

LLMs and Reasoning in Health Insurance Policy Documents: A Dataset and Argument Graph Extraction Study

Govind Arun¹ Sidharth N Krishna² Susie Xi Rao^{3,4} Ajo Babu George⁵

¹University of Maryland, College Park, MD, USA

²Indira Gandhi National Open University, New Delhi, India

³School of Engineering and Computer Science, Bern University of Applied Sciences;

⁴Department of Management, Technology and Economics (MTEC) & AI Center, ETH Zurich, Switzerland

⁵DiceMed, Cuttack, Odisha, India

govind123.ga@gmail.com sid4official@gmail.com
srao@ethz.ch drajobabu7@gmail.com

Abstract

Health insurance policy documents justify coverage decisions through clinical evidence and regulatory criteria, but the underlying reasoning is buried in dense medical and legal text. A pipeline is presented that extracts argument components and relations from these documents and assembles them into directed argument graphs. A dataset of 5,230 sentences and 5,685 relations across 53 documents from Aetna, CMS, and United Healthcare is constructed along with two custom annotation tools. Six classical baselines, three fine-tuned transformer models, and three instruction-tuned LLMs are evaluated, with the LLMs assessed under zero-shot, few-shot, and chain of thought style prompting. **PubMedBERT** and **Legal-BERT**, fine-tuned on the dataset, achieve the strongest results overall (0.731 and 0.653 Macro F1 on component classification and relation extraction, respectively). Notably, the much simpler Term Frequency–Inverse Document Frequency (TF-IDF) + Support Vector Machine (SVM) baseline is competitive with these fine-tuned transformers (0.709 and 0.649 Macro F1) and outperforms the best prompted small (4-7B parameter) instruction-tuned LLM by 12 and 21 F1 points respectively. The dataset and code are available at <https://github.com/GIND123/LLM-Driven-Argument-Mapping>.

1 Introduction

Health insurance policy documents contain structured reasoning in which coverage decisions are justified through clinical evidence and policy rules. Argument mining, the automatic identification of claims, premises, and their relations, has been applied to persuasive essays (Stab and Gurevych, 2017) and legal texts (Peldszus and Stede, 2015),

but health insurance policies remain largely unexplored. The central question is whether instruction-tuned LLMs can match classical feature based methods for argument mining in formulaic institutional text, and whether reason-oriented style prompting narrows the gap. The contributions are:

- A dataset of 53 policy documents with 5,230 component labels and 5,685 relation annotations (Krippendorff’s $\alpha=0.79$ components, $\alpha=0.74$ relations)
- Two web-based annotation tools for sentence level component labelling and Sentence-BERT based relation candidate ranking.
- Systematic comparison of six classical baselines, three fine-tuned transformer models, and three LLMs under zero-shot, few-shot, and CoT-style prompting (Wei et al., 2022), showing that fine-tuned transformers achieve the strongest results overall, while TF-IDF pipelines remain competitive and substantially outperform the tested 4–7B instruction-tuned LLMs.

2 Related Work

Stab and Gurevych (2017) established claim premise identification in persuasive essays, reporting Macro F1 of approximately 0.73 for components and 0.58 for relations. Peldszus and Stede (2015) achieved comparable results (0.60–0.70 F1) on microtext corpora. Surveys by Lippi and Torroni (2016) and Lawrence and Reed (2020) identify institutional domains as underexplored. Chain of thought prompting (Wei et al., 2022) has improved LLM reasoning and recent work has explored reasoning structure and uncertainty in clinical prediction (Arun et al., 2026), as well as increasingly sophisticated forms of social, strategic,

and multi-agent reasoning in LLMs (Wang et al., 2026). However, domain-specific lexical features remain competitive in formulaic domains (Mosbach et al., 2023). Recent comparative work in health-related text classification has similarly evaluated traditional machine learning and language model approaches (Lee et al., 2026). Health insurance documents exemplify this, with recurring phrases such as “considered experimental and investigational” serving as strong classification signals.

3 Dataset and Annotation

3.1 Data Sources

Documents are collected from three sources: **Aetna Clinical Policy Bulletins** (Aetna, 2024), **CMS National Coverage Determinations** (Centers for Medicare & Medicaid Services, 2024), and **United Healthcare medical policies** (UnitedHealthcare, 2024). These documents are medical necessity policy documents, formal institutional texts that define the clinical criteria under which specific procedures, treatments, or diagnostic tests are covered or denied.

