Efficient Validation of Long-Form Text Claims through Modular Retrieval and Entailment

Jun Hu
Cornell University
Ithaca, NY
jh2829@cornell.edu

Abstract

The rapid advancement of long-context language models has enabled the summarization of extensive documents, including book-length texts exceeding 125,000 tokens. This paper addresses a critical question in this domain: *Can we efficiently evaluate claims made about a large corpus of text, where the language model has no knowledge of (i.e., has not been trained on), usually encompassing around 125K tokens?* Current methods, which involve passing the entire document to advanced language models, incur significant computational costs. We propose an alternative method that selectively extracts the most relevant chunks from the text document and uses these excerpts as evidence to evaluate claims. By framing the problem as a text classification task with (premise, hypothesis) pairs, we aim to determine the truthfulness of the hypothesis (claim) based on the premise (selected text chunk). This approach holds significant promise for reducing computational expenses while maintaining high accuracy in claim verification.

https://github.com/jjunhu/Efficient-Validation-of-Long-Form-Text-Claims

1 Introduction

The explosion of digital content has led to the creation of vast text corpora, necessitating efficient methods for evaluating claims within these large documents. A pertinent question in this context is: *Can we efficiently evaluate claims made about a large corpus of text, where the language model has no prior knowledge, usually encompassing around 125K tokens?* The current state-of-the-art approach involves passing the entire text document to advanced language models, such as GPT-4 Turbo, to verify the claims. While this method achieves relatively high accuracy, it is computationally expensive and not scalable for extensive datasets.

We propose an alternative approach that focuses on selecting the most relevant chunks from the text document and feeding these into the language model as (premise, hypothesis) pairs. This method frames the problem as a text classification task, where the goal is to evaluate whether the hypothesis (claim) is true based on the premise (selected text chunk). This approach significantly reduces computational costs while maintaining high accuracy.

Our contributions are threefold:

- We introduce an entailment dataset specifically focused on evaluating book-length summarization claims, including both entailed and non-entailed data.
- We develop a modular retrieval and entailment pipeline that abstracts the connection between the two, allowing users to easily switch between trained retrieval and entailment models.
- We demonstrate that improved retrieval models enhance the accuracy of evaluating large-context claims, even for texts unknown to the language model.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Background and Related Work

The need for efficient evaluation of claims in long-form texts spans multiple research areas. In the field of natural language processing, various studies have focused on different aspects of this problem:

FABLES: Evaluating faithfulness and content selection in book-length summarization (1). This project aimed to create a benchmark for evaluating the factual accuracy of claims by providing a dataset of annotated fact-checking examples. The FABLES dataset helps researchers develop models that can verify claims against a large body of text.

LONGEVAL: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization (2). This study introduced a set of guidelines to improve the reliability and efficiency of human evaluations for long-form summaries. The authors proposed a finer granularity of judgment to reduce inter-annotator variance and validated their approach on two long-form summarization datasets.

MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents (3). This paper presents an innovative framework for verifying the factual accuracy of language model outputs against reference documents. The authors highlight the challenges in ensuring factual consistency in generated content and emphasize the need for efficient fact-checking methods. Through MiniCheck, they address these challenges by automating the comparison process, underscoring the importance of improving evaluation techniques for large language models.

WICE: Real-World Entailment for Claims in Wikipedia (4). This work introduced a new benchmark and dataset for verifying claims in long-context documents. The authors evaluated various retrieval and entailment models, showing that better retrieval models significantly improve claim verification accuracy.

3 Background

In recent years, the task of evaluating the factual accuracy of claims in long-form texts has gained significant attention. This section provides a technical overview of the previous work in this area, highlighting the methodologies and findings of key studies.

3.1 Human Evaluation of Summarization

Human evaluation remains the gold standard for assessing the faithfulness of automatically generated summaries. However, evaluating long-form summaries presents unique challenges due to their length and complexity. (3) LONGEVAL is a set of guidelines aimed at standardizing human evaluation practices for long-form summaries. Their approach includes finer-grained annotations to improve inter-annotator agreement and reduce workload.

