Context-Efficient Retrieval with Factual Decomposition

Anonymous ACL submission

Abstract

There has recently been considerable interest in incorporating information retrieval into large language models (LLMs). Retrieval from a dynamically expanding external corpus of text allows a model to incorporate current events and can be viewed as a form of episodic memory. Here we demonstrate that pre-processing the external corpus into semi-structured "atomic facts" makes retrieval more efficient. More specifically, we demonstrate that our particular form of atomic facts improves performance on various question answering tasks when the amount of retrieved text is limited. Limiting the amount of retrieval reduces the size of the context and improves inference efficiency.

1 Introduction

Although large language models (LLMs) demonstrate remarkable capabilities across various tasks, their inability to continuously adapt to dynamic or domain-specific knowledge without parameter updates remains a substantial limitation. To address this limitation retrieval-augmented generation (RAG) supplements models with some external knowledge source during inference (Lewis et al., 2020; Ram et al., 2023; Borgeaud et al., 2022a). Typically these models treat the external source as a set of arbitrarily segmented blocks of raw text. However, there has also been interest in using more structured external knowledge sources such as knowledge graphs (Edge et al., 2024; Peng et al., 2024), compressed documents (Xu et al., 2024) or document trees (Sarthi et al., 2024). In each case one can identify a "unit of retrieval" where one retrieves some set of such units, such as a set of documents or a set of knowledge graph triples. Various candidates for units of retrieval, or "atomic facts", have been formulated (Chen et al., 2023; Jiang et al., 2024b; Min et al., 2023; Gunjal and Durrett, 2024).

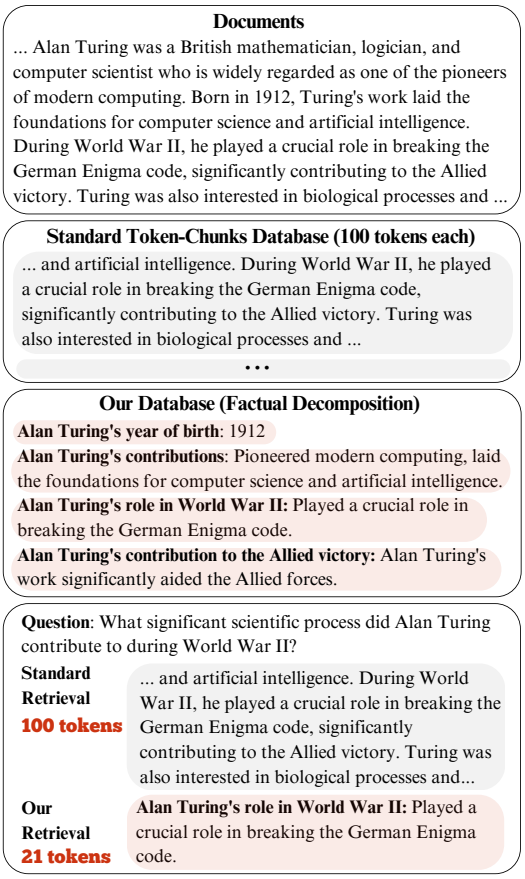

Figure 1: Example of datastore used for knowledge retrieval in our approach compared with typical fixed-size text chunks in RAG. We retrieve much shorter contexts.

In designing units of retrieval there is a tension between concise but brittle logical representations, such as knowledge graph triples, and highly expressive and nuanced, but verbose and unstructured, chunks of raw text. We propose an intermediate retrieval unit that we call an entity-description pair (EDP). This is a pair of an "entity"[1] and some form of description of that entity. For example the entity might be "Alan Turing's contributions" and the

[1] Here we take a very liberal notion of "entity" not to be confused with the narrow notion of entity used in named entity recognition.

factual description could be "Pioneered modern computing, laid the foundation of computer science and engineering". Each EDP is a structured piece of information, like in structured databases, but also enjoys the flexibility of natural language. See figure 1. We use a three-step language model prompting protocol to decompose a chunk of free text into a collection of EDPs and use the EDPs as the unit of retrieval in the resulting EDP knoweldge base (KB). See figure 2.

Our main result is a demonstration that on various challenging question answering benchmarks EDP KB retrieval achieves better accuracy when the amount of retrieval (the number of retrieved tokens) is limited. This can be phrased as improving the "context-efficiency" of RAG. We are also optimistic that our formulation of EDP KBs is a significant step toward more structured yet expressive internal representations of knowledge.

2 Related Work

Context-Efficient Retrieval As we will see in experiments, our approach achieves superior performance in context-efficient retrieval, which we define as RAG methods striking to reduce retrieved contexts for cost-effective LLM generations. Previous related work involves various compression methods. Some focus on vector-based compression, where models learn to compress long contexts into compact memory slots through end-to-end training (Ge et al., 2024; Cheng et al., 2024). Others are text-based compression, which includes training rerankers (Pradeep et al., 2023), applying extractive summarization (Xu et al., 2024), or training abstractive summarizers to compress the retrieved context (Xu et al., 2024; Jiang et al., 2024a). We also reduce retrieved contexts, but rather than compressing them *post-retrieval*, we achieve context efficiency from the outset through improved knowledge representation *pre-retrieval*. Post-retrieval context compression methods still rely on token chunks as coarse units of knowledge for retrieval, whereas we structure the knowledge more efficiently with clear, well-defined representations that maintain high expressivity.

Knowledge Representation for Retrieval Most previous works directly segment source documents into equal-length text chunks, each containing hundreds of tokens (Lewis et al., 2020; Ram et al., 2023; Borgeaud et al., 2022a). Recent research has explored alternative formats for knowledge representation, such as indexing source documents using knowledge graphs (Edge et al., 2024), hierarchical tree structures (Sarthi et al., 2024), or more relaxed versions of knowledge graphs (Liang et al., 2024).

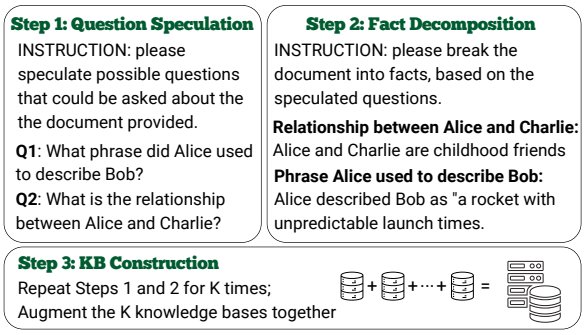

Figure 2: Overview of our method.