3.2 Annotation Tools and Scheme

Three annotators labeled the corpus: a physician with ten years of clinical experience, an AI/ML research engineer, and an applied statistics graduate student with health informatics experience, ensuring both domain expertise in interpreting clinical policy language and technical rigor in applying categorical labels consistently.

A Claim asserts a coverage decision or policy determination (e.g., “For Aduhelm, member must be enrolled in a clinical trial”). A Premise provides clinical evidence or diagnostic criteria supporting or challenging a claim (e.g., “It is a neurologic condition characterized by loss of mental ability severe enough to interfere with normal activities of daily living, lasting at least six months and not present from birth”). A Non-Argumentative sentence contains background, procedural, or inconclusive research content that does not directly support or oppose any coverage decision (e.g., “However, unless prospective studies are conducted, the issue remains open whether mean diffusivity is better as early predictor of AD than volume”).

Sentence-level annotation is adopted following established practice in argument mining for structured domains such as clinical trials (Mayer et al.,

2020) and legal judgments (Poudyal et al., 2020). Health insurance policy documents are written in a formal, regulatory style where sentences typically express a single argumentative function. A manual inspection of 532 sentences across two representative documents revealed that fewer than 13% of sentences contained discourse connectives potentially signalling multiple argument components and manual review confirmed the majority of these still expressed a single unified argumentative function. This supports sentence-level granularity as appropriate for this domain, with sub-sentential ambiguity acknowledged as a known limitation (Section 9).

The **Argument Component Annotation Tool** loads policy PDFs, segments text using spaCy, and presents each sentence for labeling. The **Argument Relation Annotation Tool** generates candidate premise–claim pairs via Sentence-BERT (SBERT) cosine similarity (Reimers and Gurevych, 2019), retaining the top 3 claims per premise. Annotators assign SUPPORT, ATTACK, or NONE. Annotators labeled only the retrieved top-3 candidates and could not introduce premise–claim pairs outside this set. Consequently, gold relations are conditioned on SBERT retrieval, and neither candidate-generation recall against an exhaustive relation set nor full argument-graph completeness is evaluated in this work (see Limitations). Existing tools such as BRAT do not natively support cross-sentence relation annotation with a document-level DAG view; the custom tools were designed to streamline the two-stage workflow and integrate directly with the dataset format and export pipeline.

Inter-annotator agreement across all three annotators, computed on 8 randomly sampled documents, yields Krippendorff’s $\alpha = 0.79$ for component classification and $\alpha = 0.74$ for relation labeling. The lower α on relations reflects the inherent difficulty of distinguishing ATTACK from SUPPORT when clinical evidence language is similar. Disagreements were resolved by majority vote under the supervising researcher’s adjudication.

3.3 Dataset Statistics

Table 1 summarises the corpus. PREMISE dominates component labels (56%), reflecting extensive clinical evidence supporting fewer coverage decisions. SUPPORT relations form the majority (58%), followed by NONE relations (31%), while ATTACK relations remain relatively rare (10%) typically appearing in policies where conflicting

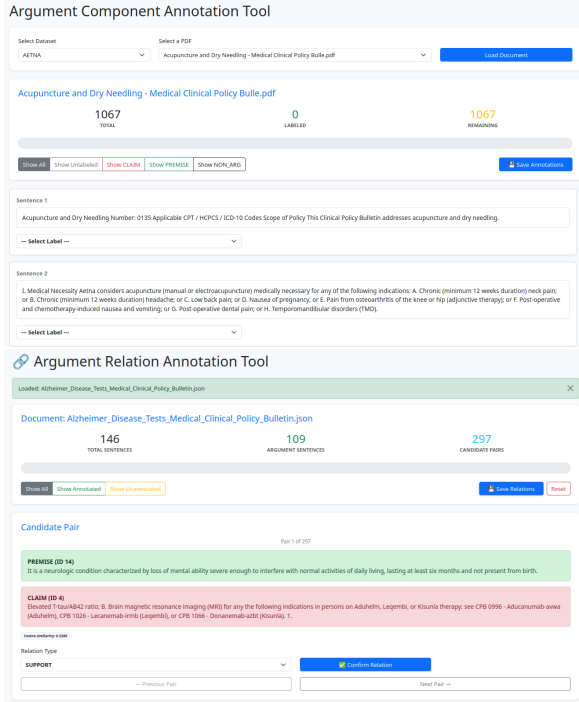


Figure 1: Argument Component Annotation Tool (top) and Argument Relation Annotation Tool with SBERT ranked candidate pairs (bottom).

