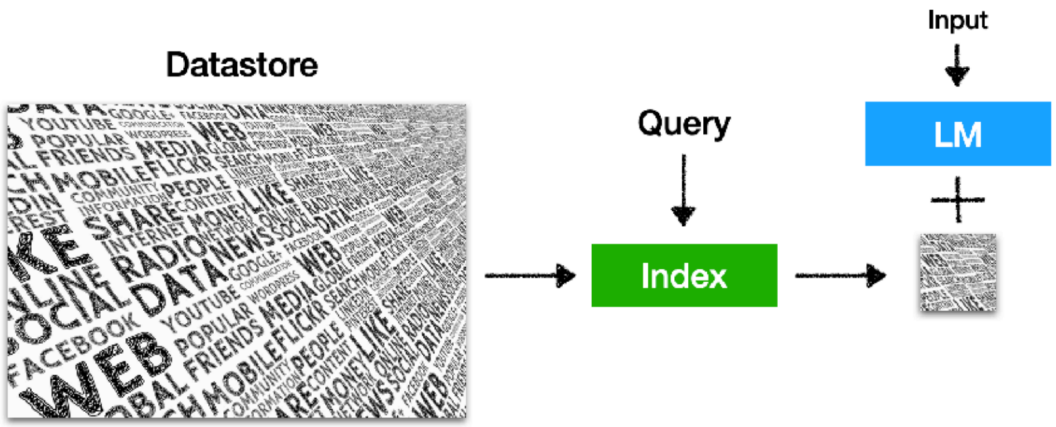
BRIEF: Bridging Retrieval and Inference for Multi-hop Reasoning via Compression

Yuankai Li*, Jia-Chen Gu*, Di Wu, Kai-Wei Chang, Nanyun Peng

Project page: https://jasonforjoy.github.io/BRIEF

Retrieval Augmented Generation

- Retriever $P(z|\ x)$: Return the top-K retrieved passages z given a query x
- Generator $P(y|\ x, z)$: Generate the target y based on the input x and the retrieved passages z

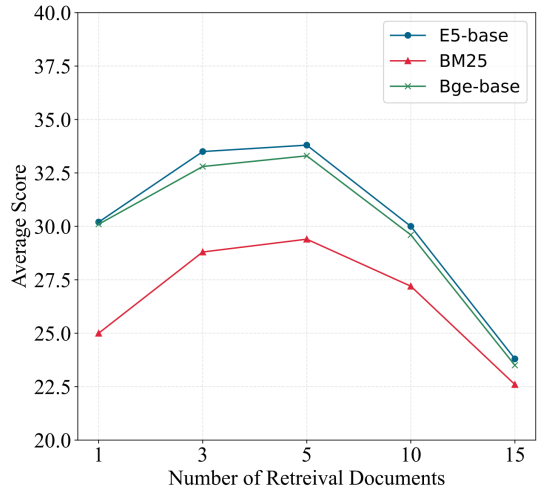


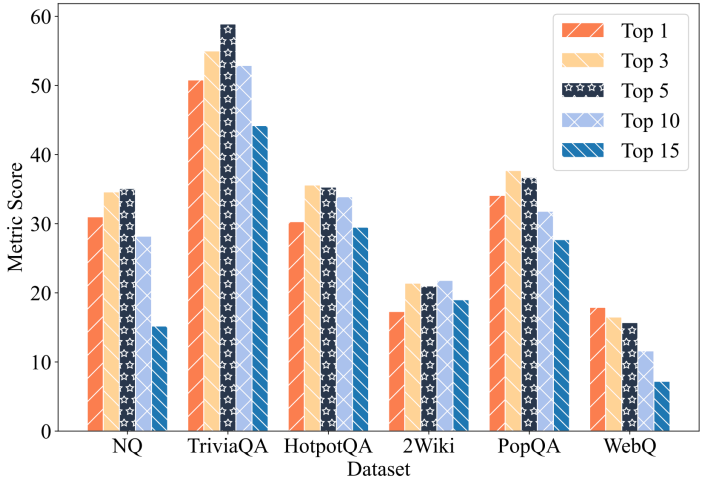
- ✓Efficiently incorporate the non-parametric knowledge into parametric LLMs
- ✓Individualization of private data
- ✓Low training and inference costs

Patrick Lewis, et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* NeurIPS 2020.

As the number of retrieved documents increases

- Input length grows linearly -> substantial increases in latency and inference costs
- Prone to introducing noise -> confusing LLMs and degrading long-context understanding
- *lost-in-the-middle* challenge -> underutilizing critical details buried deep in the middle

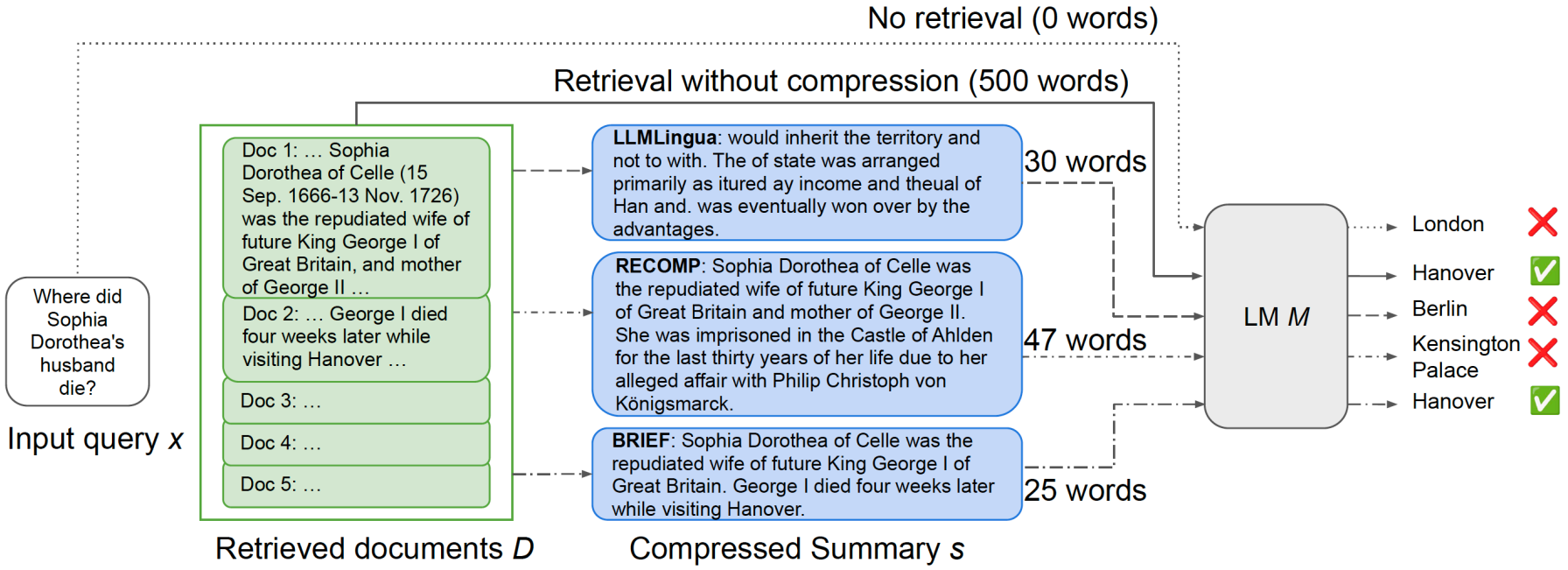




Jiajie Jin, et al. *FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research.* 2024.