Unlike these works, we don't rely on any explicit relational structures; our knowledge datastore consists of flat, semi-structured entity-description pairs The work most similar to ours is Chen et al. (2023), which decomposes each sentence in Wikipedia into individual propositions for retrieval. However, we propose a novel method where speculated queries generated by the LM guide the fact extraction, and repeated samples of factual decompositions enhance our database. These techniques yield significant performance improvements by enabling more targeted information extraction and increased coverage of constructed facts, resulting in more relevant and concise data being incorporated during inference.[2]

3 Methodology

As shown in Figure 2, our method consists of three main steps: (1) question speculation, (2) atomic fact extraction , and (3) knowledge base (KB) augmentation. This approach allows to decompose long documents into concise, factual entity-description pairs (EDPs), building up a semi-structured KB for RAG, reducing retrieval overhead on the lengths of contexts to improve inference efficiency.

3.1 Question Speculation

Let D represent a long document, which is split into N equal-length chunks $D = \{D_1, D_2, \ldots, D_N\}$, where each chunk D_i contains approximately the same number of tokens. For each chunk D_i, we prompt a LM to speculate a set of possible questions $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{iJ}\}$,

[2]More general background of RAG is in Appendix A.

where q_{ij} represents a potential question one might ask about the chunk D_i. This question speculation process helps direct the extraction of relevant knowledge and facilitates targeted information retrieval later.

Formally, given a document chunk D_i, we define the question speculation function as: $Q_i = \text{LM}_{\text{speculate}}(D_i)$, where $\text{LM}_{\text{speculate}}$ denotes the question-speculation language model. The result is a set of speculative questions Q_i for each document chunk D_i. Prompts for $\text{LM}_{\text{speculate}}$ are shown in Table 4 and Table 5 in Appendix F.

3.2 Query-Guided Factual Decomposition

Once we have the speculative questions Q_i, we feed both the set of questions Q_i and the corresponding document chunk D_i into the language model to extract relevant information that can be used to answer the questions. The goal is to retrieve concise, atomic facts that are highly specific and contextually relevant.

We prompt the language model to produce a set of EDPs for each chunk, where EDP is defined as a pair k_{im} consisting of an entity e_{im} and a fact f_{im}. The entity e_{im} represents a key concept, while the fact f_{im} encapsulates the essential information regarding e_{im}. Notably, the entity needs not be limited to a noun or an entry from a traditional knowledge graph; it can be a short noun phrase, sentence, or even a question.

The extraction process for a chunk D_i is as: $K_i = \text{LM}_{\text{extract}}(D_i, Q_i)$, where $K_i = \{k_{im} = (e_{im}, f_{im})\}_{m=1}^{M}$ is the set of EDPs for chunk D_i. This method ensures that the extracted knowledge is both flexible and informative. Note that each EDP k_{im} does not have to correspond to a particular query q_{ij} as K_i are generated collectively with guidance from all Q_i, and the total number of of EDPs M could vary, regardless of the size of Q_i. Prompts for $\text{LM}_{\text{extract}}$ are shown in Table 6 and Table 7 in Appendix F.

3.3 Sample Augmentation

To further enrich the knowledge base, we apply a sampling-based approach that augments the fact extraction across multiple runs. By repeating the extraction process multiple times using the same prompt and leveraging the inherent randomness of the LM's outputs, we capture diverse sets of EDPs and prevent information gaps. This stochasticity, similar to techniques like self-consistency with majority voting (Wang et al., 2023), allows the model to produce variant responses on each run. We aggregate the knowledge extracted from these different runs to build a more comprehensive and robust KB.

Let S denote the number of sampling runs. For each document chunk D_i, we repeat the extraction process, including both question speculation and EDP extraction, S times, yielding multiple KBs: $K_i^{(1)}, K_i^{(2)}, \ldots, K_i^{(S)}$. We then merge these KBs to form a final, augmented knowledge base K_i^{final} for chunk D_i: $K_i^{\text{final}} = \bigcup_{s=1}^{S} K_i^{(s)}$. The final knowledge base for the entire document D is then constructed by merging the augmented knowledge from all chunks: $K^{\text{final}} = \bigcup_{i=1}^{N} K_i^{\text{final}}$. K^{final} provides a rich semi-structured knowledge repository for retrieval, where units are each EDP.

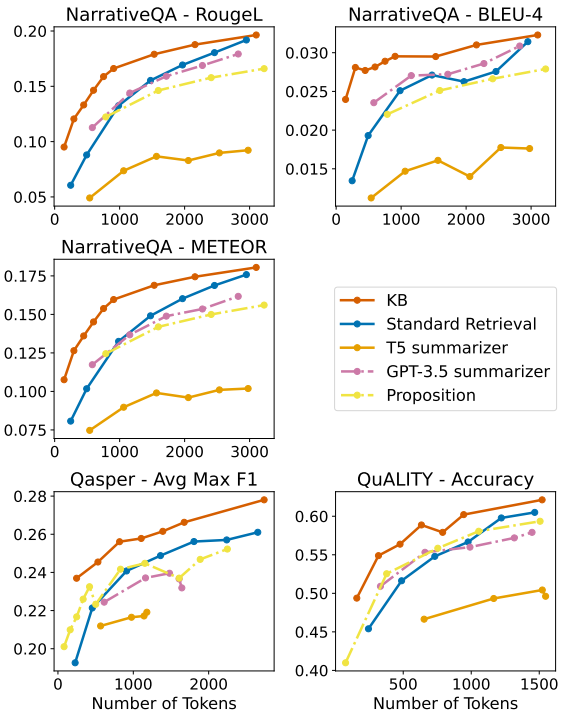

Figure 3: Results on NarrativeQA (top three plots), Qasper (bottom right), and quality (bottom left). The x-axis represents the number of tokens fixed in the retrieval context, and y-axis are different QA metrics used for each dataset.

