Retrieval Helps or Hurts? A Deeper Dive into the Efficacy of Retrieval Augmentation to Language Models

[Oral presentation]

Seiji Maekawa, Hayate Iso, Sairam Gurajada, Nikita Bhutani

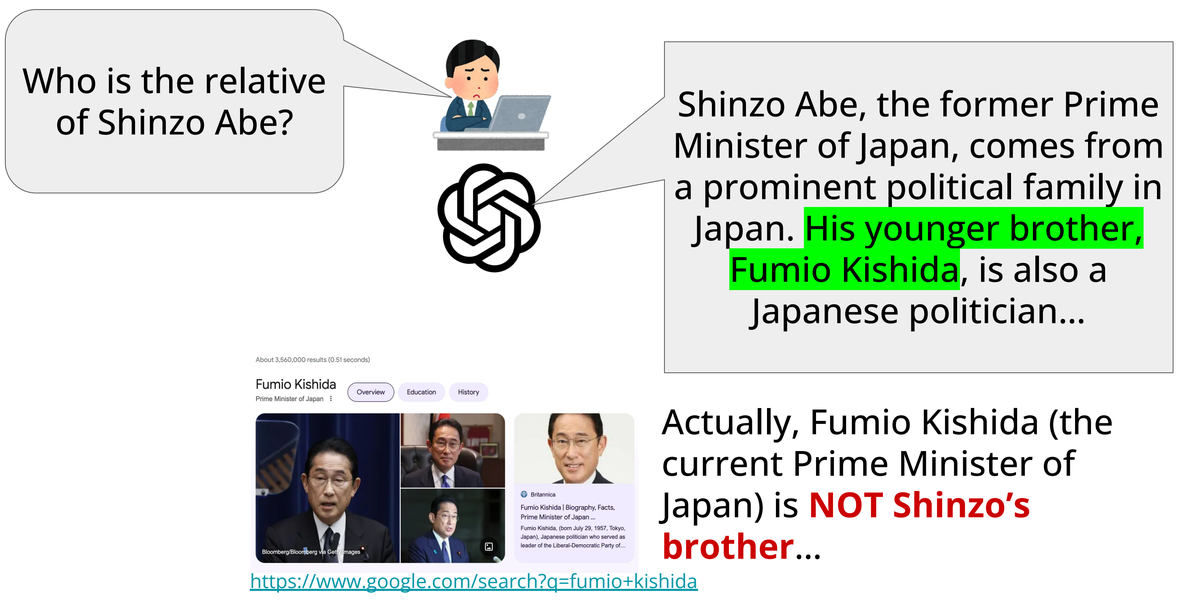
Megagon Labs

Key Findings

- Large language models (LMs) can recall frequently encountered entity-relation pairs without retrieval. However, **their performance drops significantly for minor facts**.
- Retrievers perform better than LMs for **long-tailed entity-relation pairs**. However, this does not apply to well-known pairs **(knowledge override)**.
- LMs achieve higher accuracy than retrievers for **well-known entity-relation pairs concerning long-tailed entities**, even though previous studies indicate large LMs struggle with these entities.

Problem: Entity Popularity May Not be Enough

Example of popular entity but minor fact:



Actually, Fumio Kishida (the current Prime Minister of Japan) is **NOT Shinzo's brother**...

https://www.google.com/search?q=fumio+kishida

To deeply understand LLMs, we aim to evaluate them with "**fact-centric**" datasets.