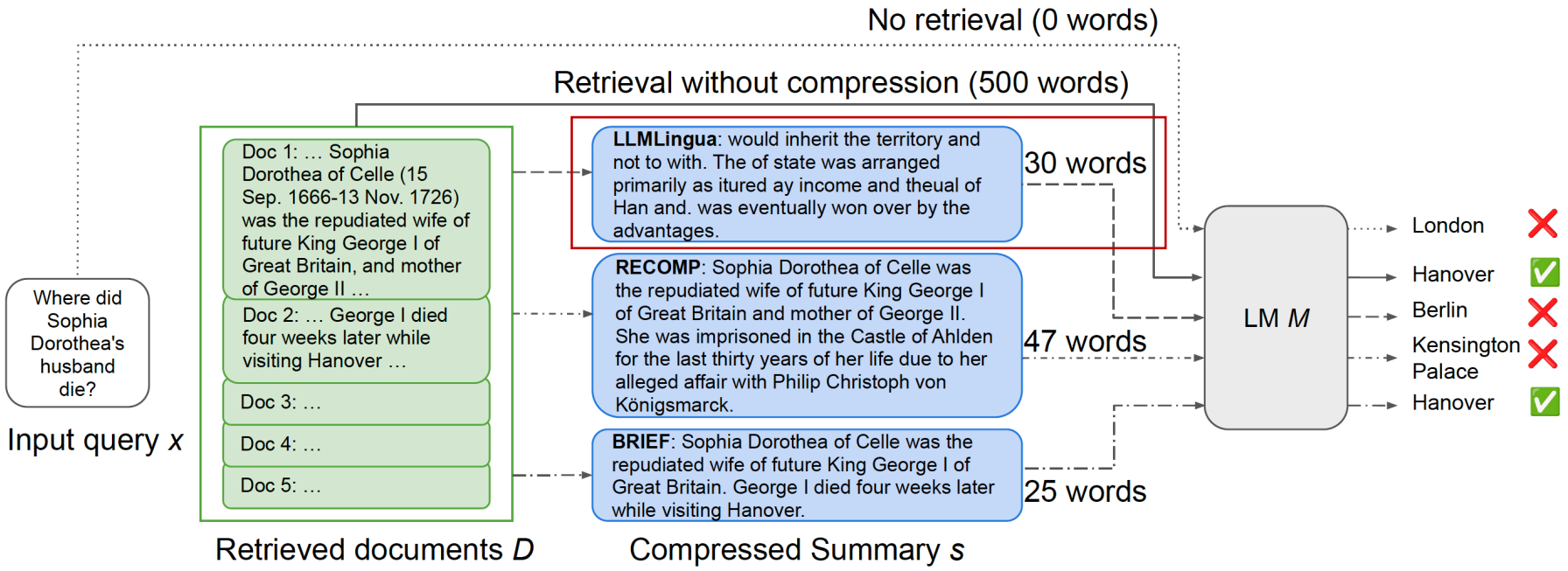
Compress Long-context into Textual Prompts

- Applicable to black-box LMs
- Reasoning across documents is required to collect necessary evidence scattered throughout various positions of the documents



Sentence-, Phrase-, or Token-level Compression

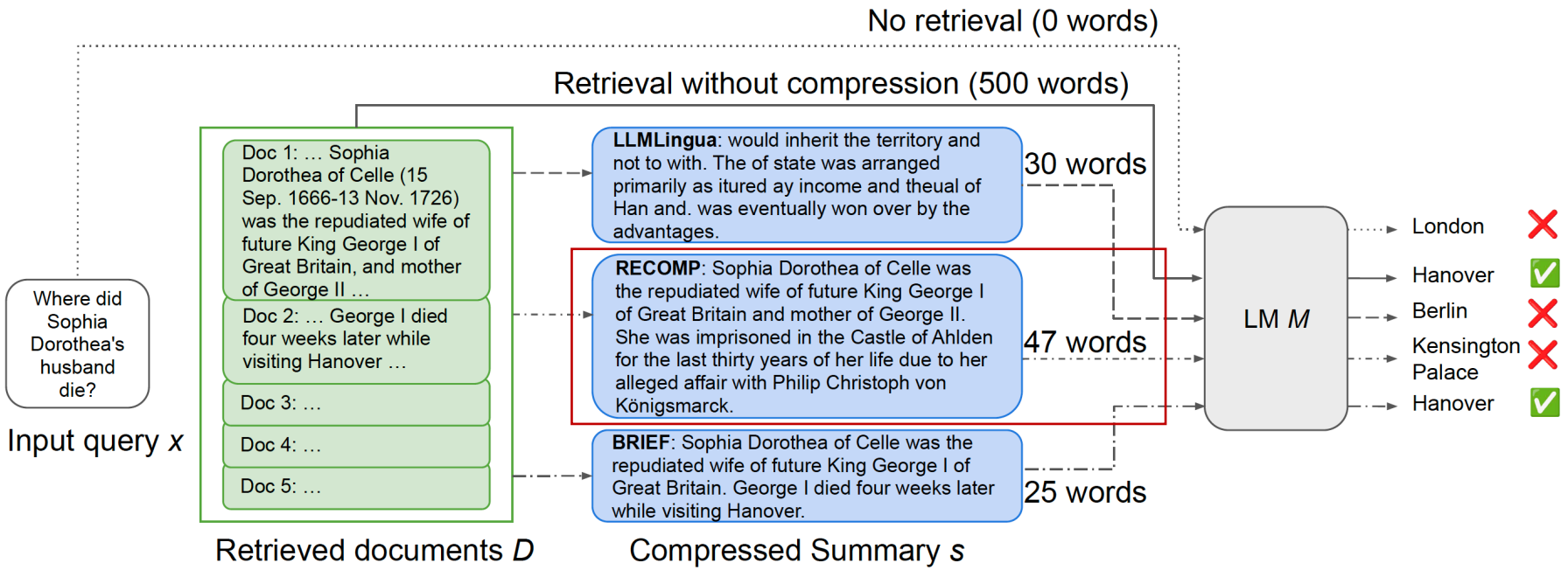
- Struggle to organize evidence in a natural language format -> ineffective for use by the follow-up reader LM



Huiqiang Jing, et al. *LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models.* EMNLP 2023.

Compression for Single-hop Queries

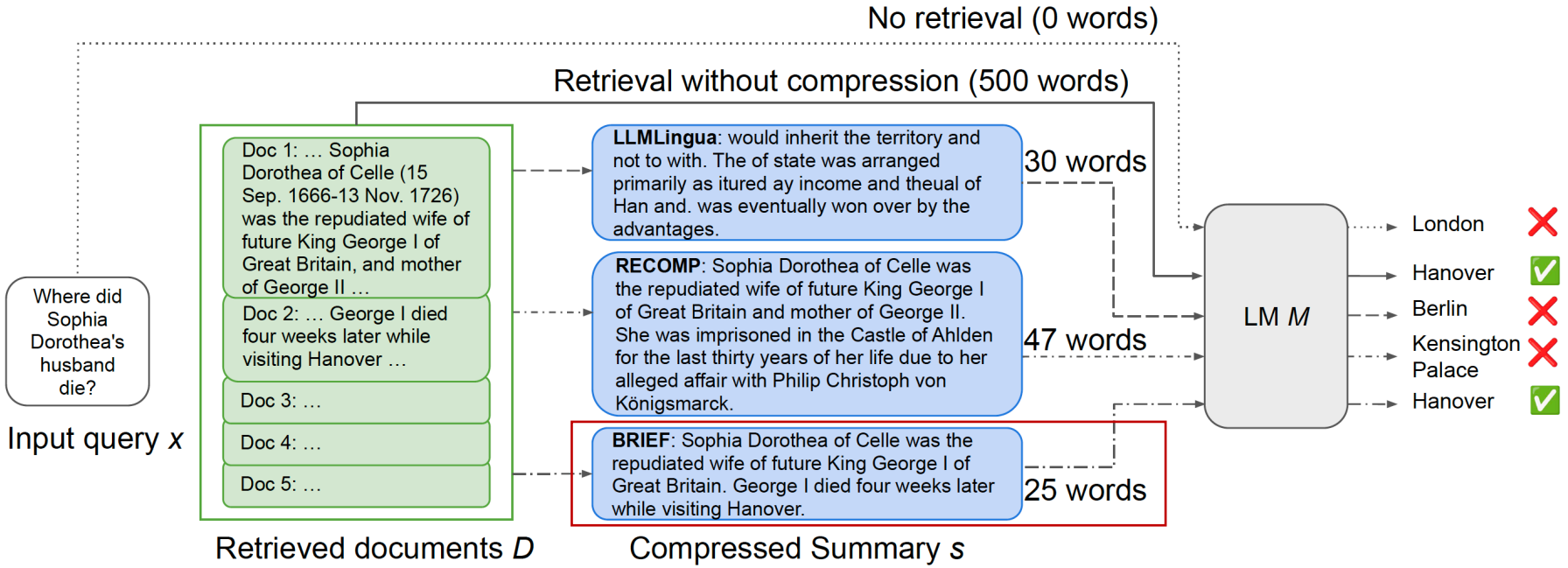
- Struggle to summarize the evidence from multiple documents for multi-hop questions, and the summary is lengthy



