A Straightforward Pipeline for Targeted Entailment and Contradiction Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Finding the relationships between sentences in a document is crucial for tasks like fact-checking, argument mining, and text summarization. A key challenge is to identify which sentences act as premises or contradictions for a specific claim. Existing methods often face a trade-off: transformer attention mechanisms can identify salient textual connections but lack explicit semantic labels, while Natural Language Inference (NLI) models can classify relationships between sentence pairs but operate independently of contextual saliency. In this work, we introduce a method that combines the strengths of both approaches for a targeted analysis. Our pipeline first identifies candidate sentences that are contextually relevant to a user-selected target sentence by aggregating token-level attention scores. It then uses a pretrained NLI model to classify each candidate as a premise (entailment) or contradiction. By filtering NLI-identified relationships with attention-based saliency scores, our method efficiently isolates the most significant semantic relationships for any given claim in a text.

1 Introduction

Large-scale text data has increased the need for tools that can help humans understand complex documents efficiently. The challenge is not in one of information retrieval, but of sens-making: in an era of sophisticated misinformation and information overload, the ability to critically evaluate a claim is essential. Whether analyzing a news article, a scientific paper, a legal contract, or a corporate report, a user's primary goal is often to verify a specific statement. This requires untangling the document's argumentative structure to answer fundamental questions: (1) What evidence is presented to support this claim? (2) Are there any statements within the text that contradict it? Manually performing this analysis is time-consuming and prone to error, creating a pressing need for automated tools that facilitate targeted critical analysis.

Large Language Models (LLMs) have demonstrated capabilities in text understanding. The self-attention mechanism, a core component of their architecture [6]. It is particularly tailored to capture dependencies between tokens. The resulting attention maps can be interpreted as a measure of token-level saliency, indicating which parts of a text are important for understanding other parts. However, these attention scores lack explicit semantic meaning: high attention between two phrases does not, on its own, tell us if one supports or contradicts the other.

While no existing work offers a direct, integrated solution for our targeted analysis task, related research falls into three main categories, each with significant limitations. The first is the field of Natural Language Inference (NLI), which has produced models specifically trained to recognize semantic relationships between a sentence pair (a *premise* and a *hypothesis*) [8]. Models like `facebook/bart-large-mnli` [4] can classify these pairs with high accuracy. The primary drawback of pure NLI is its context-agnostic nature and computational inefficiency. To find relationships for a target sentence, one would need to compare it against every other sentence in the document, a process

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

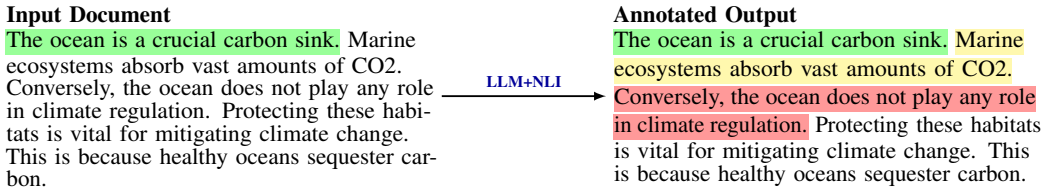

Figure 1: Given an input with a user-selected **target** sentence (green), our method uses attention saliency and NLI to automatically identify and highlight the corresponding **premise** (yellow) and **contradiction** (red).

with quadratic complexity ($O(n^2)$) that is impractical for long texts. More importantly, this approach treats all sentence pairs equally, failing to distinguish between relationships that are central to the document's discourse and those that are trivial or incidental.

A second related area is Argument Mining, which aims to automatically extract argumentative components (e.g., claims, premises) and their relations (e.g., support, contradiction) from text [5]. While conceptually similar, argument mining systems are often complex, heavyweight solutions that require extensive, domain-specific annotated data for training. They are typically designed for structured, formal argumentation and may not be easily adaptable as a lightweight, general-purpose tool for interactive text exploration. Their complexity stands in contrast to our goal of a straightforward, model-agnostic pipeline.