3.2 Automatic Evaluation Metrics

Several automatic metrics have been developed to evaluate the factual consistency of summaries. These metrics, such as ROUGE, BLEU, and BERTScore, compare the generated summary against reference summaries to assess its accuracy. However, these metrics often fall short in capturing the nuanced errors in long-form summaries (5).

3.3 Fact Verification in Long-Context Documents

The task of fact verification involves checking the truthfulness of claims against a large knowledge source. (2) Tang introduced a benchmark and dataset specifically for long-context claim verification. They evaluated various retrieval and entailment models and found that improved retrieval techniques significantly enhance verification accuracy.

3.4 Retrieval and Entailment Models

Retrieval models aim to identify the most relevant passages from a large corpus that can support or refute a given claim. Entailment models then determine whether the retrieved passages entail the

claim. Combining these two steps, (1) Kim proposed a modular pipeline that abstracts the connection between retrieval and entailment, allowing for easy integration of different models.

4 Method and Experimental Setup

4.1 Data Collection and Preparation

We scraped data from 256 books, each containing over 256,000 tokens, from the BookSum dataset. This dataset includes summaries at four levels of depth (0, 1, 2, and 3) and input chunks, each with 512 tokens. Each input chunk is mapped one-to-one with a depth 3 summary. Depth 2 summaries are generated by summarizing chunks of depth 3 summaries.

To create our entailment dataset, we first broke down each depth 3 summary into individual claims and mapped each claim to the corresponding input chunk of the book, resulting in over 44,000 data points. Additionally, we mapped depth 2 data to input chunks by conducting a 2-gram overlap between depth 2 claims and depth 3 summary chunks, then associating the depth 2 claim with the input chunk linked to the depth 3 summary chunk with the highest overlap.

The inclusion of depth 2 data poses a more challenging task compared to depth 3 summaries due to the necessity of inferring claims from multiple sections of the text, reflecting more realistic and complex real-world applications.

For negative examples, we used the BART-based sequence-to-sequence model from Hugging Face to generate sentences that are the opposite of the original input chunk. Our investigation confirmed that these generated sentences accurately represent the opposite meaning of the original text, making them suitable for creating non-entailed data points. The inclusion of negative examples helps balance the dataset and addresses performance discrepancies observed in previous works like FABLES, where F1 scores for negative and positive examples differed significantly.

4.2 Pipeline and Models

We implemented a pipeline incorporating two retrieval models—Dense Passage Retrieval (DPR) and BM25—and one entailment model, Facebook's BART-large fine-tuned on the MNLI dataset. Our primary focus was on the retrieval models, based on the hypothesis that better retrieval improves the overall performance of faithfulness evaluation.

- **Retrieval Models**:
  - **BM25**: A classic term frequency-inverse document frequency (TF-IDF) based retrieval model.
  - **DPR**: A dense retrieval model known for its capability to be fine-tuned, potentially yielding better retrieval results when adjusted for specific tasks.
- **Entailment Model**:
  - **BART-large fine-tuned on MNLI**: This model performs the final entailment evaluation, determining whether the hypothesis (claim) is true based on the premise (retrieved text chunk).

We assessed the standalone performance of BM25 and DPR as retrieval models, comparing their effectiveness in retrieving relevant text chunks. The objective was to demonstrate that a superior retrieval model would enhance the pipeline's overall performance in evaluating the faithfulness of claims.

4.3 Oracle Scenario

In addition to evaluating the retrieval models, we ran an oracle scenario where the entailment model received the most relevant chunk directly, bypassing the retrieval model. This setup allowed us to measure the performance difference between using a retrieval model and an oracle setup. By highlighting the gap between these methods, we aimed to emphasize the potential for improved retrieval to enhance the pipeline's overall accuracy.