Fangyuan Xu, et al. *RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation.* ICLR 2024.

Challenge: Evidence Fusion for Multi-hop Queries

- BRIEF: **B**ridging **R**etrieval and **I**nference through **E**vidence **F**usion
- Query-aware multi-hop reasoning to compress retrieved documents into highly dense textual summaries



Overview of BRIEF

- Multi-hop reasoning-aware context compression
- Lightweight, T5-based compressor (770M) reduces costs by over 70%
- Cost-effective with synthetic data built entirely by open-source models
- Concise summaries enable a range of LMs to achieve exceptional open-domain QA performance

BRIEF Inference

- An off-the-shelf retriever returns a set of retrieved documents
- **COMPRESSOR** ([query; documents]) -> summary
  - ✓return an empty string if the retrieved documents are considered irrelevant
- An off-the-shelf LM reads [query; compressed summary]

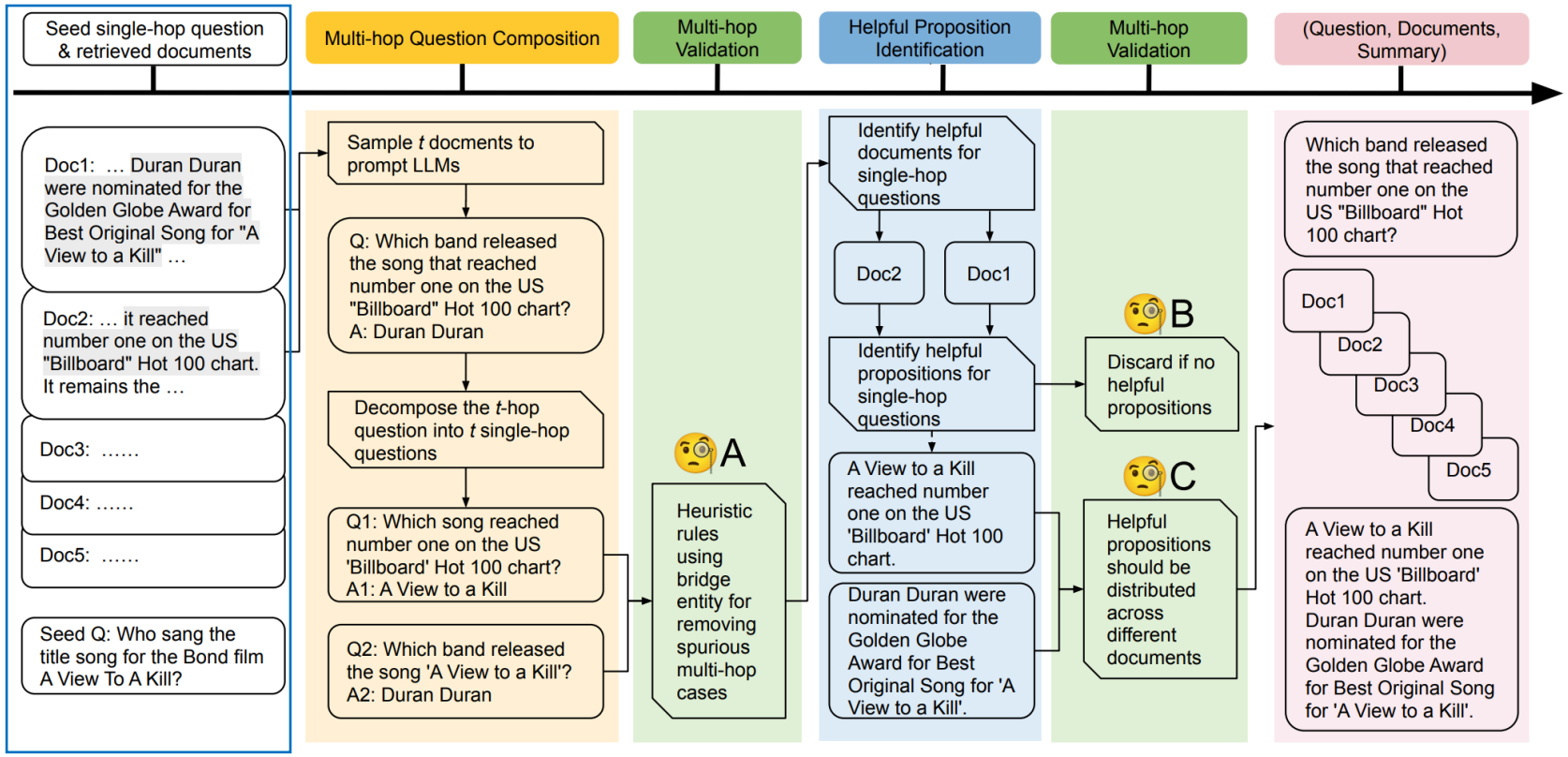
Challenges of Data Collection for Training

1) Collecting human annotations
2) Distilling the summarization knowledge of proprietary LLMs

- Expensive, not reproduce-friendly and impractical to scale up data generation due to the high costs