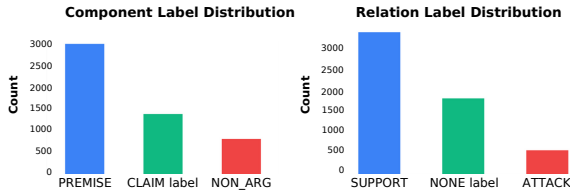


Figure 2: Distribution of argument component labels (left) and relation labels (right) in the dataset.

clinical evidence is discussed.

4 Methodology

Argument Component Classification: Each sentence is classified as CLAIM, PREMISE, or NON_ARGUMENTATIVE. Classical baselines apply Logistic Regression (LR), Linear Support Vector Machine (SVM), and Random Forest (RF) on TF-IDF vectors (unigrams to trigrams, 20K features) or SBERT embeddings (Reimers and Gurevych, 2019), with balanced class weights. LLM classifiers use instruction-tuned models (Section 5).

Relation Candidate Generation and Classification: SBERT cosine similarity retains the top 3 most similar claims per premise within each document. Biomedical pretrained language models have also been explored for relation

Statistic	Components	Relations
Documents	53	53
Total instances	5,230	5,685
CLAIM / SUPPORT	1,414	3,313
PREMISE / NONE	2,949	1,784
NON_ARG / ATTACK	867	588
Train / Test	4,184 / 1,046	4,548 / 1,137

Table 1: Dataset statistics for argument components and relations.

extraction in specialized biomedical text (Bhasuran, 2022). Each pair is classified as SUPPORT, ATTACK, or NONE using the same model families, with input as Source: [premise] Target: [claim].

5 Experiments

Models: The full pipeline, shown in Figure 3, proceeds in four stages: component classification, candidate relation generation, relation classification, and argument graph construction. Classical baselines pair TF-IDF or SBERT with LR, SVM, and RF. Three LLMs are evaluated: Qwen2.5-7B-Instruct (Qwen Team, 2024), Mistral-7B-Instruct v0.3 (Jiang et al., 2023), and Nemotron-Mini-4B-Instruct (NVIDIA, 2024), loaded in float16 on an A100 40GB GPU with chat templates. Three transformer-based models are additionally fine-tuned as stronger baselines: BERT-base (Devlin et al., 2019), Legal-BERT (Chalkidis et al., 2020), pre-trained on legal corpora, and PubMedBERT (Gu et al., 2021), pre-trained on biomedical abstracts. Each model is fine-tuned with a linear classification head for 5 epochs ($\text{lr}=2\text{e-}5$, batch size=16, balanced class weights) on the same train/test splits as the classical baselines. For relation extraction, sentence pairs are encoded as [CLS] premise [SEP] claim [SEP].

Prompting Strategies:

- **Zero-shot:** task definitions only.
- **Few-shot:** 3 real training examples per class for component classification, and 2 per class for relation classification (due to the scarcity of ATTACK examples in the training set).
- **CoT-style:** the model is instructed to output the predicted label first, followed by a brief justification of its decision. For LLM evaluation, sentences are truncated to 100 tokens, and 200 test instances are sampled per task.