Finally, a third line of work involves the interpretation and visualization of attention mechanisms themselves [7]. These studies provide valuable insights into how LLMs process information but are primarily diagnostic in nature. They reveal the existence of connections but do not assign explicit semantic labels to them. A high attention score signals relevance but fails to specify the nature of that relevance, be it support, contradiction, elaboration, or simple co-reference.

This work addresses the gap left by these approaches. We propose a two-stage method that integrates causal LLM attention with NLI to create a system that is both context-aware and semantically explicit. Our core contribution is the idea that a meaningful semantic relationship between a target sentence S_{target} and another sentence S_j should be characterized by both: (1) a high degree of mutual attention, indicating contextual saliency, and (2) a clear "entailment" or "contradiction" label from an NLI model. By combining these signals, our method efficiently filters for relevant premises and contradictions, enabling a practical and targeted analysis of any claim within a document.

2 Method

To map the logical structure surrounding a specific claim in a document, our method operates on a simple yet powerful principle: a significant relationship between a target sentence and another sentence must be both **contextually salient** and **semantically explicit**. Relying on one signal alone is insufficient. Raw attention scores from a language model, while indicative of textual connections, are often noisy and only show that two passages are related, but not *how*. Conversely, applying an NLI model to compare a target sentence against all others is computationally expensive and context-agnostic.

Our pipeline integrates these two signals to overcome their individual limitations. It is a three-step process for analyzing a chosen target sentence, S_{target}: (1) we use attention as a wide-angle spotlight to efficiently find potentially important connections to S_{target}, (2) we use NLI as a precision lens to classify their meaning relative to S_{target}, and (3) we fuse these signals to render a final verdict on the core premises and contradictions.

2.1 Identifying Contextual Saliency with Attention

The first step is to efficiently identify which sentences are worth comparing to our target sentence, S_{target}. The attention mechanism of a causal language model provides a robust signal for this task. We extract attention weights from the **final layer** of the transformer, as these layers are known to encode more abstract and semantic dependencies crucial for understanding discourse structure [2].

The process begins by feeding the entire document into a pre-trained language model and extracting the token-to-token attention matrix, A_{tok} of the entire document. To create a sentence-level view, we aggregate these token-level scores with sentence tokenizer [1]. The saliency of sentence S_j relative

to sentence S_i is defined as the mean attention score from all tokens in S_i to all tokens in S_j:

$$A_{sent}[i, j] = \frac{1}{|S_i||S_j|} \sum_{k \in S_i} \sum_{l \in S_j} A_{tok}[k, l] \tag{1}$$

This yields a saliency matrix, A_{sent}, which serves as a weighted, directed graph. To find candidate sentences for our target, S_{target}, we select all other sentences S_j where $A_{sent}[\text{target}, j]$ exceeds a certain threshold, indicating a strong, contextually-aware connection that warrants further semantic analysis.

2.2 Classifying Semantic Relationships with NLI

While the saliency matrix tells us **where** to look, it does not tell us **what** we are seeing. The second stage addresses this by assigning precise semantic labels to the filtered candidate sentences. For this, we employ a pre-trained NLI model.

For each candidate sentence S_c identified in the previous step, we form two ordered pairs with our target sentence, S_{target}, to check for both premises and contradictions:

1. **To find premises**: We treat the candidate as the premise and the target as the hypothesis, (S_c, S_{target}). If the NLI model classifies this pair as 'entailment', we label S_c as a **premise** for our target.
2. **To find contradictions**: We treat the target as the premise and the candidate as the hypothesis, (S_{target}, S_c). If the model returns 'contradiction', we label S_c as a **contradiction**.

Pairs classified as 'neutral' are discarded. This stage produces a relationship matrix, R, which exhaustively labels every potential semantic link to the target sentence.

2.3 Combining Saliency and Semantics

The final step confirms the significance of the identified relationships. The NLI process, even on filtered candidates, can sometimes flag logically valid but contextually weak connections. The attention-based saliency scores provide the necessary final filter.