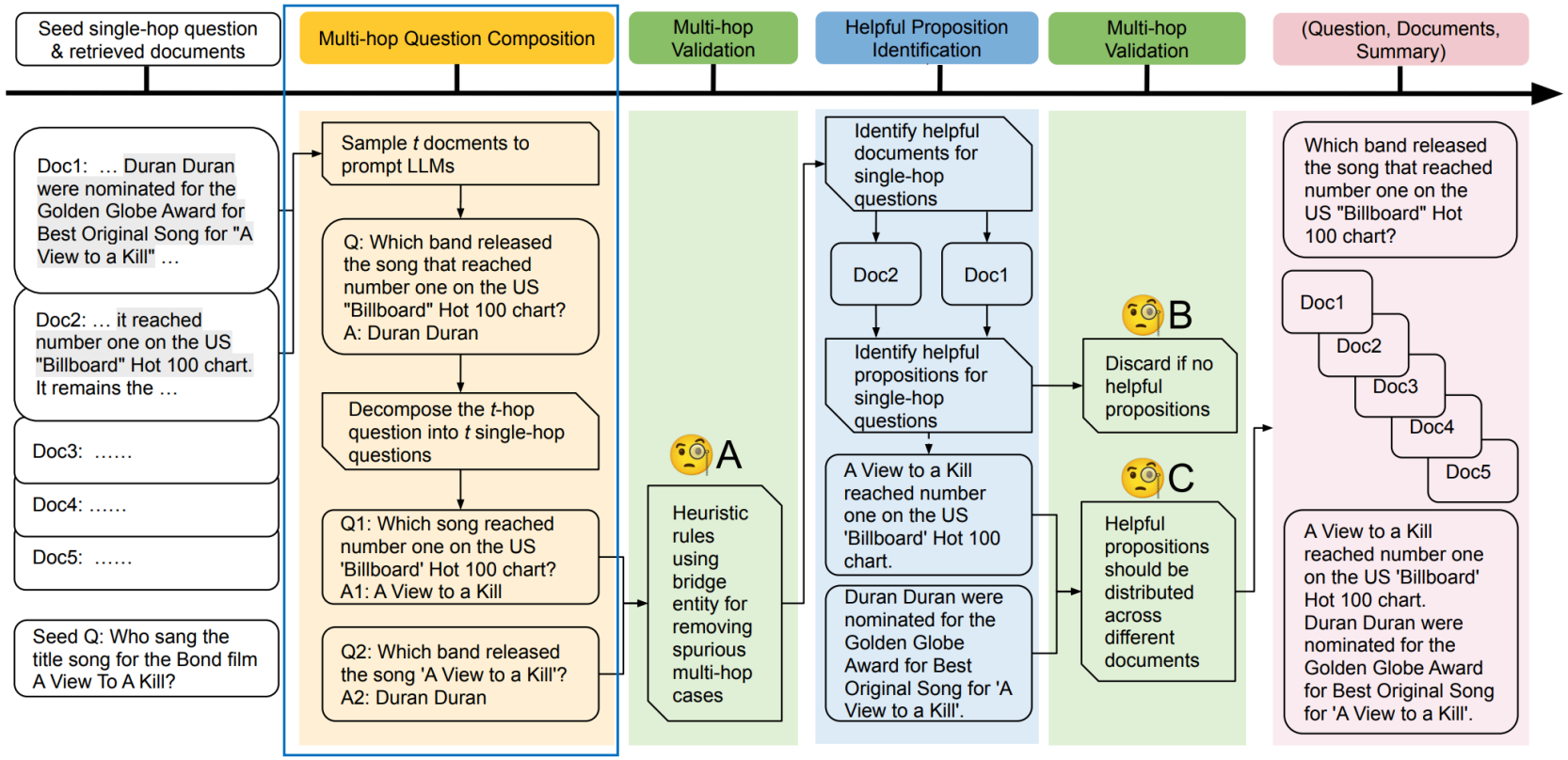
Start with the Seed Single-hop Q and Docs

- Single-hop questions can be easily obtained from existing datasets



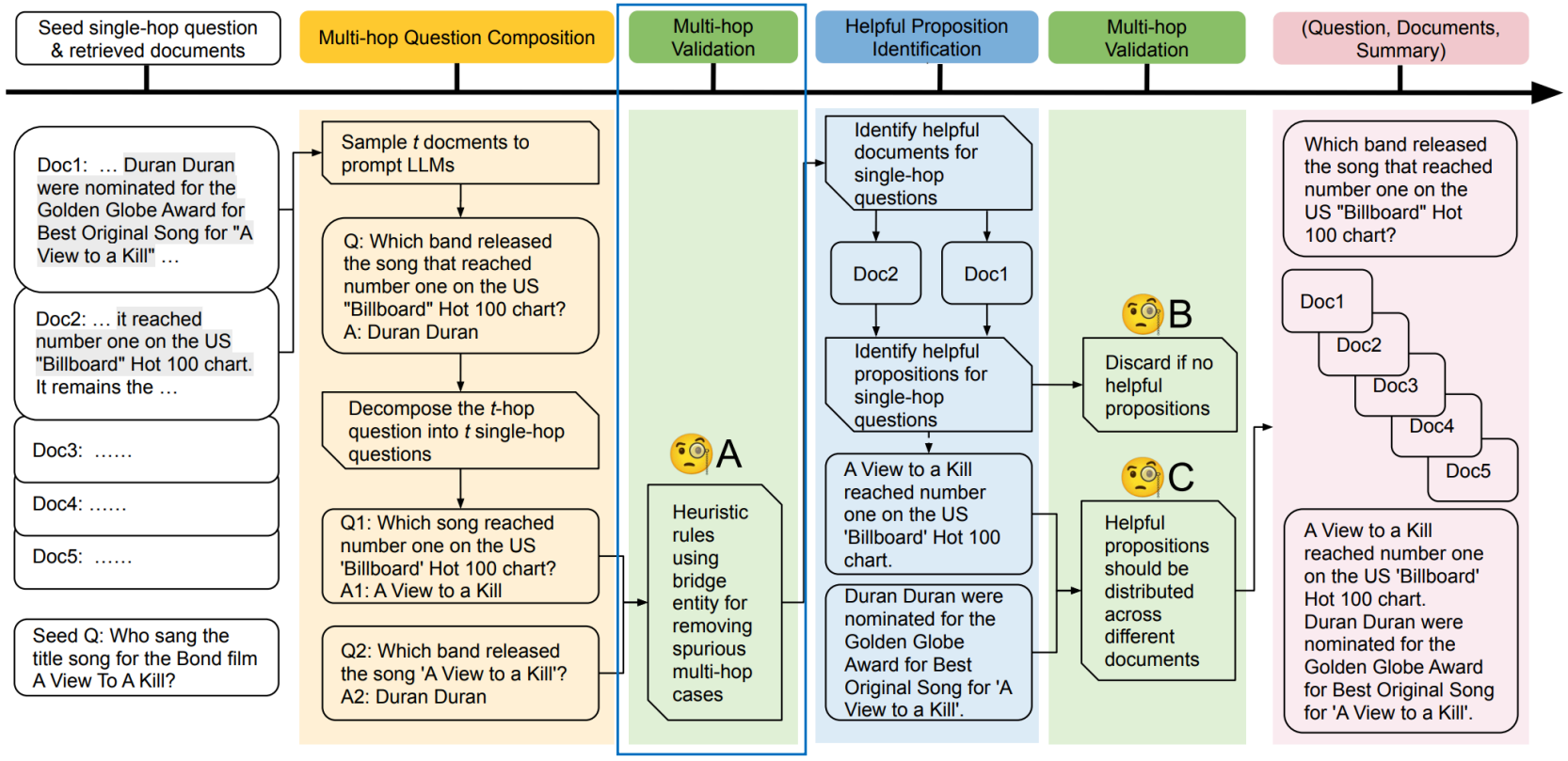
Multi-hop Question Composition

- *t* documents are randomly sampled
- Prompt LLMs to compose a *t*-hop question and its answer



