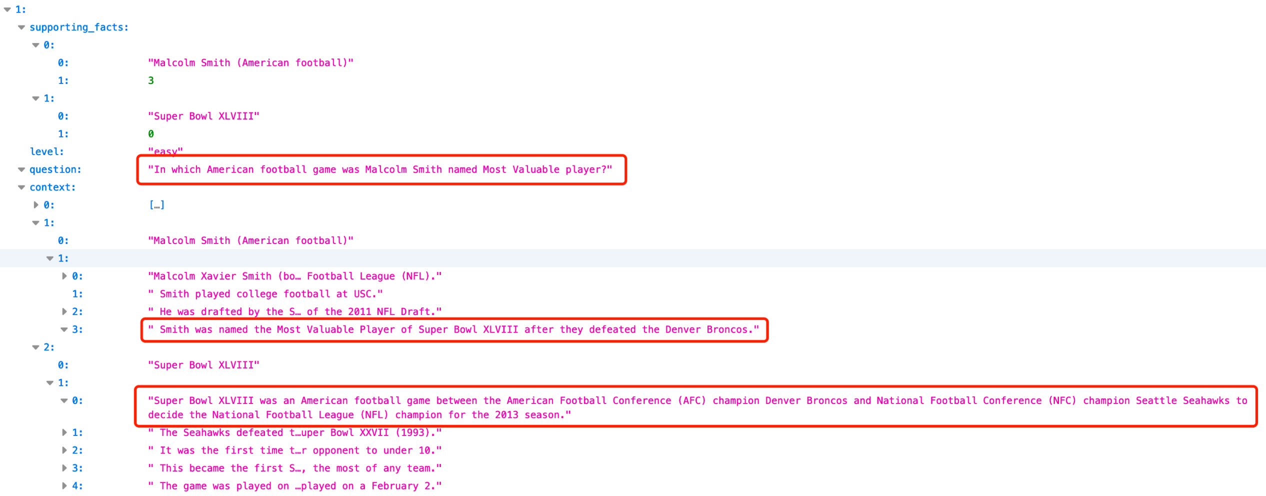
- The two test sets *test-distractor* and *test-fullwiki* are used in two different benchmark settings. 评价指标不同。
- **distractor**：以问题为查询，从维基百科上检索出的8段话作为干扰项，加上两段正确的gold paragraphs（用来让众包工作者生成问题和答案的段落）；
- **full wiki**：？？从所有维基百科中定位事实查找答案？？最后给出10个context，所以支撑事实不一定在当前问题的context中。 In the second setting, we fully test the model’s ability to locate relevant facts as well as reasoning about them by requiring it to answer the question given the first paragraphs of all Wikipedia articles without the gold paragraphs specified.
- **full wiki** 的支撑事实不一定在当前问题的context中。hotpot_dev_fullwiki_v1.json中有7405个数

据，其中支撑事实不在context中的有5316个。



```
▼ supporting_facts:
  ▼ 0:
      0:          "Scott Derrickson"
      1:          0
  ▼ 1:
      0:          "Ed Wood"
      1:          0
▼ context:
  ▼ 0:
      0:          "Adam Collis"
    ▶ 1:          […]
  ▼ 1:
      0:          "Ed Wood (film)"
    ▶ 1:          […]
  ▼ 2:
      0:          "Tyler Bates"
    ▶ 1:          […]
  ▼ 3:
      0:          "Doctor Strange (2016 film)"
    ▶ 1:          […]
  ▼ 4:
      0:          "Hellraiser: Inferno"
    ▶ 1:          […]
  ▼ 5:
      0:          "Sinister (film)"
    ▶ 1:          […]
  ▼ 6:
      0:          "Deliver Us from Evil (2014 film)"
    ▶ 1:          […]
  ▼ 7:
      0:          "Woodson, Arkansas"
    ▶ 1:          […]
  ▼ 8:
      0:          "Conrad Brooks"
    ▶ 1:          […]
  ▼ 9:
      0:          "The Exorcism of Emily Rose"
```