4 Experiments

4.1 Setup

We evaluate our method on three long-context QA datasets: **NarrativeQA** (questions based on books and movie transcripts (Kočiský et al., 2018)), **Qasper** (questions from NLP research papers (Dasigi et al., 2021)), and **QuALITY** (multiple-

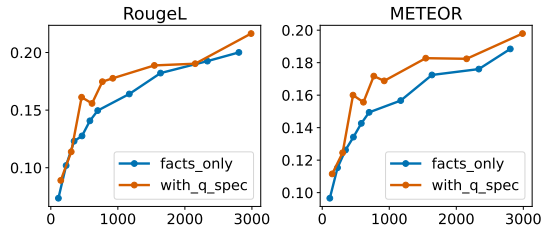

Figure 4: Comparison of performance (y-axis) vs. number of retrieved tokens (x-axis) between Fact-Only KB construction and Question-speculated KB construction on a subset of NarrativeQA's validation set.

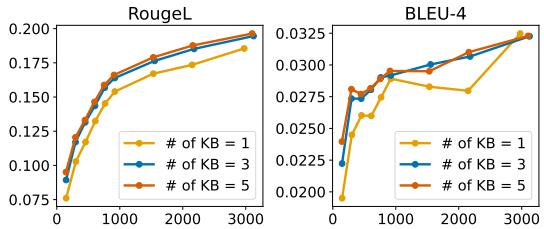

Figure 5: Performance (y-axis) vs. number of retrieved tokens (x-axis) on NarrativeQA for the number of resampled KBs equal to 1, 3, and 5.

choice questions with English articles, such as sci-fi, magazine articles, and nonfiction (Pang et al., 2022)). See Appendix B.1 for more details on data and evaluation metrics.

Following Xu et al. (2024), we compare our approach to the following baselines: **decomposition into propositions** (Chen et al., 2023) using ChatGPT (gpt-4o-2024-08-06), **retrieve-then-summarize** (Xu et al., 2024) using off-the-shelf summarizers (T5-large and GPT-3.5), and **standard retrieval** on raw chunks. Please refer to Appendix B.2 for more details.

4.2 Results

Figure 3 shows results on NarrativeQA, Qasper, and QuALITY. We see that our method consistently outperforms all baselines when the number of tokens in the context is kept at *all* different levels, shown with our curve above all others. Our method performs especially well on short-context regime when the retrieved tokens are very limited. This can effectively reduce LLM inference cost with more efficient usage of context in RAG. We also find that decomposing sentences into propositions (Chen et al., 2023) does not generalize well to domains with lower fact density, such as novels or scientific papers. Some quantitative examples of retrieved documents for each method are provided in Appendix C.

4.3 Analysis

Here we ablate each component of our method (Section 3) on their contribution to the overall performance.

Why Question Speculation? Despite extensive prompt tuning, providing speculated queries to LM when generating KB results consistently yields better performance compared to letting the LM extract facts without guidance (see Figure 4). To investigate this, we randomly select 20 stories (617 associated queries) from NarrativeQA and compute the similarities between the speculated questions and the real queries. Surprisingly, using similarity thresholding heuristics and manual inspection, we find that 11.18% of the speculated questions closely align with or rephrase the real queries, and 53.97% focus on the same topic (for more details, see Appendix D). This significant overlap aligns with previous research showing that LLMs are effective at generating synthetic queries (Wu and Cao, 2024). In fact, these speculated questions function like a chain-of-thought process (Wei et al., 2024), allowing the LM to gather relevant information before answering the query.

Why KB Augmentation? We observe that the questions speculated and facts extracted vary between different runs due to the LM's inherent stochasticity. Figure 5 shows that augmenting KBs improves performance, indicating that the sampling process effectively captures a more diverse range of meaningful knowledge pairs. Full results for NarrativeQA, Qasper, and QuALITY are in Appendix E.

5 Conclusion

We have demonstrated that on various challenging question answering benchmarks EDP KB retrieval achieves better accuracy when the amount of retrieval (the number of retrieved tokens) is limited. This improves the "context-efficiency" of RAG, which has real cost-effective implications for LLM inference. We are also optimistic that our formulation of EDP KBs is a significant step toward more structured yet expressive internal representations of knowledge, and we encourage future research to build upon and expand this approach.

Limitations

While our approach demonstrates improved context-efficiency in retrieval-augmented generation for question answering tasks, several limitations warrant discussion. First, our method relies heavily on the performance of large language models for both question speculation and factual decomposition. Any biases or errors inherent in these models could propagate through the process, potentially affecting the quality and reliability of the extracted entity-description pairs.

Second, the stochastic nature of our sampling-based augmentation introduces variability in the generated knowledge bases. Although multiple samples help capture a broader range of information, this approach may lead to inconsistencies across different runs. Further research is needed to assess the stability and reproducibility of the results when applying our method in diverse settings.

In summary, while our method enhances context-efficiency, it remains vulnerable to inherent LLM biases and sampling-induced variability. Addressing these issues is crucial for improving the reliability and consistency of our approach in various applications.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *Advances in Neural Information Processing Systems*, 36.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022a. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022b. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2023. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. *Preprint*, arXiv:2405.13792.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Tao Ge, Hu Jing, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in llm fact verification. *Preprint*, arXiv:2406.20079.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024a. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024b. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xun Liang, Simin Niu, Zhiyu li, Sensen Zhang, Shichao Song, Hanyu Wang, Jiawei Yang, Feiyu Xiong, Bo Tang, and Chenyang Xi. 2024. Empowering large language models to set up a knowledge retrieval indexer via self-learning. *Preprint*, arXiv:2405.16933.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankVicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv:2309.15088*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Mingrui Wu and Sheng Cao. 2024. Llm-augmented retrieval: Enhancing retrieval models through language models and doc-level embedding. *arXiv preprint arXiv:2404.05825*.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. In *International Conference on Learning Representations*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Appendix

A Background on RAG

Retrieval-Augmented Generation Retrieval-augmented generation (Lewis et al., 2020) (RAG) is the process of dynamically adding additional information at inference time through a similarity search process in order to improve generation quality. It is typically used in domains where it may be difficult for the language model to rely on parametric knowledge alone, for example long-tail question answering, or for current events past the training data cutoff date. The simplest form of RAG is to add text related to the query directly to the input (Ram et al., 2023). There are also vector-based variants, for example injecting information at deeper layers of network (Borgeaud et al., 2022b; Wu et al., 2022; Bertsch et al., 2023), or interpolating with a nearest neighbor generation (Khandelwal et al., 2019). Some works tune with retrieval-augmentation (Guu et al., 2020), or to induce retrieval behavior (Asai et al., 2024). Though retrieval-augmentation generally quite beneficial, language models can be distracted depending on the order (Liu et al., 2024) or content (Yoran et al., 2023) of the data retrieved.