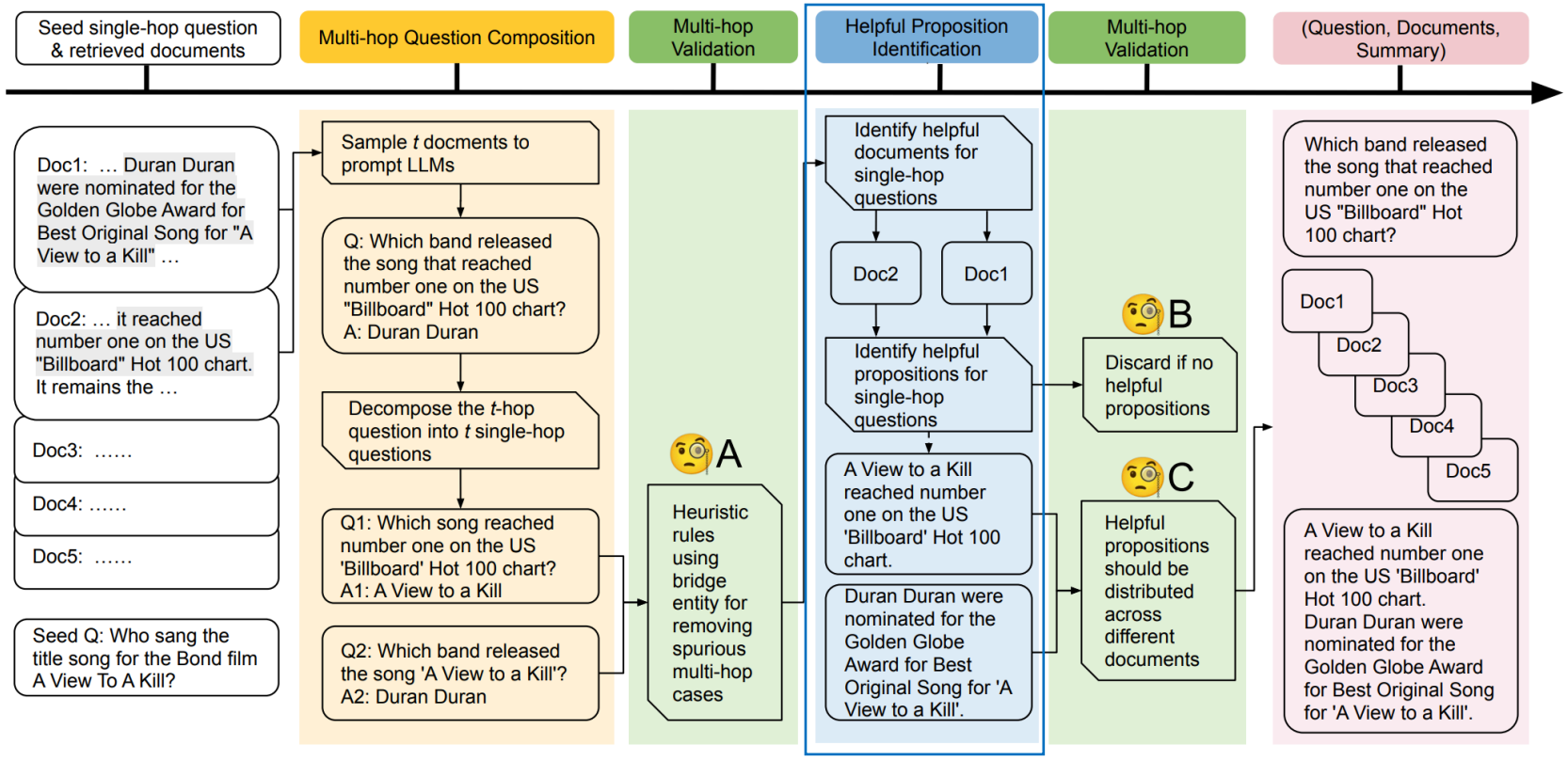
Multi-hop Validation

- May appear complex but fail to require genuinely multi-hop reasoning
- Automatic validation to eliminate spurious multi-hop questions



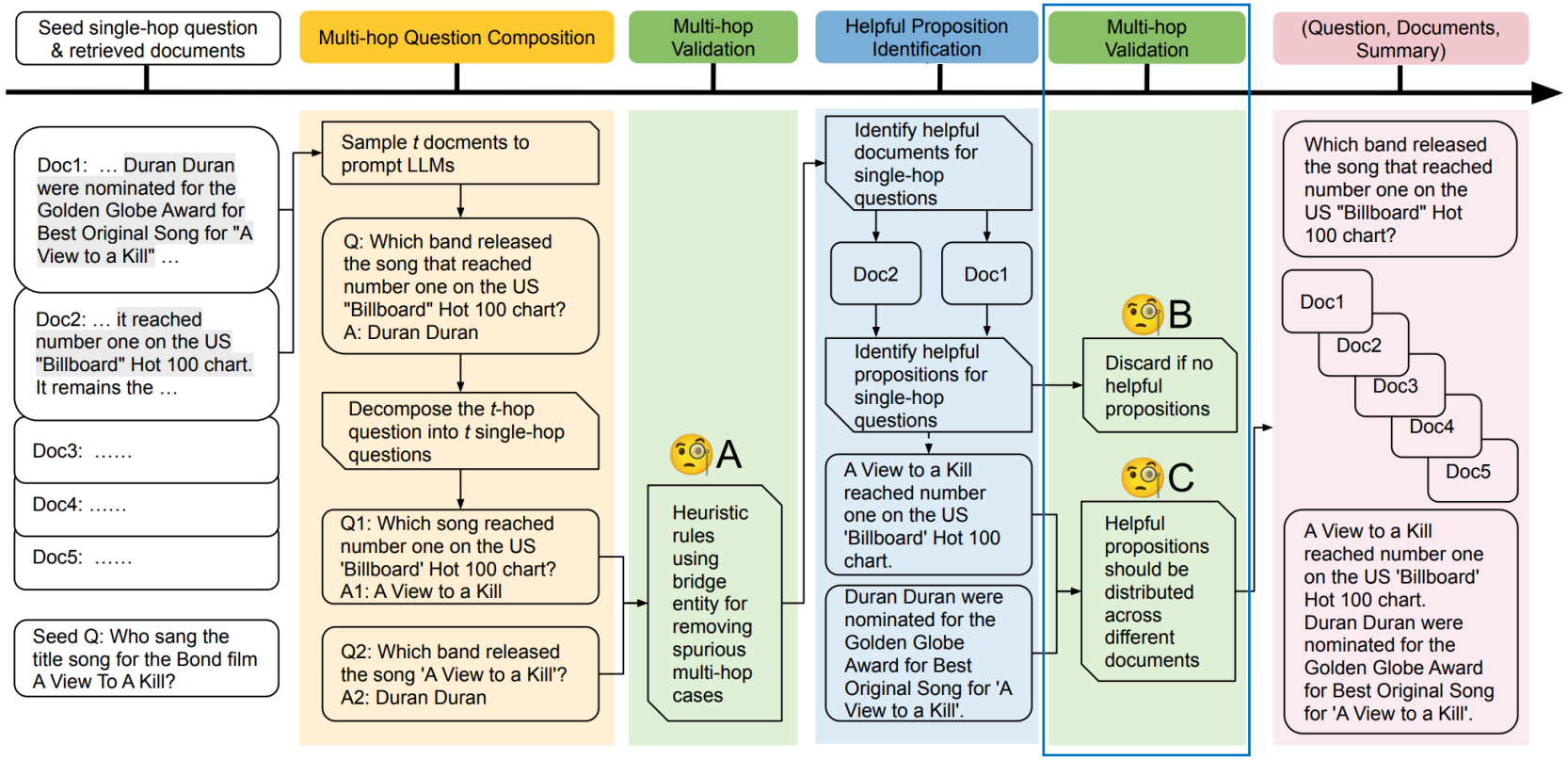
Helpful Proposition Identification

- Identify the most helpful evidence propositions for decomposed single-hop questions (will be used later for target summary)