We formalize our rule: a sentence S_c is confirmed as a premise or contradiction of S_{target} only if it receives the appropriate NLI label from stage two **and** its saliency score $A_{sent}[\text{target}, c]$ surpasses a predetermined threshold. This fusion ensures that our final output only contains relationships to the target sentence that are both semantically unambiguous and validated as important by the language model's contextual understanding of the entire document.

3 Experiments and Results

To show the plausibility and effectiveness of our targeted analysis, we designed three test cases with increasing complexity. For each case, we selected a target sentence and applied our pipeline to identify its corresponding premises and contradictions within the text.

To validate our pipeline and underscore its accessibility, all experiments were performed on a conventional laptop (Dell XPS with an Intel i7 processor and 16 GB of RAM), deliberately avoiding the need for specialized GPU hardware. For the attention-based saliency analysis described in we utilized the `Qwen/Qwen3-1.7B` [9]. Following this, for the semantic classification stage, we evaluated several publicly available NLI models. We empirically determined that the `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` [3] model delivered the most robust and accurate performance across our diverse test cases.

Case 1: Direct Factual Relationships. We first analyzed a short text containing a simple factual claim to test the system's ability to identify direct support and contradiction. The target sentence selected was "*The sun is a star.*"

The sun is a star. It is the center of our solar system. The sun is a planet. All planets revolve around it.

Figure 2: Analysis of the target sentence (green). The system correctly identifies one premise (yellow) and one contradiction (red) that are both semantically valid and contextually salient.

As shown in Figure fig:example1, our system correctly identified two key relationships. The sentence "*It is the center of our solar system*" was flagged as a premise (entailment), while "*The sun is a planet*" was flagged as a contradiction. Both sentences exhibited high attention scores relative to the target, allowing them to pass the saliency filter and be confirmed by the NLI model. Specifically, the attention between the target and the contradiction was 0.1288, significantly above the text's average inter-sentence attention of 0.0349.

Case 2: Implied Contradiction. Next, we crafted a narrative text where a contradiction arises from a logical inconsistency rather than a direct factual opposition. The target sentence was the initial statement of intent: "*Mark decided to build a bookshelf from scratch.*"

> Mark decided to build a bookshelf from scratch. He started by carefully measuring the space in his living room. Next, he bought high-quality oak wood and cut each piece to the exact size. He spent a full weekend sanding, assembling, and staining the bookshelf. He found that IKEA fits perfectly to his requirements. In the end, the bookshelf was sturdy, fit perfectly in the space, and looked professionally made.

Figure 3: Detection of an implied contradiction. The outcome (finding an IKEA solution) logically contradicts the initial intention to build from scratch (target, green).

The system successfully identified the subtle contradiction shown in Figure fig:example2. The statement "*He found that IKEA fits perfectly to his requirements*" logically undermines the initial goal. This relationship was captured because the attention score between the two sentences (0.0768) was a standout, more than double the standard deviation (0.0284) above the mean attention (0.0135) for this text. This demonstrates the model's ability to connect distant but semantically opposed concepts.

Case 3: Complex Argument Analysis. Finally, we analyzed a more nuanced text involving a central claim supported by evidence and challenged by a counter-argument. We targeted the sentence presenting specific evidence: "*In fact, several pilot programs have reported savings of up to 60% on lighting expenses after switching to LEDs.*"

> Many cities are exploring the idea of replacing traditional streetlights with smart LED systems. These smart lights are far more energy-efficient than conventional bulbs, helping municipalities cut electricity costs. In fact, several pilot programs have reported savings of up to 60% on lighting expenses after switching to LEDs. However, the data systems that control these lights require regular software updates and cybersecurity measures, which have added unexpected ongoing costs for some cities. In some cases, these challenges have led municipalities to abandon LED upgrades altogether and return to conventional lighting.

Figure 4: A unified analysis of a complex argument. For the target evidence (green), the system identifies the broader claim it supports as a premise (yellow) and a counter-argument that challenges its implications as a contradiction (red).