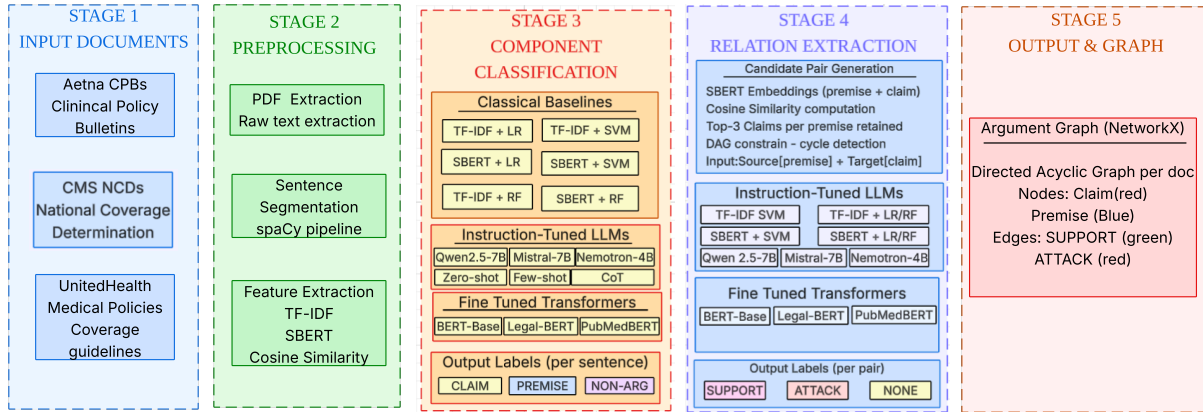


Figure 3: System pipeline: component classification, SBERT based relation candidate generation, relation classification, and argument graph construction.

[System] You are an expert in argument mining for health insurance policy documents. Classify each sentence into one of the predefined argument component labels.

[Input] Sentence: {sentence}

[Instruction] Depending on the prompting strategy: (i) zero-shot uses only the input sentence, (ii) few-shot prepends representative labelled training examples, and (iii) CoT-style prompting asks the model to output the predicted label first and then provide a brief justification.

[Output] One of: CLAIM, PREMISE, or NON_ARGUMENTATIVE. For CoT-style prompting, output the predicted label followed by a brief justification.

Figure 4: Prompt template used for argument component classification. Curly braces denote runtime-substituted variables.

All prompts share a task-specific system message providing label definitions and domain context (Jin et al., 2026). For component classification, the system message defines CLAIM, PREMISE, and NON_ARGUMENTATIVE in terms of health insurance policy language, while for relation classification it defines SUPPORT, ATTACK, and NONE in terms of evidential direction. Zero-shot prompts append only the input sentence or sentence pair with a Label: cue. Few-shot prompts prepend the specified number of representative training examples per class, sampled from the training split using a fixed random seed and restricted to sentences of at most 60 words to control prompt length. CoT-style prompts instruct the model to output a label first, followed by a brief justification. All sentences are truncated to 100 tokens. Figures 4 and 5 illustrate the prompt templates used for argument component and relation classification, respectively.

Setup: Stratified 80/20 splits yield 1,046 component and 1,137 relation test instances. Macro F1 is the primary metric. Across prompting strategies, Macro F1 varied by up to 30 points for the same model (e.g., Nemotron-4B: 0.194 zero-shot

[System] You are an expert in argument mining for health insurance policy documents. Determine the argumentative relation between two sentences.

[Input] Source: {source_sentence}
Target: {target_sentence}

[Instruction] Depending on the prompting strategy: (i) zero-shot uses only the input pair, (ii) few-shot prepends representative labelled source-target examples, and (iii) CoT-style prompting requests a brief justification after predicting the relation label.

[Output] One of: SUPPORT, ATTACK, or NONE. For CoT-style prompting, output the relation label followed by a brief explanation.

Figure 5: Prompt template used for relation classification. Curly braces denote runtime-substituted variables.

to 0.499 few-shot), confirming that prompt formulation substantially impacts performance in this domain. The impact of prompting strategy is analyzed in Section 6.3.

6 Results

6.1 Component Classification

Table 2 presents results. TF-IDF + SVM achieves 0.709 Macro F1, competitive with the 0.73 reported by Stab and Gurevych (2017) on shorter persuasive essays. Among fine-tuned transformers, PubMedBERT achieves the highest component classification **F1 (0.731)**, surpassing TF-IDF + SVM by 2.2 points, suggesting that biomedical pretraining captures domain-specific clinical terminology better than surface lexical features alone. Few-shot prompting consistently improves LLMs: Qwen2.5-7B rises from 0.466 to 0.566, Nemotron-4B from 0.194 to 0.499. CoT improves over zero-shot for some models (e.g., Nemotron-4B: 0.194 to 0.389) but does not surpass few-shot prompting for any model. LLMs show complementary strength: Qwen2.5-7B few-shot achieves 0.79 recall on NON_ARGUMENTATIVE versus 0.53 for TF-IDF + SVM. Among the 200 sampled instances used for LLM evaluation, the most frequent error is