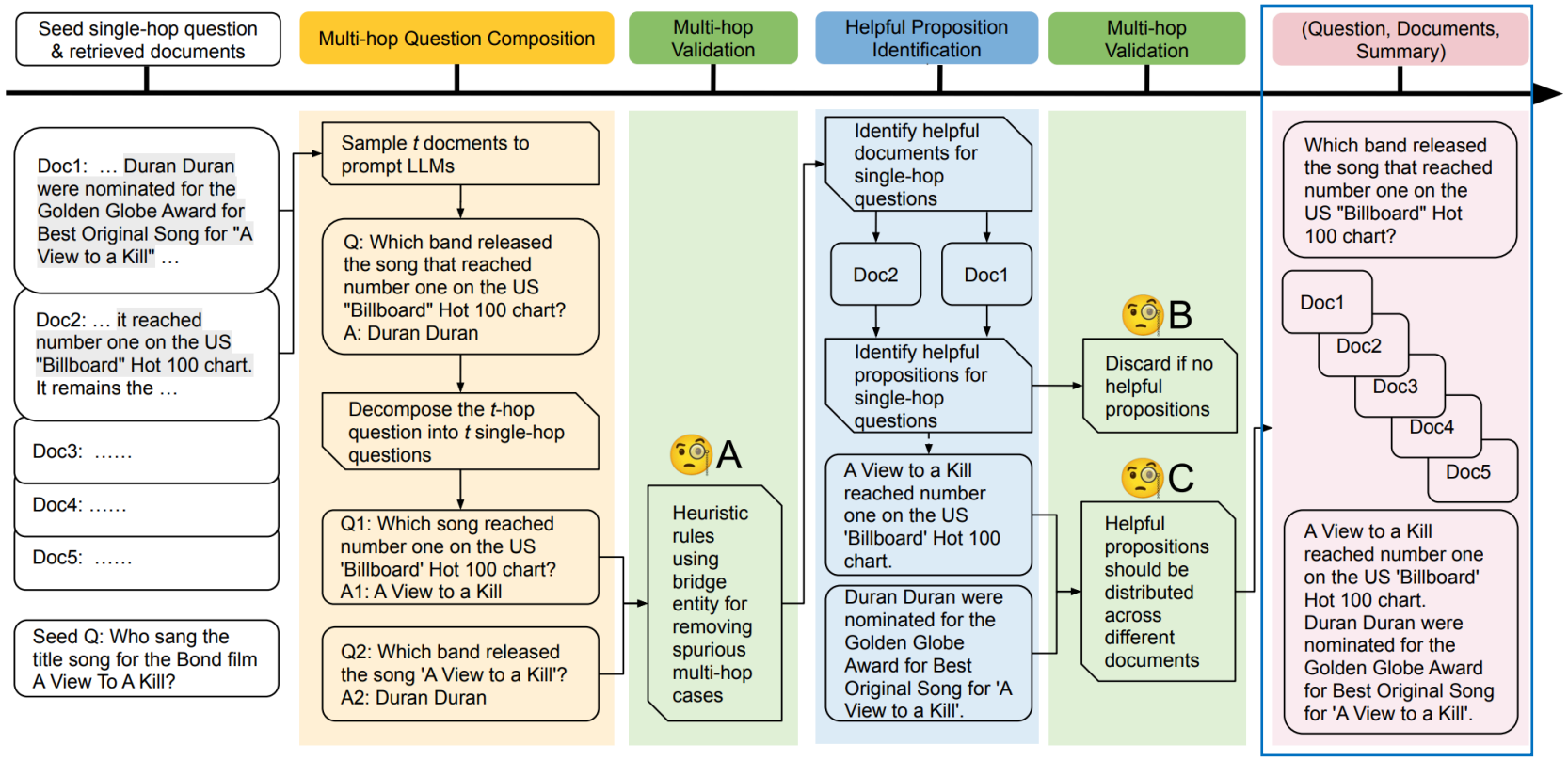
Multi-hop Validation (cont'd)

- Remove spurious multi-hop data based on identified propositions
- Ensure the model collects relevant evidence from multiple sources



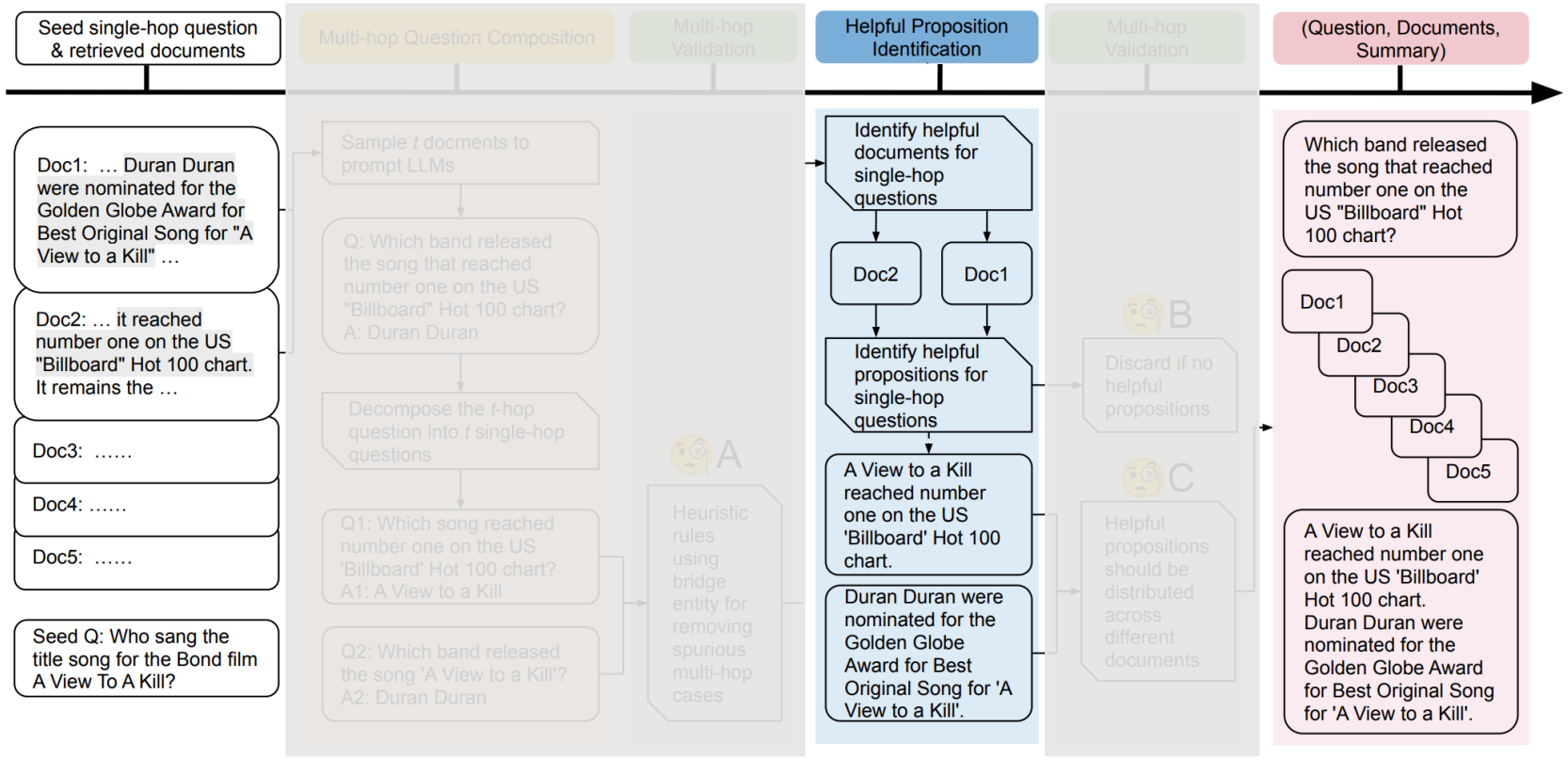
Synthetic (Question, Documents, Summary)

- The target summary is defined as the concatenation of the identified helpful propositions



Simplification for Single-hop Data

- Bypassing the modules to generate data for handling single-hop queries
- Allow for the compression for questions of varying complexity



BRIEF Training

- Utilize the curated synthetic dataset to fine-tune a T5-large (770M) compressor for a fair comparison with RECOMP
- The fine-tuning process follows the standard next token objective
- Maximize P(summary | query, documents)