WiTQA dataset

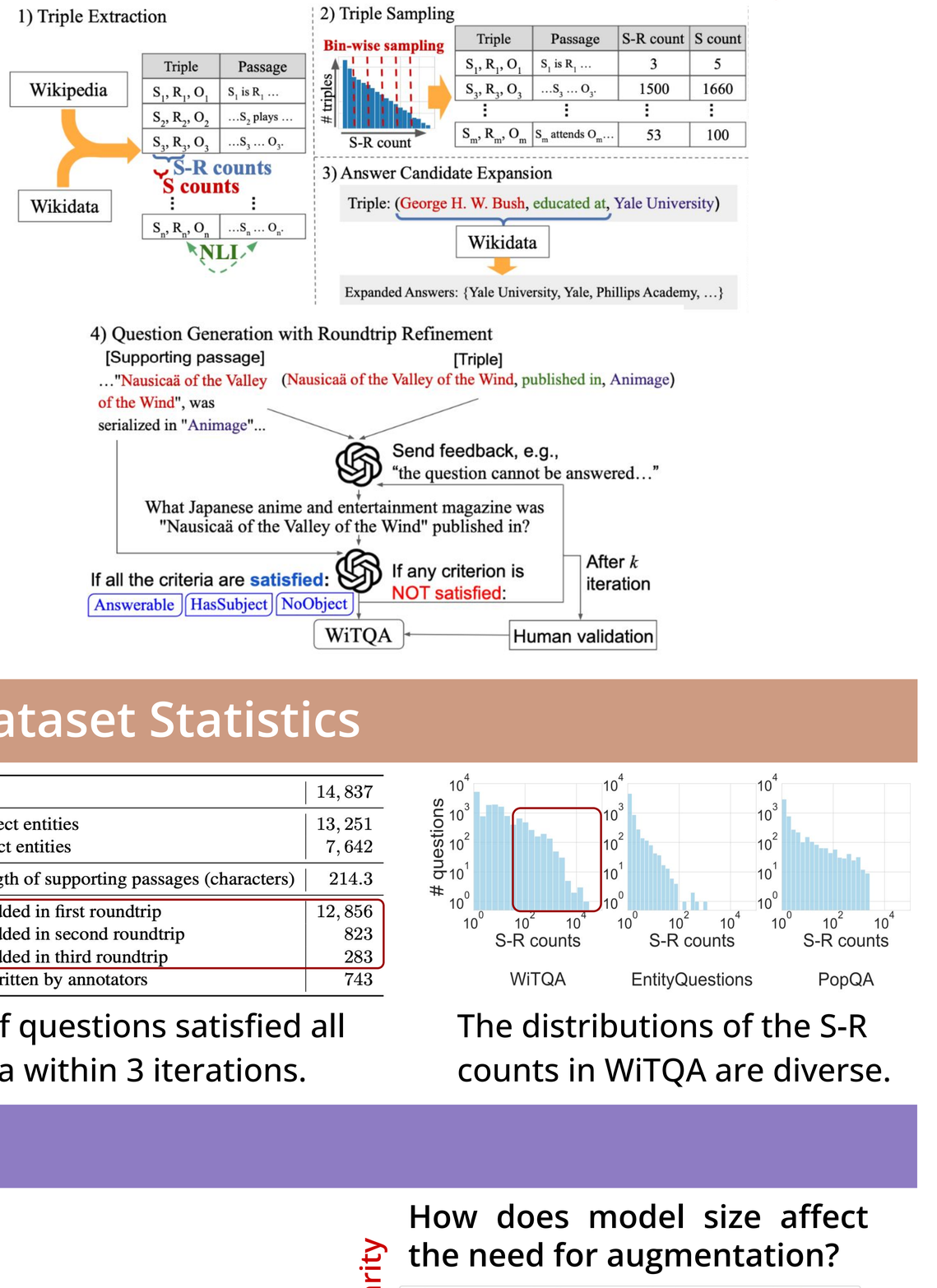
Dataset	Page view	S count	S-R count	Supporting passages	# of Relation Type	Question form
EntityQuestions (Sciavolino et al., 2021)	✗	✗	✗	✗	24	Template
PopQA (Mallen et al., 2023)	✓	✗	✗	✗	16	Template
WiTQA (Ours)	✓	✓	✓	✓	32	Model-assisted

Entity level popularity — Fact level popularity

Triple: (The Criminal Code, screenwriter, Fred Niblo, Jr.)
Subject — Relation — Object

Verbalize

Question example: Who was the screenwriter for The Criminal Code?
Relation — Subject