Model	Prompt	F1	P	R
TF-IDF + SVM	baseline	.709	.726	.697
TF-IDF + LR	baseline	.703	.697	.711
TF-IDF + RF	baseline	.626	.774	.593
SBERT + SVM	baseline	.622	.617	.630
SBERT + LR	baseline	.604	.597	.630
SBERT + RF	baseline	.482	.741	.478
BERT-base	fine-tune	.708	.706	.709
Legal-BERT	fine-tune	.709	.722	.700
PubMedBERT	fine-tune	.731	.724	.738
Mistral-7B	<u>few</u>	<u>.585</u>	<u>.590</u>	.588
Qwen2.5-7B	<u>few</u>	<u>.566</u>	<u>.570</u>	<u>.626</u>
Mistral-7B	cot	.550	.546	.586
Mistral-7B	zero	.543	.543	.560
Nemotron-4B	few	.499	.568	.570
Qwen2.5-7B	zero	.466	.464	.469
Qwen2.5-7B	cot	.442	.470	.457
Nemotron-4B	cot	.389	.393	.395
Nemotron-4B	zero	.194	.291	.340

Table 2: Component classification (macro averaged). Best overall: PubMedBERT (.731). Best LLM: Mistral-7B, few-shot (.585). **Bold** = best overall, underline = best LLM.

CLAIM predicted as PREMISE (52/178); full error analysis across the complete test set is presented in Section 6.4.

6.2 Relation Extraction

Table 3 presents results. TF-IDF + SVM achieves 0.649, exceeding the 0.58 reported by [Stab and Gurevych \(2017\)](#). **Legal-BERT** achieves the best result overall on relation extraction (**0.653**), narrowly surpassing TF-IDF + SVM (0.649), suggesting that relation classification in this domain relies more on shallow lexical co-occurrence patterns than on contextual representations, though contextual pretraining still yields a modest edge. The best LLM is Mistral-7B CoT at 0.443. ATTACK (14% of the 200-instance LLM evaluation sample) is most confused with SUPPORT (25 errors in that sample). The confusion likely arises because both relation types share clinical evidence language, differing only in whether the evidence corroborates or contradicts the target claim. Notably, CoT-style prompting produces the best LLM results on this task (Mistral-7B: 0.443, Qwen2.5-7B: 0.440), suggesting that the CoT-style prompting step helps when the model must assess evidential direction rather than simply classify a single sentence.

6.3 Prompting Strategy Analysis

Figure 7 compares strategies. Few-shot prompting is the strongest strategy for component classification across all three models, improving over

Model	Prompt	F1	P	R
TF-IDF + SVM	baseline	.649	.643	.657
TF-IDF + RF	baseline	.634	.650	.625
TF-IDF + LR	baseline	.624	.617	.650
SBERT + RF	baseline	.587	.691	.564
SBERT + SVM	baseline	.545	.540	.555
SBERT + LR	baseline	.532	.530	.568
Legal-BERT	fine-tune	.653	.646	.663
BERT-base	fine-tune	.633	.621	.660
PubMedBERT	fine-tune	.634	.625	.651
Mistral-7B	<u>cot</u>	<u>.443</u>	<u>.545</u>	.448
Qwen2.5-7B	cot	.440	.444	.450
Mistral-7B	zero	.428	.452	<u>.485</u>
Nemotron-4B	cot	.356	.445	.373
Nemotron-4B	few	.356	.364	.388
Mistral-7B	few	.333	.452	.451
Qwen2.5-7B	zero	.314	.429	.406
Qwen2.5-7B	few	.302	.522	.385
Nemotron-4B	zero	.272	.323	.318

Table 3: Relation extraction (macro averaged). Best overall: Legal-BERT (.653). Best LLM: Mistral-7B, CoT (.443). **Bold** = best overall, underline = best LLM.

zero-shot by up to 30 F1 points (Nemotron-4B: 0.194 → 0.499). For relation extraction, the pattern reverses: CoT is the best or joint-best strategy for all three models (e.g., Mistral-7B: 0.443 CoT vs. 0.333 few-shot), suggesting CoT style prompting helps more when assessing evidential direction between sentence pairs than when classifying a single sentence.