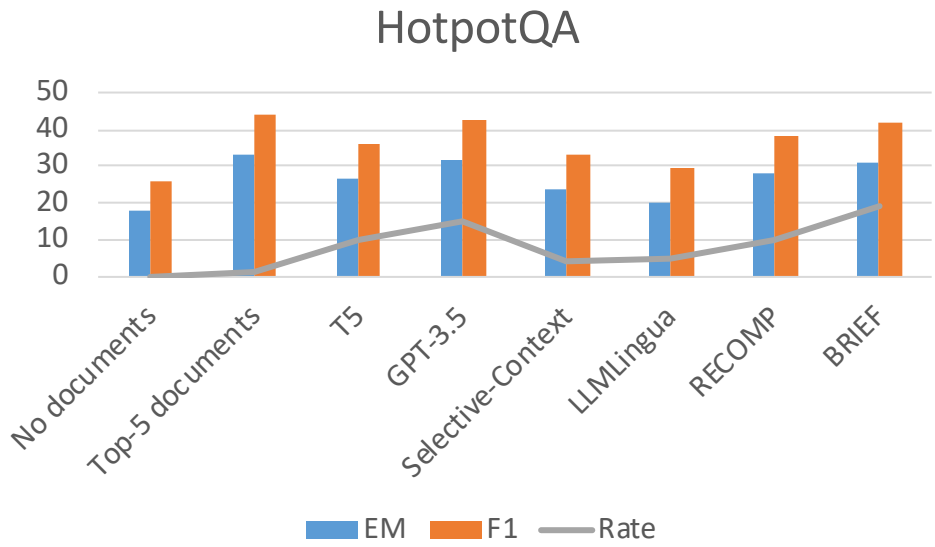
Fangyuan Xu, et al. *RECOMP: Improving Retrieval-Augmented LMs with Context Compression and Selective Augmentation.* ICLR 2024.

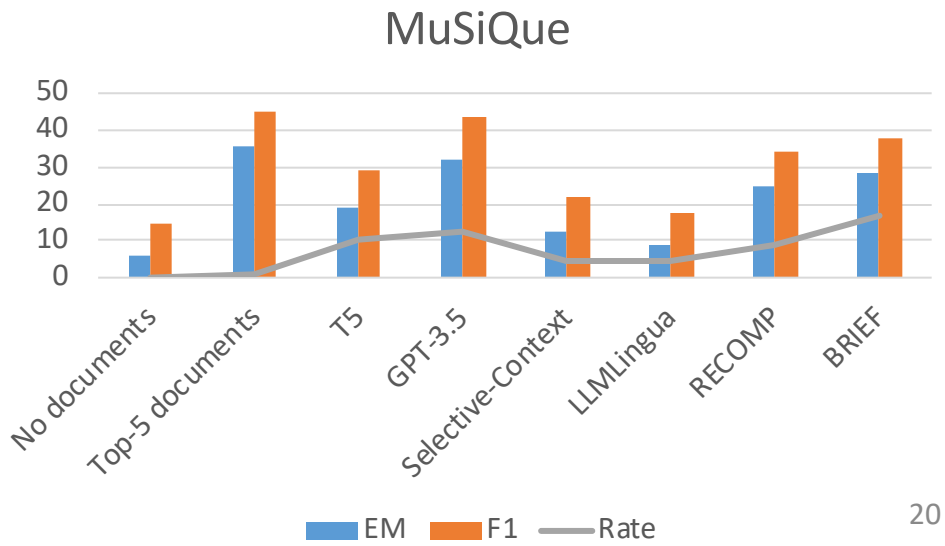
Experimental Settings

- Datasets:
  - ✓**Single-hop**: Natural Questions (NQ), and TriviaQA
  - ✓**Multi-hop**: HotpotQA, MuSiQue, MultiHop-TriviaQA and MultiHop-NQ
- Metrics:
  - **QA performance**: EM and F1 of answer strings
  - **Compression rate**: #words in retrieved documents / #words in summary, a higher compression rate indicates a shorter summary
- Baselines: (1) The off-the-shelf T5-large, (2) LLMLingua, (3) Selective Context, (4) RECOMP, and (5) GPT-3.5
- The compressed summaries produced by different methods are all fed to the same reader LM Flan-UL2

Results: Multi-hop QA

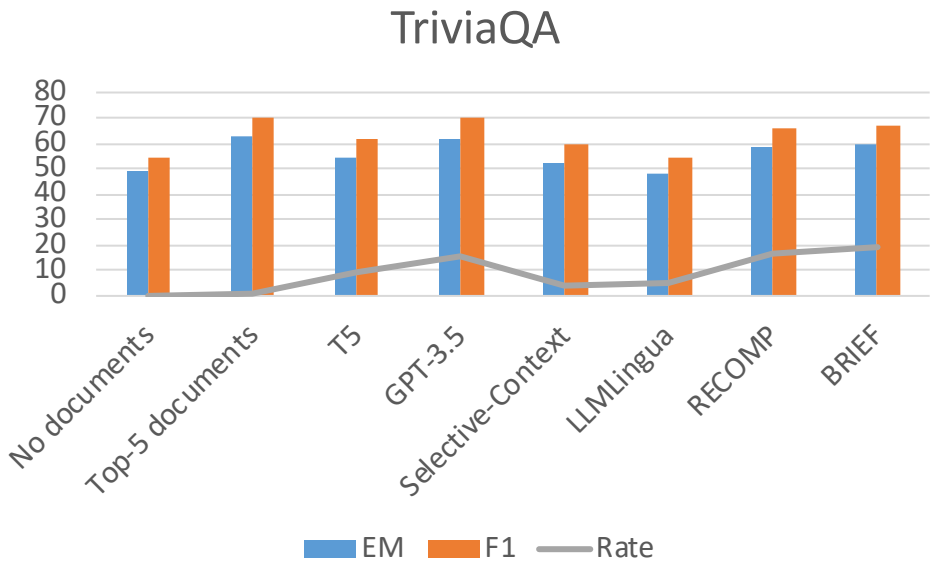
- BRIEF achieves 19.19x, with only a 1.60% EM and 1.83% F1 decrease compared to **No Compression** on HotpotQA
- Higher than **RECOMP**'s 10.02x, while still outperforming it by 3.00% EM and 4.16% F1
- Higher than **GPT-3.5**'s 14.77x, while delivering nearly similar QA results

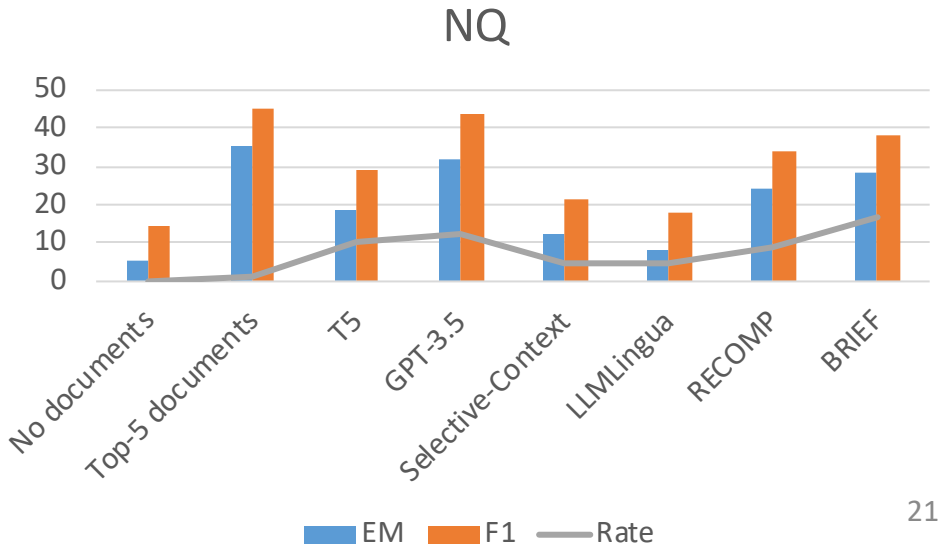




Results: Single-hop QA

- BRIEF achieves 29.76x, with only a 2.55% EM and 3.49% F1 decrease compared to **No Compression** on TriviaQA
- Higher than **RECOMP**'s 16.23x, while still outperforming it
- Higher than **GPT-3.5**'s 11.33x, while delivering competitive QA performance on NQ