B Experiment Setup

B.1 Datasets

- **NarrativeQA** (Kočiskỳ et al., 2018) is a dataset containing 1,572 documents, including books and movie transcripts. It requires answering questions based on the full text of these narratives. The task tests the model's ability to comprehend entire stories, with performance measured using BLEU (B-1, B-4), ROUGE (R-L), and METEOR metrics. We report BLEU-4, ROUGE-L and METEOR on the entire test set.

- **QASPER** (Dasigi et al., 2021) consists of 5,049 questions drawn from 1,585 NLP papers, with answers categorized as Answerable/Unanswerable, Yes/No, Abstractive, and Extractive. The questions focus on extracting detailed information embedded within the full text of the papers. Accuracy is evaluated using the F1 metric, reported on the entire test set.

- **QuALITY** (Pang et al., 2022) contains multiple-choice questions, each paired with context passages averaging around 5,000 tokens. Since the QuALITY test set is not public, accuracy is reported on the validation set.

B.2 Details on Setup

For the standard retrieval baseline, we experiment with different token counts within a chunk (see Appendix B.3) and select the best-performing one as the final baseline. In all experiments, we follow Sarthi et al. (2024), using CL100K_BASE from Tiktoken as the tokenizer to split source documents into chunks and compute final token usage. We use BM25 as the retriever, ChatGPT (gpt-4o-2024-08-06) for our EDP-based KB construction, and Mixtral-8x7B-Instruct-v0.1 for question answering.

B.3 Best Chunk Length for Standard Retrieval Baseline

We perform comprehensive ablation studies to find the optimal chunk length for each retrieved document (see Figure 6, Figure 8, Figure 7). We test chunk lengths of 50, 100, 150, 200, 250, 300, and 350 tokens, ensuring sentence boundaries are respected when chunking the book into fixed-size documents. For each chunk length, we select 5-10 different numbers of documents. We find that a chunk length of 250 tokens achieves the best performance on NarrativeQA, Qasper, and QuALITY, and we use this as the naive retrieval baseline reported in the main text.

C Quantitative Examples of Retrieved Documents

We provide datastore examples that are retrieved when answering a question from NarrativeQA. Table 1 shows our retrieval compared to the standard retrieval, and Table 2 shows the retrieval following Chen et al. (2023)'s proposition method. We find that our retrieval leads to the best final answer, while the other two approaches struggle to retrieve the correct information from their datastores. The standard baseline fails to find the relevant chunk from the book, and the proposition baseline decomposes all human dialogue into even smaller units, which makes the information more scattered and harms retrieval.

D Ablations on Question Speculation

Table 3 shows the similarity between real queries and speculative queries in a subset of NarrativeQA.

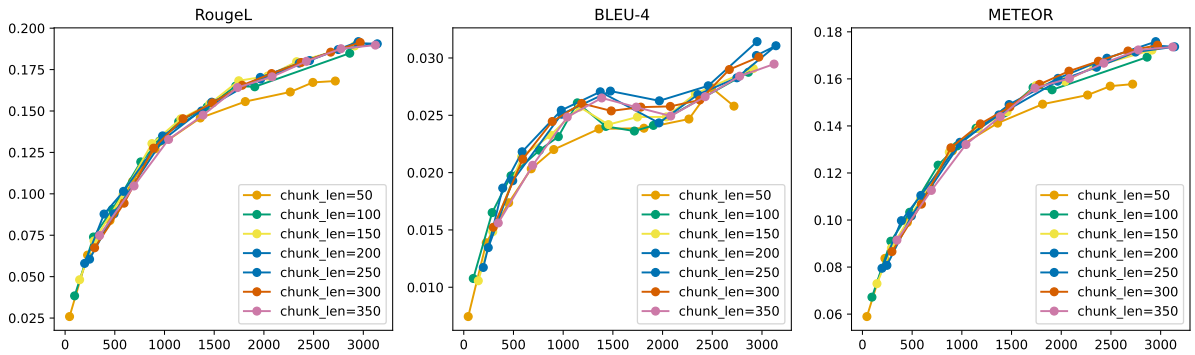

Figure 6: Metrics vs Number of Tokens for Different Chunk Lengths on NarrativeQA

Question: What does Mark urge his listeners to do?
Ground Truth: to do something about their problems instead of committing suicide

Our retrieval:
- "how does happy harry hardon describe his listeners?: happy harry hardon refers to his listeners as 'horny'.",
- "what actions does happy harry hardon encourage his listeners to take when contacting him?: happy harry hardon encourages his listeners to send him their most pathetic moments, as long as they are real, and wants details like size, shape, feel, and smell.",
- "who is happy harry hardon and what role does he play?: happy harry hardon is a radio host, engaging with his listeners over the air, and helping chris discuss his experiences and feelings.",
- "what contest does happy harry hardon mention during his show and how does he engage his listeners?: happy harry hardon runs a contest on the best way to put parents out of their misery and challenges listeners to amaze and discuss the sensational content of his show.",
- "what concern does mark have about the effects of his message?: mark feels ill about the situation and perceives it as chaotic, tangled with his life",
- "what does mark express about his feelings in the conversation with his parents?: he finds it difficult to talk to girls in his new place and feels different.",
- "what reaction does mark have to the police car pulling up?: mark is not worried and thinks the police are just dropping in on his neighbor.",
- "what internal conflict does mark express after his interaction with nora?: mark expresses that the whole thing is making him ill and that it's his life that nora is messing with.",
- "what critical message does mark want to convey to his audience about taking control of their lives?: mark conveys that it's their life, they should take charge of it, and they should fill the air and speak out.",
- "english class question: jan emerson asks mark to share his feelings about what he wrote."

Our answer: take charge of their lives.