5 Results

5.1 Examples of Entailed and Contradicted Sentences

To illustrate the differences between entailed and contradicted sentences, we present several examples. For a detailed mapping of depth 3 summaries to input chunks and depth 2 summaries to input chunks, please refer to the github repository.

- Laurella decide to name the baby Johnnie, after her sister.
- Laurella decides to name the baby Johnnie, after her husband.
- Pros finishes repair the cradle and removes the baby from the inside.
- Pros finishes repairing the cradle and places the baby inside.
- Johnnie's mother, Laurella, is absent from the kitchen preparing breakfast.
- Johnnie's mother, Laurella, is in the kitchen preparing breakfast.
- It's a clear November day in London.
- It's a foggy November day in London.

model_size	book_num	summary_sentence_num	summary_sentence	text_chunk	Entailment
175b	0	0	Uncle Pros is helping Laurella, his sick niece, by borrowing a cradle for her newborn baby.	THE BIRTH OF A WOMAN-CHILD	entailment
175b	0	1	Laurella explains that her husband, Consadine, is a poor provider and that he often goes off to make his fortune elsewhere.	selfish--said she'd like to know how I was	entailment
175b	0	2	Pros takes the cradle outside to get the first of the evening sun on his task.	rich, broken light from the cavernous	contradiction
175b	0	3	Mavity sends Bud and Mandy Ann to ask her father, Gideon Himes, to look in a green chest for a spotted poke that contains something he wants.		entailment
175b	0	4	Pros finishes repair the cradle and removes the baby from the inside.	she wants--ain't ye, Pretty?"	contradiction
175b	0	5	Laurella decide to name the baby Johnnie, after her sister.	hit the name that should 'a' went with	contradiction
175b	0	6	Johnnie kneels and gazes at a pink moccasin flower growing by the spring.	thread of water which trickled from the	entailment
175b	0	7	Johnnie's mother, Laurella, is absent from the kitchen preparing breakfast.	walk. Her mother would get up too, and	contradiction
175b	0	8	Laurella reluctantly refuses to let Johnnie go to work at the cotton mill.	and able."	contradiction
175b	0	9	Johnnie kisses the baby and says goodbye to her mother before leaving to go to Cottonville to work in the cotton mill.	Sis's goin' to--"	entailment
175b	1	0	It's a clear November day in London.	London. Michaelmas term lately over,	contradiction
175b	1	1	The Lord High Chancellor sits in his court, surrounded by crimson curtains and a foggy glory.	sitting here--as here he is--with a foggy	entailment
175b	1	2	Besides the Lord Chancellor, the counsel in the cause, a few other lawyers, and the solicitors, there are a few other people in the court this afternoon.	here!"	entailment
175b	1	3	The Jarndyce and Jarndyce case has been dragging on for years, with no end in sight.	minutes without coming to a total	entailment
175b	1	4	The suitors in the case have been kept at bay by the solicitors' boys who have been telling them that Mr. Chizzle, Mizzle, or otherwise is too busy to see the	trickery, evasion, procrastination,	entailment
175b	1	5	The Chancellor rises, and the bar refuses to follow suit.		contradiction
175b	1	6	It's a foggy afternoon, and the world of fashion is not unlike the Court of Chancery.	It is but a glimpse of the world of fashion	entailment
175b	1	7	My Lady Dedlock, who is childless, has left her estate in Lincolnshire to the rain, crows, rabbits, deer, and partridges.	on the stone terrace in the foreground	entailment
175b	1	8	Sir Leicester is ceremonious, stately, and arrogant.	and whiskers, his fine shirt-frill, his pure-	contradiction
175b	1	9	Mr. Tulkinghorn, the legal adviser of the Dedlocks, is led by a powdered Mercury to my Lady's presence.	Chancery, who has the honour of acting	entailment
175b	2	0	Phileas Fogg lives in a house in Burlington Gardens in London in 1872.		entailment
175b	2	1	Phileas Fogg has lived in London for many years.	It was at least certain that Phileas Fogg	contradiction
175b	2	2	Phileas Fogg is a very punctual man, and at exactly half past eleven he will leave his house to go to the Reform Club.	almost superhumanly prompt and	entailment
175b	2	3	Passepartout observes Mr. Fogg carefully during their brief interview, and comes to the conclusion that he is a calm, well-balanced man who is very exact		entailment
175b	2	4	Passepartout is a weak, thin man with brown hair.	his figure almost portly and well-built,	contradiction
175b	2	5	Passepartout inspects the house and is impressed by the order and regularity of everything.	inspection, proved to be a programme of	entailment