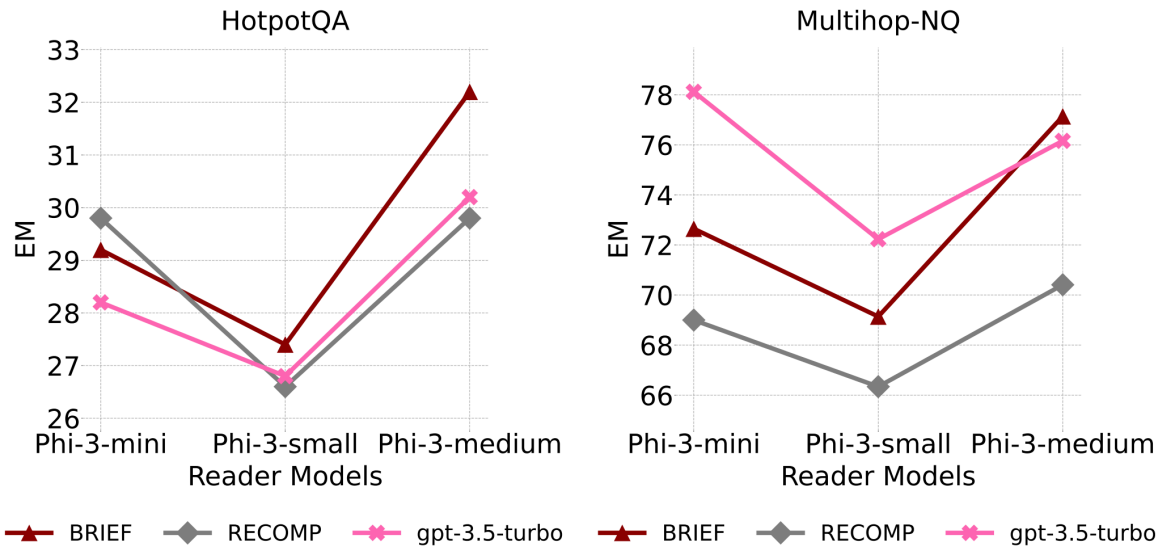




Analysis (1): The transfer ability of compressed summaries across LMs

- Maintain the core semantics relevant to the query, while compatible with a wider range of LMs

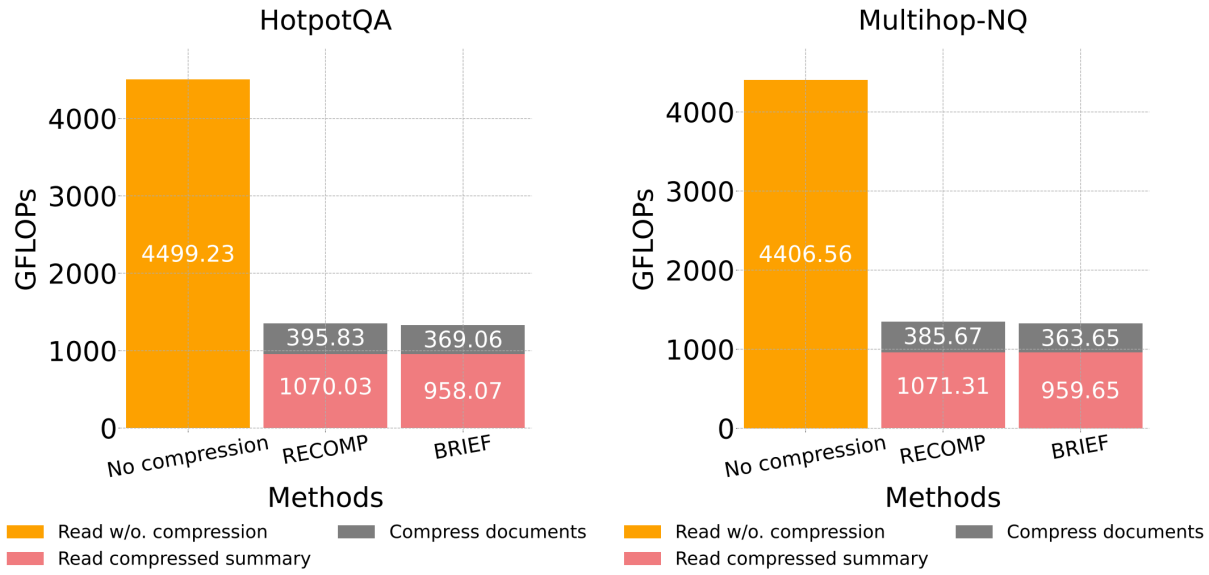
- BRIEF is more compatible
  - ✓drop less from *mini* to *small*
  - ✓enlarge more from *small* to *medium*



Analysis (2): The improvement of latency in terms of the overall computational overhead

- Computation is reduced to less than 30% of that before compression

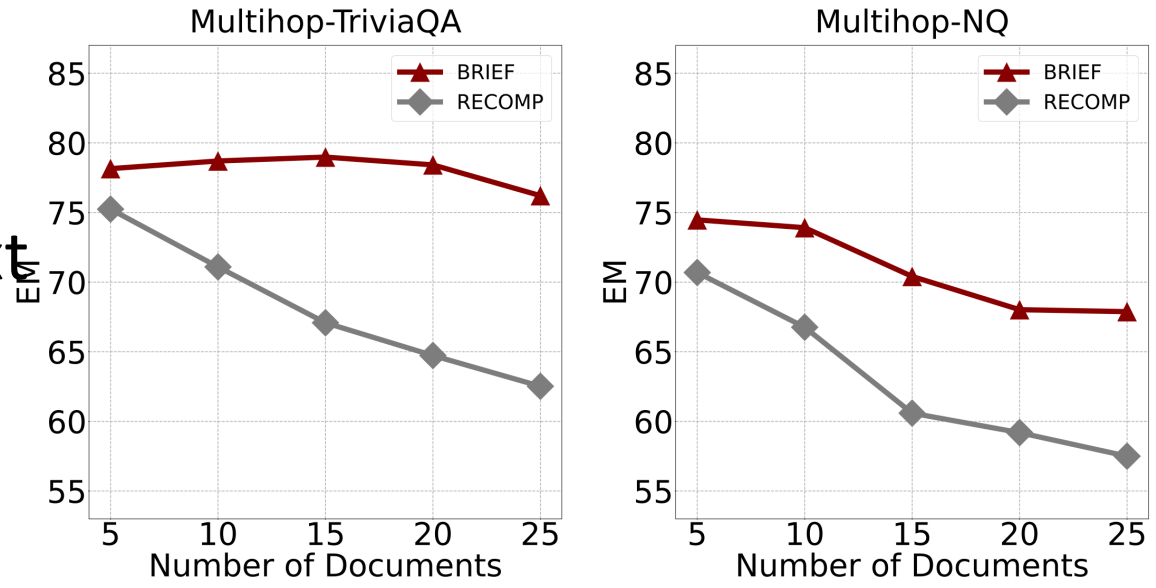
- Highlight BRIEF's potential to optimize inference, especially for large-scale document retrieval and understanding



Analysis (3): The scalability to compress longer documents

- Expand the scope of documents from the top-5 to the top-25

- BRIEF demonstrates better scalability, identifying evidence within a longer and noisy context



Conclusion

- **BRIEF, a context compressor** tailored for document compression to enable multi-hop reasoning with RAG
- **Synthetic data**, built entirely by open-source models, is high-quality and cost-effective synthetic data for learning context compression
- **More concise** summaries while enabling LMs to show **better QA** than compression methods and **competitive QA** than proprietary GPT-3.5

Thanks! Q&A

Homepage: https://JasonForJoy.github.io

Contact: gujc@ucla.edu

Github: https://github.com/JasonForJoy