Standard retrieval baseline:
"happy harry hardon - you interview a student and then you rat on her, you betray her trust, isn't that right sir <deaver hangs up> well as you can see, these guys are played out society is mutating so rapidly that anyone over the age of twenty has really no idea err alright, back down to business ï share a room with my older brother and nearly every night after he turns off his light he come over to my bed and gives me a few arm nookies and stuff and then makes me scratch his back and other refinementsït's about time we had some refinements on this show ïhen sooner or later he gets worked up and further a do he rubs his thing and makes me watch ŝigned ï'm just screwed upẅell first of all you're not screwed up, your an unscrewed up reaction to a screwed up situation feeling screwed up at a screwed up time, in a screwed up place does not make you necessarily screwed up, if you catch my drift well as you know dear listeners if you enclose your number a reply is guaranteed <rings miss screwed up>",
"creswood - it's the trouble makers, you can't run a top school with trouble makers in the mix brian - okay, so what exactly is a trouble maker creswood - someone who has no interest in education brian - oh c'mon that includes every teenager i know creswood - can't you understand that nothing is more important than a good education brian - except for the basic right to it creswood - the point is i have the highest s a t scores in the state brian - yeah but how creswood - i stand by my record <the school field> shep sheppard - mr watts, shep sheppard channel six news here watts - good evening shep sheppard - how does washington intend to deal with this situation watts - we at the f c c feel that democracy is about protecting the rights of the ordinary citizen un-regulated radio would result in programming of the lowest common denominator, the rule of the mob <watts looks round to see one of his f c c vans has picked up some graffiti> this is vandalism, not free expression <everyone gathers outside the school to listen to the happy harry hardon show>"
Standard retrieval's answer: mark does not urge his listeners to do anything

Table 1: Example of answering a question from NarrativeQA: In our method, the blue highlight represents the evidence most likely to contribute to the final answer.

Question: What does Mark urge his listeners to do?
Ground Truth: to do something about their problems instead of committing suicide

Proposition Baseline (Chen et al., 2023):
"creswood states, s̈o what does this prove, not everyone goes to college.",
"mark goes to collect his post.",
"the speaker asserts, ḧappy harry hardon will go to any language to keep his three listeners glued with huwy bluwy to their radios.",
"happy harry hardon asks, äre you willing to tell my listeners what you told me here in this letter?",
"happy harry hardon invites listeners to share their most real moments.",
"nora finds mark burning his happy harry hardon letters.",
"mark adds, ï know exactly what it means.",
"mark protests, ï swear, what are you doing?",
"mrs. kaiser invites malcolm to join his parents downstairs.",
"happy harry hardon states that listeners are interested in the decision to expel cheryl bates.",
"mark explains that it is his mom's jeep and that she kind of loaned it to him.",
"chris expresses, ï didn't know what to do.",
"mark says, n̈o it's outside,änd shows nora his converted radio jeep.",
"happy harry hardon asks david deaver to explain his work.",
"mark collects his post from the postal center, exits, and starts to read the eat me beat me lady's letter.",
"donald shakes his head in disgust.",
"chris asks, s̈o what are we going to do about this?¨",
"happy harry hardon suspects a lie if miss screwed up does not remember or tell the truth.",
"happy harry hardon concludes, b̈ut you know what you have to do.¨",
"happy harry hardon lists, ÿou have parents, teachers telling you what to do.¨",
"nora pulls mark into the clayroom and reassures, ït's cool, it's safe. guess what i heard?¨",
"back outside the lockers, doug asks donald, s̈o what did they do to you?¨",
"nora questions, m̈ark what is with you?¨",
"malcolm's mother, mrs. kaiser, asks malcolm about his homework.",
"happy harry hardon continues, ÿou have movies, magazines, and tv telling you what to do.¨",
"happy harry hardon questions what david deaver says to young people about the world's trustworthiness.",
"detective denny, holding up his badge, implies that the postal clerk can give the information to him.",
"mark asks, c̈lose to what?¨", "malcolm tells mrs. kaiser that he has finished his homework.",
"happy harry hardon notes, n̈ow they've all run home to tune in and listen to what they've all been talking about.¨",
"mark comments, ÿeah, back to you.¨",
"happy harry hardon addresses his audience as äll my horny listeners.¨",
"marla hunter asks brian hunter, ḧave you noticed his behaviour lately?¨",
"brian questions, ökay, so what exactly is a troublemaker?¨",
"nora points out, f̈.c.c. you know what that means.¨",
"happy harry hardon asks, s̈o what did you do?¨",
"happy harry hardon prompts, s̈o tell us what happened.¨",
"mark adds, ï can't talk to them!¨",
"mark mentions having something to show nora.",
"mark comments to nora, ÿou're so different.¨",
"mark clarifies, ï can't talk to you.¨", "nora greets, ḧi! what are you doing? you having fun?¨",
"brian asks, l̈oretta what the hell is going on here?¨",
"cheryl asks, c̈an you tell me what this is about?¨",
"creswood asserts, n̈onsense, she doesn't know what she's talking about.¨",
"happy harry hardon claims, ḧappy harry just happens to have in his very hands a copy of a memo written by mr.",
"mark asserts, ï can't talk to you people.¨",
"mark declares, s̈teal it, it belongs to you.¨", "happy harry hardon acknowledges äll of my horny listeners would love it if i would call up the eat me beat me lady.¨",
"jan reveals, l̈ast night one of our students, malcolm kaiser, took his own life.¨"
Proposition Baseline's answer: Mark does not urge his listeners to do anything. No specific action is mentioned.

Table 2: Example for answering one question from NarrativeQA.

The similarity is measured by computing the similarity between embeddings encoded with the all-MiniLM-L6-v2 model from Hugging Face's sentence-transformers. We examined 617 questions and found that 11.18% of the speculated questions closely align with or rephrase the real queries, while 53.97% focus on the same topic.

E Ablations on KB Augmentation

Figure 9 shows the effect of different numbers of KBs in NarrativeQA. Figure 10 shows the effect of different numbers of KBs in Qasper. Figure 11 shows the effect of different numbers of KBs in QuALITY.

Real Question	Speculated Question	Similarity
Closely Related / Rephrase of the Question (Similarity $\geq$ 0.85)		
Why does Helen return to Grassdale?	Why does Helen eventually return to Grassdale alone?	0.9637
What name does Klaatu use at the boarding house?	Where does Klaatu come from before entering the boarding house?	0.9013
What object did Tom find in Klaatu's room?	What does Tom find on the floor of Klaatu's room?	0.8852
How does Data finally defeat the Borgs?	What actions does Data take to thwart the Borg's attempts?	0.8640
What gift did the Borg Queen offer Data?	What does the Borg Queen want from Data?	0.8614
Questions on the Same Topic (Similarity 0.7 - 0.85)		
What did Klaatu say would happen if his message was ignored by Earth's people?	What does Klaatu want to discuss with representatives from Earth?	0.7529
	What is Klaatu's demeanor when he discusses the stakes for Earth's future if his message is not heeded?	0.7783
	How does Klaatu react to the replies from world leaders regarding the meeting?	0.7284
	What alternative does Klaatu say Earth would face if his proposals are rejected?	0.7517
	What message does Klaatu ask to be delivered and to whom?	0.7136
	What ultimatum is being given to the audience in Klaatu's message?	0.7272
Who did Bobby suggest was the greatest living person?	How does Bobby respond to Klaatu's question about the greatest man in America?	0.7343
	Who does Bobby identify as the greatest scientist in the world?	0.7316