Figure 1: Examples of Entailment Data

5.2 Retrieval Results

We compared the performance of BM25 and DPR retrieval models. In the recall@k=1 setting, BM25 slightly outperforms DPR. However, as the value of k increases, DPR's performance surpasses BM25. This indicates that if the context window of the entailment model is sufficiently large, it can effectively capture the relevant input chunk.

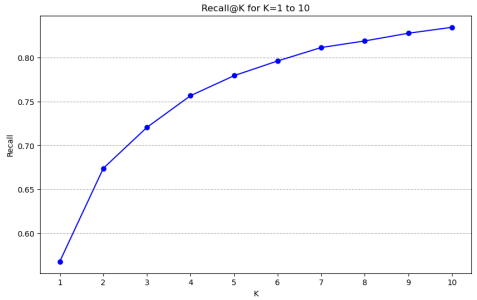

Figure 2: Retrieval performance of DPR across different recall@k settings.

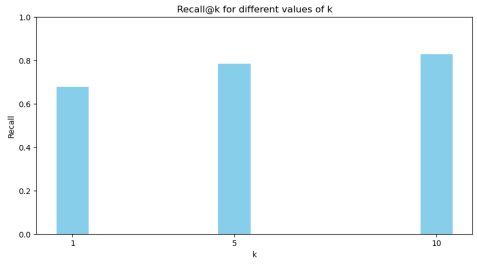

Figure 3: Retrieval performance of BM25 across different recall@k settings.

5.3 Pipeline Performance

When integrated into the pipeline, the performance of BM25 and DPR retrieval models is similar (refer to Figure 8 for the confusion matrix). However, when compared to the oracle scenario—where the entailment model receives the most relevant chunk directly—the oracle retrieved chunks significantly outperform both BM25 and DPR in terms of recall scores (refer to Figure 6).

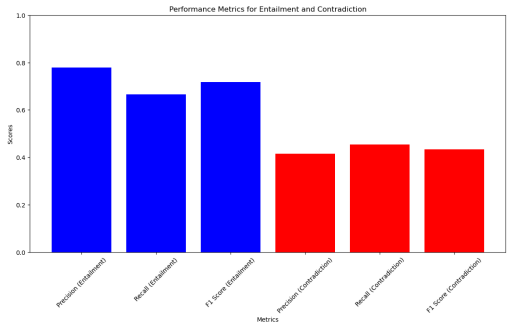

Figure 4: Performance Metrics for Pipeline with BM25

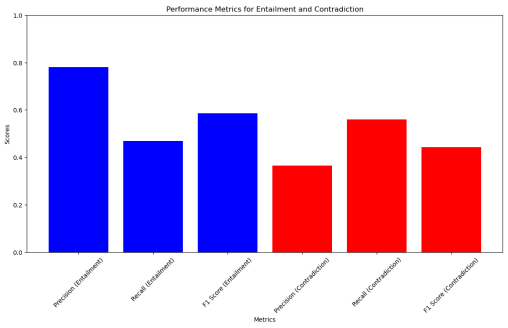

Figure 5: Performance Metrics for Pipeline with DPR

6 Analysis

6.1 Fine-Tuning DPR

The results suggest that the DPR model requires further fine-tuning to improve its performance over BM25 in the retrieval task. Fine-tuning DPR could potentially enhance the overall pipeline performance. Given that DPR has the capability to learn and adapt to specific domains, targeted fine-tuning on our dataset could yield significant improvements in retrieving relevant chunks.