The results in Figure fig:example3 showcase the system's ability to parse a multi-faceted argument. When analyzing the target evidence, it correctly identified the preceding general claim ("*These smart lights are far more energy-efficient...*") as its premise. The attention score between this pair (0.0024) was sufficient to proceed to NLI, which confirmed the entailment. Furthermore, it identified the sentence about cities abandoning the upgrades as a contradiction. Although its attention score was lower (0.0008), it was still deemed salient enough in the local context to be evaluated, and the NLI model confirmed its contradictory nature relative to the successful pilot programs.

4 Discussion and Future Work

The results from our test cases suggest that this two-staged approach is a promising direction for interactive text analysis tools. By allowing users to select a target sentence, our method enables focused and meaningful exploration of a document's argumentative structure.

By combining contextual saliency from attention with explicit classification from NLI models, our method offers a more robust analysis than either technique alone. The attention filter narrows the search space to sentence pairs the model deems interconnected, while NLI labels provide the semantic grounding that raw attention scores lack. An additional advantage is that our approach is not tied to any specific high-quality LLM; since it relies on the attention mechanism common to transformer models, it can be applied broadly and will likely improve automatically as LLM capabilities advance.

While experimenting, we identified two main limitations. (1) Its performance depends on the quality of the underlying models: attention patterns may vary, and models not tuned for structured reasoning can produce noisy saliency maps. (2) NLI models, while accurate, still struggle with nuance, sarcasm, or complex syntax. Our aggregation of token-level attention to the sentence level via arithmetic mean is also a simplification; more sophisticated methods could yield more precise saliency scores. Future work should explore user-centric design, enabling users to adjust the saliency threshold to control analysis granularity.

References

[1] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72, 2006.

[2] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL*, 2019. URL `https://aclanthology.org/P19-1356.pdf`.

[3] Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100, 2024.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[5] Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[7] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.

[8] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[9] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are directly supported by the methods described and the results presented in the experimental section.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The 'Discussion and Future Work' section explicitly addresses the limitations of the proposed method, including its dependency on the quality of the underlying language and NLI models.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper presents an empirical methodology and does not introduce new theoretical results that would require mathematical proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides a detailed description of the methodology, including the formula for aggregating attention scores. Furthermore, the case studies include specific attention values and statistical context (mean, std. dev.) to aid in the interpretation and potential reproduction of the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A Streamlit script used for evaluation and experimentation will be provided as supplemental material. It is a self-contained, interactive tool that can be run directly to reproduce our analysis. Code avilable at https://tinyurl.com/causcineuralips

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: As our method does not involve training, there are no training-specific hyperparameters. We clearly specify the pre-trained models used for both attention extraction and NLI classification, as well as the hardware environment.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper presents qualitative results on illustrative case studies rather than a quantitative evaluation on a benchmark dataset. Therefore, statistical significance tests or error bars are not applicable to the presented results.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the 'Experiments and Results' section specifies the hardware used (CPU type, RAM), confirming that the method can be run on consumer-grade equipment.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Answer: [Yes]

   Justification: Yes, the research adheres to the NeurIPS Code of Ethics. The work utilizes publicly available models for a general-purpose text analysis task.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction discusses the positive societal impact of enabling users to detect contradictions and verify claims. While the tool is designed for beneficial purposes like fact-checking, a potential negative impact could involve misusing the tool to selectively find apparent contradictions out of context to support a misleading narrative. However, the interactive nature of the tool encourages a deeper exploration of the text, which can mitigate this risk.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work uses publicly released, pre-trained language models and does not introduce new datasets or models that pose a high risk of misuse. Therefore, specific safeguards for model releases are not applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all models and architectures used are properly credited in the paper via citations to the original publications.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There are no new assets introduced in this paper.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research did not involve any crowdsourcing or experiments with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This research did not involve human subjects, so IRB approval was not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: Large Language Models were used for ancillary tasks such as improving the clarity of the writing and assisting in code development. They were not a component of the core methodology itself.