Table 3: Examples of Speculated Questions and Their Similarity to Real Questions

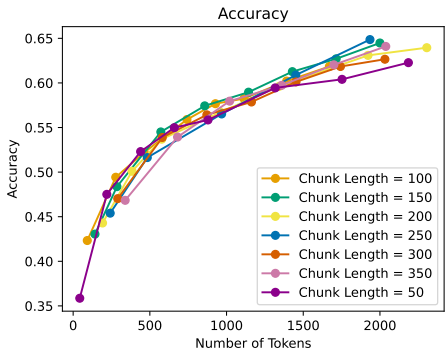

Figure 7: Accuracy vs number of retrieved tokens for different chunk lengths as retrieval units in the standard RAG approach on QuALITY dataset.

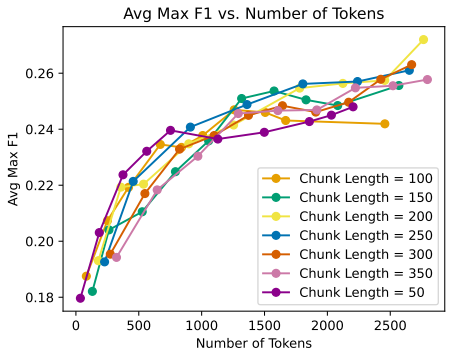

Figure 8: Avg Max F1 vs number of retrieved tokens for fifferent chunk lengths as retrieval units in the standard RAG approach on Qasper dataset.

F Prompts

We detail all the prompts used in our method and baselines. For our method, prompts for question

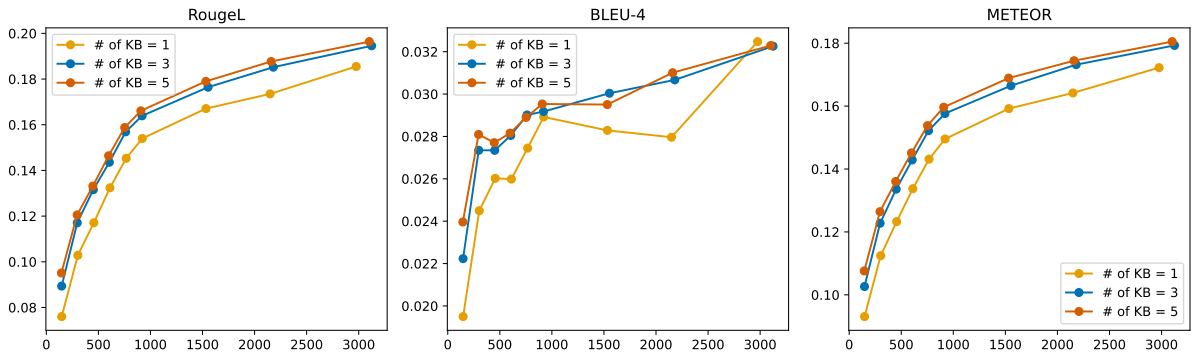

Figure 9: Results on different number of KBs on NarrativeQA.

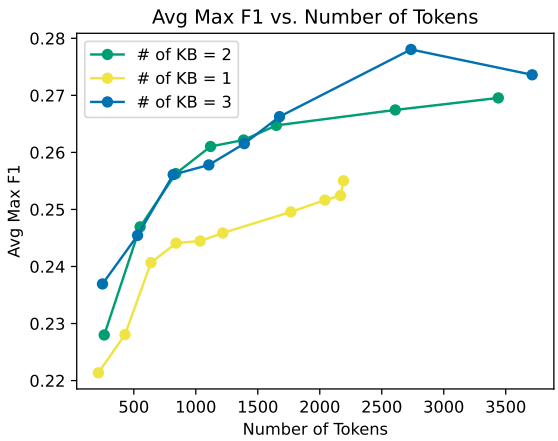

Figure 10: Results on different number of KBs on Qasper.

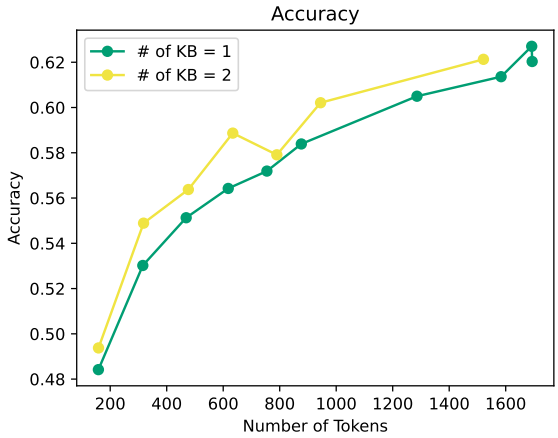

Figure 11: Results on different number of KBs on QuALITY.

speculation are shown in Table 4 and Table 5. Prompts for EDP KB construction are shown in Table 6 and Table 7. Prompts for question answering are shown in Table 8, Table 9, Table 10, and Table 11. Note that for NarrativeQA, we use a two-step prompting approach to obtain the final answer: first, perform regular question answering based on the query and retrieved documents; second, compress the answer to make it more concise. This is because answers in NarrativeQA are typically just a few words, but Mixtral tends to generate lengthy responses regardless of prompt adjustments, prompting us to adopt a two-step process.

NarrativeQA & QuALITY (Question Speculation)

System: You are a highly attentive assistant focused on generating specific and concise questions about the narrative elements of a text. Your goal is to produce clear and direct questions that help a reader deeply understand the concrete aspects of the story.

User: Task: Generate Specific, Concrete, and Contextual Narrative Questions

Objective: Given a section of text from the book, generate a set of specific, concise, and detailed questions that are directly related to the narrative elements—such as characters, actions, events, settings, and their historical or cultural significance. If the text contains irrelevant information like publisher details, web content, or other non-narrative elements, do not generate questions and instead return 'no questions extracted.'