6.2 Efficiency of Pre-trained Models

Even with a pre-trained model, DPR performs significantly better than no-context entailment. This indicates that the use of retrieval models, even in their pre-trained state, adds substantial value to the entailment process. The oracle performance, where the most relevant chunk is directly provided to the entailment model, approaches the performance of a model like GPT-4 Turbo that processes the entire book (>=125K tokens) as the premise. This comparison underscores the efficiency of our approach, which uses only 512 or 1024 tokens for the premise in entailment, demonstrating a more resource-effective solution without substantial loss in accuracy.

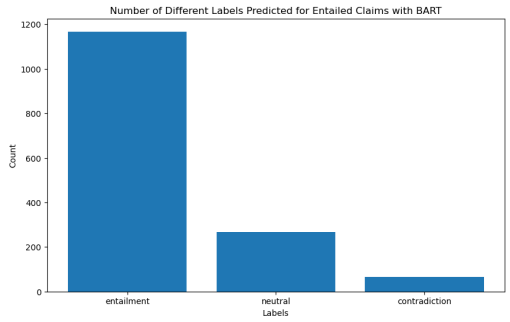

Figure 6: Distribution of Different Labels Predicted for Entailment Data

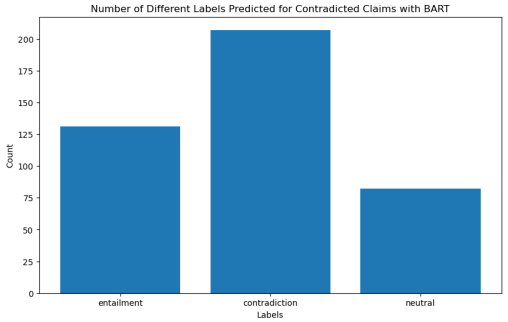

Figure 7: Distribution of Different Labels Predicted for Contradiction Data

6.3 Challenges with Depth 2 Summaries

The depth 2 summaries prove to be much harder to evaluate than depth 3 summaries when used in the pipeline. Depth 2 summaries involve higher levels of abstraction and inference, making them more challenging to map accurately to input chunks. This presents an interesting area for future research, as depth 2 summaries are harder to retrieve and less directly associated with the actual text. Improving retrieval techniques for such complex summaries could further enhance the accuracy and applicability of our pipeline in real-world scenarios.

References

[1] Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, Mohit Iyyer. FABLES: Evaluating faithfulness and content selection in book-length summarization. arXiv preprint arXiv:2404.10774, 2024. URL: `https://arxiv.org/abs/2404.10774`.

[2] Liyan Tang, Philippe Laban, Greg Durrett. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. arXiv preprint arXiv:2404.01261, 2024. URL: `https://arxiv.org/abs/2404.01261`.

[3] Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, Kyle Lo. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. arXiv preprint arXiv:2301.13298, 2023. URL: `https://arxiv.org/abs/2301.13298`.

[4] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, Greg Durrett. WiCE: Real-World Entailment for Claims in Wikipedia. arXiv preprint arXiv:2303.01432, 2023. URL: `https://arxiv.org/abs/2303.01432`.

[5] Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. *Proceedings of the ACL 2020*, 7881-7892.