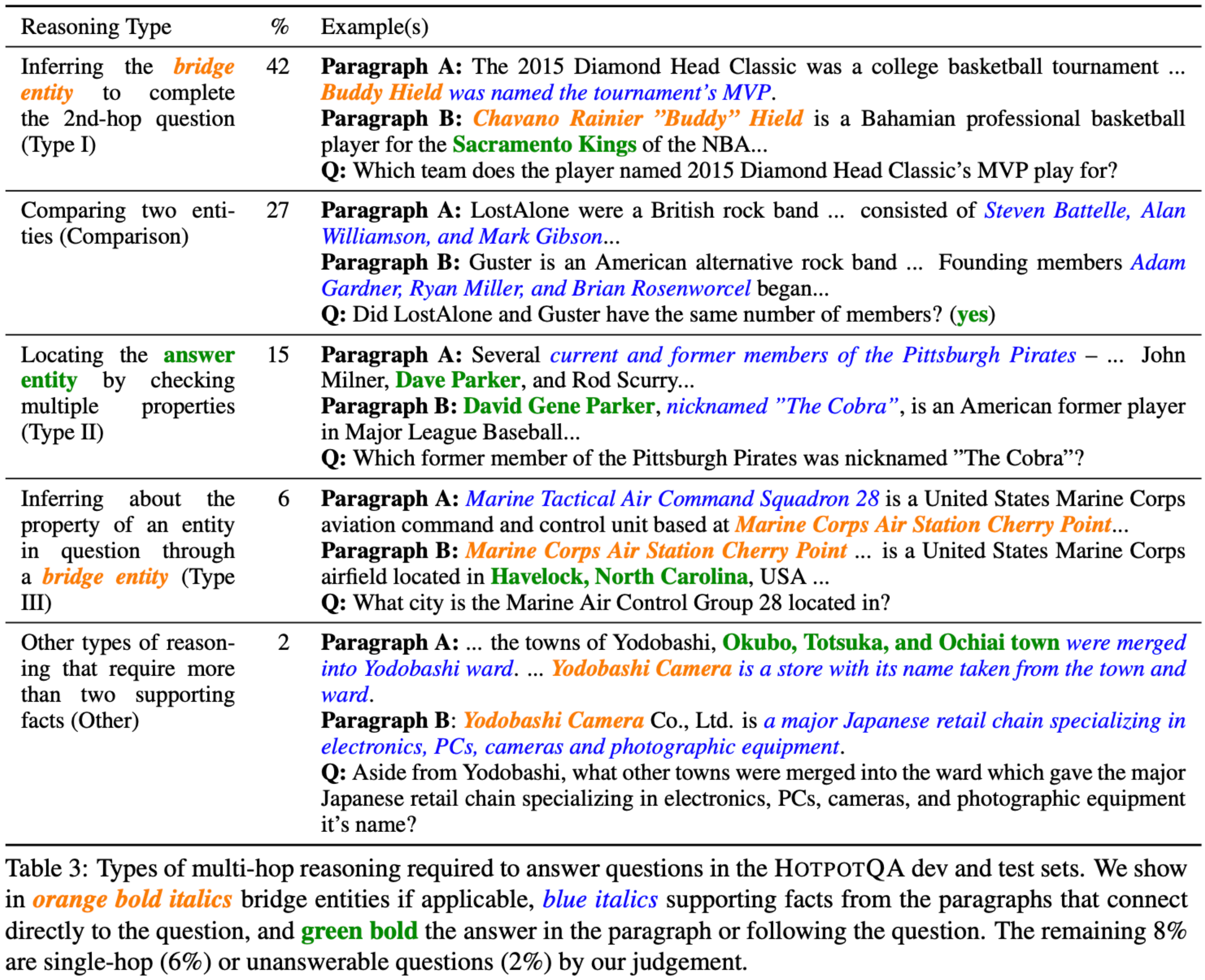
- `train-easy`: **single-hop questions**. an overwhelming percentage in the sample only required **reasoning over one of the paragraphs**.



```
▾ 1:
  ▾ supporting_facts:
    ▾ 0:
        0:            "Malcolm Smith (American football)"
        1:            3
    ▾ 1:
        0:            "Super Bowl XLVIII"
        1:            0
    level:            "easy"
  ▾ question:         "In which American football game was Malcolm Smith named Most Valuable player?"
  ▾ context:
    ▸ 0:              […]
    ▾ 1:
        0:            "Malcolm Smith (American football)"
      ▾ 1:
        ▸ 0:          "Malcolm Xavier Smith (bo… Football League (NFL)."
          1:          " Smith played college football at USC."
        ▸ 2:          " He was drafted by the S… of the 2011 NFL Draft."
        ▾ 3:          " Smith was named the Most Valuable Player of Super Bowl XLVIII after they defeated the Denver Broncos."
    ▾ 2:
        0:            "Super Bowl XLVIII"
      ▾ 1:
        ▾ 0:          "Super Bowl XLVIII was an American football game between the American Football Conference (AFC) champion Denver Broncos and National Football Conference (NFC) champion Seattle Seahawks to
                      decide the National Football League (NFL) champion for the 2013 season."
        ▸ 1:          " The Seahawks defeated t…uper Bowl XXVII (1993)."
        ▸ 2:          " It was the first time t…r opponent to under 10."
        ▸ 3:          " This became the first S…, the most of any team."
        ▸ 4:          " The game was played on …played on a February 2."
```