Instructions:
1. **Read the Text Carefully**: Pay close attention to the provided section of the text to fully understand the narrative context, including any historical or cultural references.
2. **Check for Irrelevant Information**: Identify whether the text contains non-narrative elements such as publisher details, web content, disclaimers, or any information not directly related to the narrative. If such content is found, return 'no questions extracted.'
3. **Identify Key Narrative and Contextual Elements**: If the text is free from irrelevant information, focus on identifying the key events, actions, characters, settings, and any historical or cultural references. Consider what is happening, who is involved, where and when these events are taking place, and the historical or symbolic significance of these elements.
4. **Formulate Questions**: Create questions that are specific to the identified narrative and contextual elements. Ensure each question is concise, detailed, factual, and directly connected to the content of the narrative, including its historical, cultural, or symbolic context.
5. **Question Variety and Depth**: Aim for a diverse set of questions that cover various aspects of the narrative, including specific locations, character roles, relationships, and cultural or historical context. Avoid redundancy by ensuring each question explores a different element or angle of the narrative.
6. **Avoid Abstract and Meta-Content**: Refrain from generating questions about abstract themes, philosophical ideas, or meta-information such as publication details or background information unrelated to the narrative itself.

Example:

Here is an excerpt from the book:
—
T̈he Great Peace towards which people of good will throughout the centuries have inclined their hearts, of which seers and poets for countless generations have expressed their vision, and for which from age to age the sacred scriptures of mankind have constantly held the promise, is now at long last within the reach of the nations. For the first time in history it is possible for everyone to view the entire planet, with all its myriad diversified peoples, in one perspective. World peace is not only possible but inevitable. It is the next stage in the evolution of this planet—in the words of one great thinker, 'the planetization of mankind'. Whether peace is to be reached only after unimaginable horrors precipitated by humanity's stubborn clinging to old patterns of behaviour, or is to be embraced now by an act of consultative will, is the choice before all who inhabit the earth. At this critical juncture when the intractable problems confronting nations have been fused into one common concern for the whole world, failure to stem the tide of conflict and disorder would be unconscionably irresponsible."
—

Example Questions:

- Where is the Great Peace expected?
- Who has expressed the vision of the Great Peace?
- What does 'planetization of mankind' mean?
- How does the text describe the current world state?
- What critical choice is presented?

Your Turn:

Now, using the provided section of text, check for any irrelevant information. If you find any, return 'no questions extracted.' If not, generate a list of specific, concise questions covering various narrative elements such as characters, actions, settings, historical or cultural references, and symbolic meanings.
—

Section of the book

[INSERT EXCERPT HERE]"

Table 4: Prompts for generating speculative questions on NarrativeQA and QuALITY.

Qasper (Question Speculation)

System: You are an AI language model that generates insightful and analytical questions about a given passage. Your goal is to create questions that encourage deeper understanding and critical thinking about the content, themes, and details within the passage. The questions should resemble the style of the example questions provided.

User:
Instructions:
1. Carefully read the passage provided, paying special attention to any mention of the experimental design, dataset details, evaluation methods, and results.
2. Generate a list of questions focusing on the following aspects: - Experimental setup - Dataset characteristics (e.g., size, composition) - Evaluation methods and metrics - Results and conclusions
3. The questions should be clear, specific, and thought-provoking, encouraging a deep understanding of the methodology and results presented.
4. **Each question must contain only one question.**
5. **Extract as many questions as possible.**

Example:
Passage:
"Minimally Supervised Learning of Affective Events Using Discourse Relations
Recognizing affective events that trigger positive or negative sentiment has a wide range of natural language processing applications but remains a challenging problem mainly because the polarity of an event is not necessarily predictable from its constituent words. In this paper, we propose to propagate affective polarity using discourse relations. Our method is simple and only requires a very small seed lexicon and a large raw corpus. Our experiments using Japanese data show that our method learns affective events effectively without manually labeled data. It also improves supervised learning results when labeled data are small.
Introduction
Affective events are events that typically affect people in positive or negative ways. For example, getting money and playing sports are usually positive to the experiencers; catching cold and losing one's wallet are negative. Understanding affective events is important to various natural language processing (NLP) applications such as dialogue systems, question-answering systems, and humor recognition. In this paper, we work on recognizing the polarity of an affective event that is represented by a score ranging from -1 (negative) to 1 (positive).
Learning affective events is challenging because, as the examples above suggest, the polarity of an event is not necessarily predictable from its constituent words. Combined with the unbounded combinatorial nature of language, the non-compositionality of affective polarity entails the need for large amounts of world knowledge, which can hardly be learned from small annotated data.
In this paper, we propose a simple and effective method for learning affective events that only requires a very small seed lexicon and a large raw corpus. As illustrated in Figure 1, our key idea is that we can exploit discourse relations to efficiently propagate polarity from seed predicates that directly report one's emotions (e.g., "to be glad" is positive). Suppose that events x_1 are x_2 are in the discourse relation of Cause (i.e., x_1 causes x_2). If the seed lexicon suggests x_2 is positive, x_1 is also likely to be positive because it triggers the positive emotion. The fact that x_2 is known to be negative indicates the negative polarity of x_1. Similarly, if x_1 and x_2 are in the discourse relation of Concession (i.e., x_2 in spite of x_1), the reverse of x_2's polarity can be propagated to x_1. Even if x_2's polarity is not known in advance, we can exploit the tendency of x_1 and x_2 to be of the same polarity (for Cause) or of the reverse polarity (for Concession) although the heuristic is not exempt from counterexamples. We transform this idea into objective functions and train neural network models that predict the polarity of a given event.
chatWe trained the models using a Japanese web corpus. Given the minimum amount of supervision, they performed well. In addition, the combination of annotated and unannotated data yielded a gain over a purely supervised baseline when labeled data were small."
Example Questions:
1. What is the seed lexicon?
2. How are relations used to propagate polarity?
3. How does their model learn using mostly raw data?
4. How big is the Japanese data?
5. How large is the raw corpus used for training?
6. How big is the seed lexicon used for training?
7. What are the results?
8. What are the labels available in the dataset for supervision?
9. How significant are the improvements of supervised learning results trained on smaller labeled data enhanced with the proposed approach compared to the basic approach?
—
Task:
Now, read the following passage and generate a list of questions that resemble the style of the example questions.
Passage:
[INSERT EXCERPT HERE]

Table 5: Prompt for generating speculative questions on Qasper.