Appendix

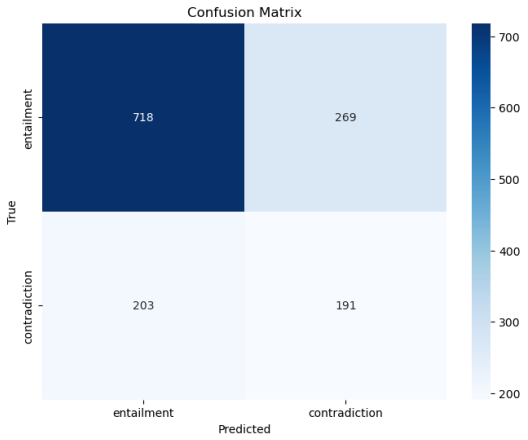

Figure 8: Confusion Matrix for BM25 Pipeline

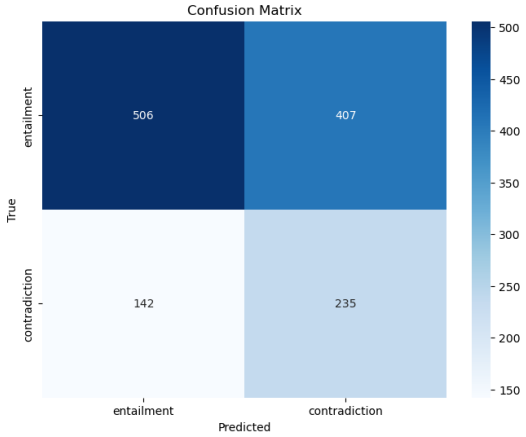

Figure 9: Confusion Matrix for DPR Pipeline

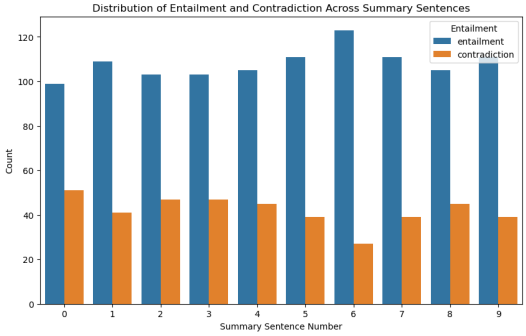

Figure 10: Data Distribution Entailment

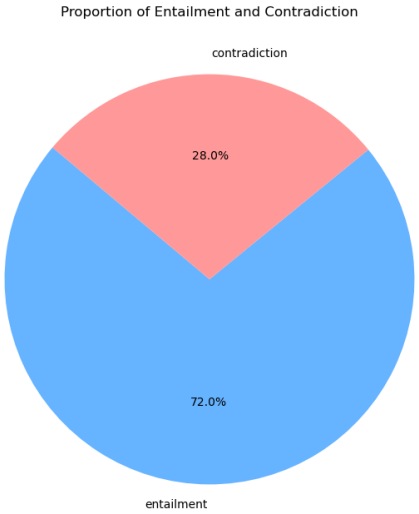

Figure 11: Pie Chart Data Distribution

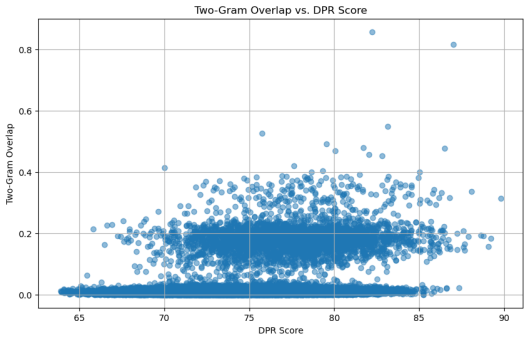

Figure 12: Two-Gram Overlap vs. DPR Score

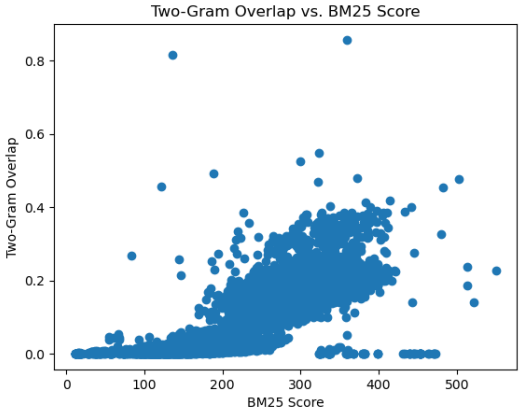

Figure 13: Two-Gram Overlap vs. BM25 Score