Dataset Statistics

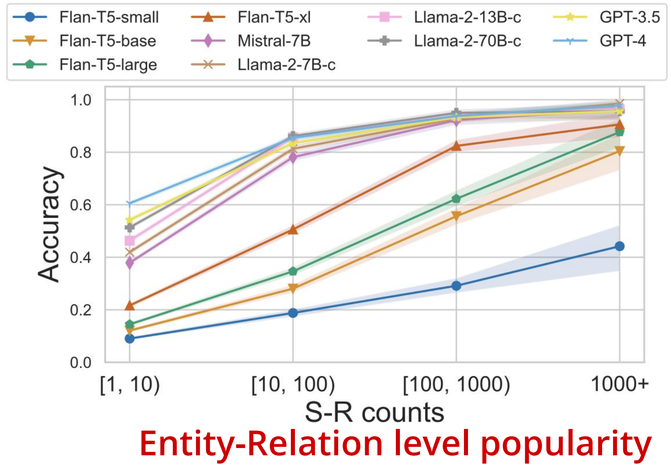
Questions	14,837
Unique subject entities	13,251
Unique object entities	7,642
Average length of supporting passages (characters)	214.3
Questions added in first roundtrip	12,856
Questions added in second roundtrip	823
Questions added in third roundtrip	283
Questions written by annotators	743

95% of questions satisfied all criteria within 3 iterations.

The distributions of the S-R counts in WiTQA are diverse.

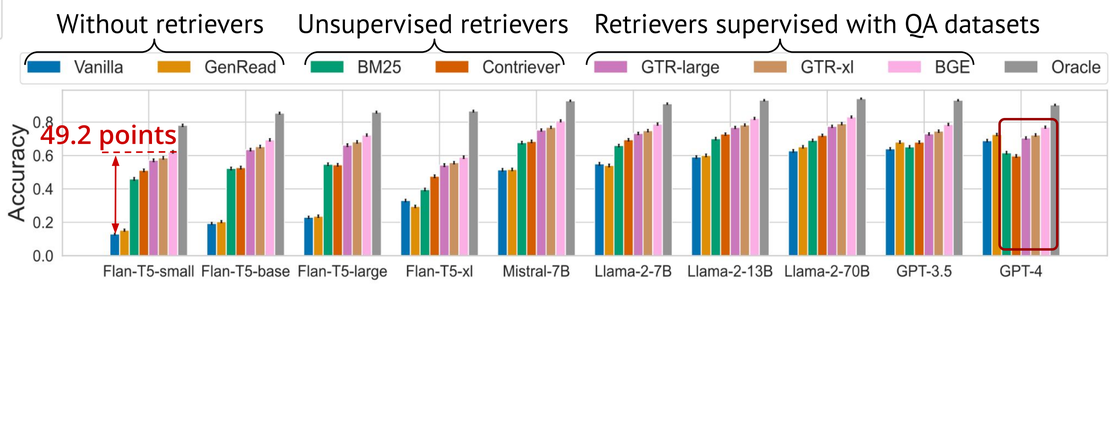
Experiments: Recall or Retrieve

Analysis of model's recall ability over entity-relation level popularity



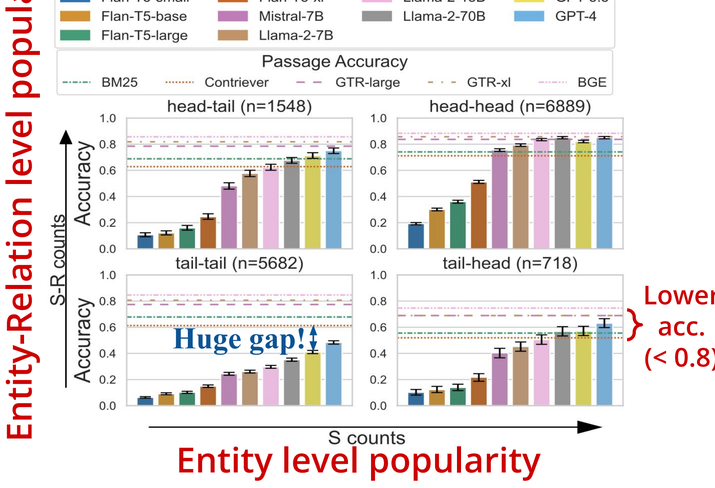
- Generally, all models demonstrate good recall of popular facts.

When do retrievers help?



- Retrieval augmentation enhances model performance, particularly for small models.
- Larger models often avoid answering when the retrieved passages are insufficient

How does model size affect the need for augmentation?



- Retrievers face challenges in identifying a specific fact when numerous passages contain references to the related entities.
- For the tail-tail group, retrieval augmentation always helps.

References:

- Simple Entity-Centric Questions Challenge Dense Retrievers (Sciavolino et al., EMNLP 2021)
- When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Mallen et al., ACL 2023)

For any questions, email: seiji@megagon.ai

Dataset link (Github)