NarrativeQA & QuALITY (KB Construction)
System: You are a helpful assistant. **User**: Please extract all relevant entities and facts from the provided passage that are useful for answering specific questions. Only return entity and facts for information that is explicitly mentioned in the passage. If a question does not have a corresponding fact in the passage, omit that entity and fact entirely. For example, if the question is "Who visits the philosopher at the beginning of the story?" and the passage mentions that a friend visits the philosopher, the response should be (Visitor, A friend visits the philosopher). However, if the passage does not provide specific information on a question and there is no mention of the location, do not include anything in your response for that question. Your returned output should be a series of tuples, like (Visitor, A friend visits the philosopher), (Philosopher's stance on law, Breaking the law is equivalent to betraying a contract with the state). Passage: [INSERT EXCERPT HERE] Questions: [INSERT SPECULATED QUESTIONS HERE]

Table 6: Prompt for constructing knowledge bases using speculative questions from NarrativeQA or QuALITY.

Qasper (KB Construction)
System: You are a helpful assistant. **User**: Please provide answers to the following questions based on the passage. Whenever possible, prioritize using **direct quotes** from the passage instead of summarizing. Only summarize when a direct quote does not provide a clear answer. Format each answer as a pair of: (Question, Answer) If a direct quote is used, place it within quotation marks. Example format: (What is the seed lexicon?, A vocabulary of positive and negative predicates that helps determine the polarity score of an event.) (How big is the Japanese data?, 7,000,000 pairs of events were extracted from the Japanese Web corpus, and 529,850 pairs of events were extracted from the ACP corpus.) (How does the proposed method compare to previous techniques?, "Compared to existing methods, the proposed approach 'achieves a 15% increase in classification accuracy while reducing computational complexity by approximately 30%.' This substantial improvement highlights the efficiency and effectiveness of the new algorithm in large-scale data settings.") Passage: [INSERT EXCERPT HERE] Questions: [INSERT SPECULATED QUESTIONS HERE]

Table 7: Prompt for constructing knowledge bases using speculative questions from Qasper.

NarrativeQA (Question Answering - round 1)
System: You are a helpful assistant. **User**: Please answer the question below using the provided context. Your response must be a phrase that directly answers the question or the phrase 'I don't know'—no further explanation should be added. Do not provide additional context or clarification in your response. Keep the replies concise and short. Do not repeat things. Do not over-explain yourself. Reply in under 10 words. Example 1: Context: [(the morning star, The entity known as 'the morning star' is also referred to by another name in astronomy.)] Question: What is another name for the morning star? Answer: Venus. Example 2: Context: [(The battle of Hastings, The battle of Hastings was fought in the year 1066.)] Question: When was the battle of Hastings fought? Answer: 1066. Example 3: Context: [(the foundational document, The document foundational to the laws of the United States is the Constitution.)] Question: What is the foundational document of the United States? Answer: The Constitution. Please answer the question below using the provided context. Your response must be either a phrase that directly answers the question or the phrase 'I don't know'—no further explanation should be added. Do not provide additional context or clarification in your response. Context: [INSERT RETRIEVED DOCUMENTS HERE], Question: [INSERT QUESTION HERE]

Table 8: Prompt for answering questions from Qasper.

NarrativeQA (Question Answering - round 2)

System: You are a helpful assistant.
User: For the question-answer pair provided below, shorten the answer by removing any redundant elements that merely repeat information from the question. Only shorten the answer if it includes unnecessary details or redundant phrasing, ensuring that all essential information is retained. Use these provided examples as a guide for the style and level of conciseness expected in the responses.
Examples:
1. **Question:** Who was Socrates visited by at the beginning of the story?
- **Original Answer:** I don't know. The context provided does not mention anyone visiting Socrates at the beginning of the story.
- **Shortened Answer:** I don't know.
2. **Question:** What does Socrates tell Crito not to worry about?
- **Original Answer:** Socrates tells Crito not to worry about the voices of the crowd regarding Socrates' choices, and not to concern himself with the fairness of the laws.
- **Shortened Answer:** The voices of the crowd.
3. **Question:** Who announces the events that are to come to the dismay of the others on stage?
- **Knowledge Base:** The character who announces the events that are to come; Identity, Phantastes.
- **Shortened Answer:** Phantastes.
4. **Question:** Where do the dancers purify themselves?
- **Original Answer:** In the temple of Apollo.
- **Shortened Answer:** In the temple of Apollo.
5. **Question:** Where is Echo's glade?
- **Original Answer:** Echo's glade is in the forest of Arden.
- **Shortened Answer:** Arden.
6. **Question:** What challenge does Phronimus propose to all comers?
- **Original Answer:** Phronimus proposes a wit duel to all comers.
- **Shortened Answer:** Wit duel.
7. **Question:** How long has Michael lived in New York?
- **Original Answer:** Michael has lived in New York for fifteen years.
- **Shortened Answer:** Fifteen years.
8. **Question:** Who wins the sparring match between Johnny and Tom?
- **Original Answer:** Tom wins the sparring match between Johnny and Tom.
- **Shortened Answer:** Tom.
Question: [INSERT QUESTION HERE]
- **Original Answer:** [INSERT ANSWER FROM ROUND 1]
- **Shortened Answer:**
Context: [INSERT RETRIEVED DOCUMENTS HERE], Question: [INSERT QUESTION HERE]

Table 9: Prompt for answering questions from Qasper.

Qasper (Question Answering)

System: You are a helpful assistant.
User: **Instructions:**
1. If you find direct evidence from the context, extract the relevant span as your answer. Ensure it is concise and faithful to the text.
2. If the answer requires a rephrasing or cannot be directly extracted, use your own words to provide a clear, concise response.
3. For yes/no questions, simply respond with 'Yes' or 'No' based on the context.
4. If no answer is found within the context, output 'Unanswerable.'
Context: [INSERT RETRIEVED DOCUMENTS HERE]
Question: [INSERT QUESTION HERE]

Table 10: Prompt for answering questions from Qasper.

QuALITY (Question Answering)

System: You are a helpful assistant.
User: Please answer the following multiple-choice question based on the context provided.
Context: [INSERT EXCERPT HERE]
Question: [INSERT QUESTION HERE]
Options: 1. options[0] 2. options[1] 3. options[2] 4. options[3]
Choose the option that seems most appropriate based on the context, even if you're unsure. Respond with only the number of the selected option and do not provide any additional text or explanation.

Table 11: Prompt for answering questions from QuALITY.