- `train-medium`: the models(baseline) were able to correctly answer 60% of the questions with high confidence. 损失最小的前60%的多跳问题。
-

问题推理类型



Reasoning Type	%	Example(s)
Inferring the ***bridge entity*** to complete the 2nd-hop question (Type I)	42	**Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... ***Buddy Hield*** *was named the tournament's MVP*. **Paragraph B:** ***Chavano Rainier ”Buddy” Hield*** is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA... **Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for?
Comparing two entities (Comparison)	27	**Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*... **Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began... **Q:** Did LostAlone and Guster have the same number of members? (**yes**)
Locating the **answer entity** by checking multiple properties (Type II)	15	**Paragraph A:** Several *current and former members of the Pittsburgh Pirates* – ... John Milner, **Dave Parker**, and Rod Scurry... **Paragraph B: David Gene Parker**, *nicknamed ”The Cobra”*, is an American former player in Major League Baseball... **Q:** Which former member of the Pittsburgh Pirates was nicknamed ”The Cobra”?
Inferring about the property of an entity in question through a ***bridge entity*** (Type III)	6	**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?
Other types of reasoning that require more than two supporting facts (Other)	2	**Paragraph A:** ... the towns of Yodobashi, **Okubo, Totsuka, and Ochiai town** *were merged into Yodobashi ward*. ... ***Yodobashi Camera*** *is a store with its name taken from the town and ward*. **Paragraph B**: ***Yodobashi Camera*** Co., Ltd. is *a major Japanese retail chain specializing in electronics, PCs, cameras and photographic equipment*. **Q:** Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese retail chain specializing in electronics, PCs, cameras, and photographic equipment it's name?

Table 3: Types of multi-hop reasoning required to answer questions in the HotpotQA dev and test sets. We show in ***orange bold italics*** bridge entities if applicable, *blue italics* supporting facts from the paragraphs that connect directly to the question, and **green bold** the answer in the paragraph or following the question. The remaining 8% are single-hop (6%) or unanswerable questions (2%) by our judgement.

1. `Type I` 两跳问题的尾节点（某比赛-[MVP]->(person)-[效率于]->（? 哪支球队）），其中（person）节点被称为*bridge entity*；
2. `Comparison` 比较两个实体的属性，是否相等（大于、小于）；
3. `Type II` 通过多个属性，确定一个实体，星状查询。
4. `Type I` 两跳问题的尾节点的属性；
5. `Other` 需要多于两个的支撑事实；