6.4 Error Analysis

For component classification, TF-IDF + SVM produces 298 errors on 1,046 test instances (28.5%). The dominant confusion is CLAIM predicted as PREMISE (74 cases). Short assertive sentences with clinical language are particularly vulnerable: for example, “*There is evidence to support the use of acupuncture in migraine*” (true CLAIM) is misclassified as PREMISE because it contains the phrase “evidence to support”, a strong lexical marker of premise patterns. The second largest confusion is NON-ARGUMENTATIVE predicted as PREMISE (65 cases), where methodological sentences such as “*Results were compared to phenotypic drug susceptibility testing to assess analytical robustness*” share surface features with clinical evidence statements. PREMISE predicted as CLAIM (60 cases) occurs when premises use assertive, policy-like language, e.g., “*Diminished blood flow to the fracture site will often suppress the healing response*”, which resembles a coverage assertion. Misclassified sentences are on average shorter (38.3 words) than cor-

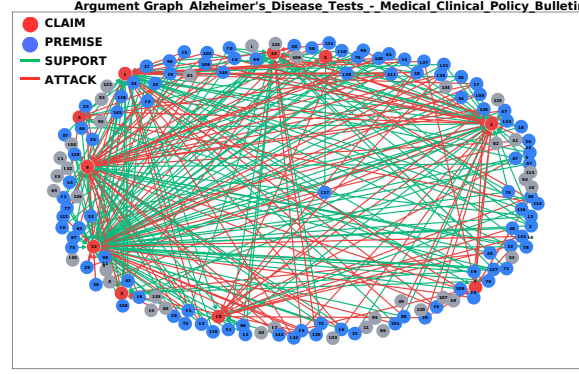


Figure 6: Argument graph for Alzheimer’s Disease Tests. Red: claims, blue: premises. Green: SUPPORT, red: ATTACK.

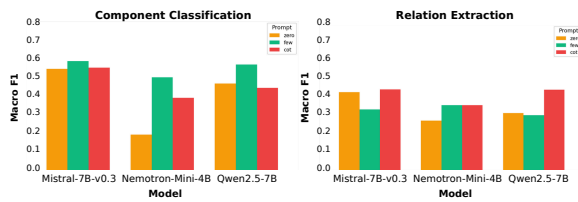


Figure 7: Macro F1 by prompting strategy for component classification (left) and relation extraction (right).

rectly classified ones (42.6 words), suggesting that shorter sentences lack sufficient lexical context for disambiguation. Error rates vary by source: UHC documents show the highest error rate (35.9%), followed by CMS NCD (33.3%), while Aetna Clinical Policy Bulletins show the lowest (22.0%), likely reflecting greater terminological consistency in the latter.

For relation extraction, ATTACK is the hardest class ($F1 = 0.41$) on the full test set, with 47 instances misclassified as SUPPORT and 21 as NONE (compare the smaller counts in Section 6.2, drawn from the 200-instance LLM sample). Attacks in policy documents are expressed through subtle hedging rather than explicit contradiction: for instance, the source sentence “*Thus, use of LDCT screening cannot yet be recommended outside of the investigational setting*” attacks the target claim that LDCT is medically necessary, yet the model predicts SUPPORT because both sentences share LDCT terminology. SUPPORT→NONE is the most frequent relation error (101 cases), where weakly-worded supporting evidence lacks the strong causal markers the model relies on. Full confusion matrices and all error examples are available in the repository.

7 Argument Graph Analysis

As a downstream application of the pipeline, extracted components and relations are assembled into directed argument graphs for document-level visualization of reasoning patterns in coverage decisions. Directed argument graphs are constructed per document using NetworkX (Hagberg et al., 2008). Figure 6 shows the Alzheimer’s Disease Tests policy (146 nodes, 297 edges). Across 53 documents, premises converge on few claim nodes averaging 2.4 incoming support edges per claim. ATTACK relations concentrate where evidence conflicts: cancer vaccines (17), back pain (41), ablative spinal treatments (70).

8 Conclusion

A pipeline for argument mining in health insurance policy documents is presented with a dataset of 5,230 sentences across 53 documents and two custom annotation tools. Fine-tuned transformers achieve the strongest results overall, with PubMedBERT best on component classification (0.731) and Legal-BERT best on relation extraction (0.653). The much simpler TF-IDF + SVM baseline remains competitive (0.709 and 0.649) and outperforms all prompted configurations of the tested (4–7B) instruction-tuned LLMs (0.585 and 0.443). LLMs show complementary strengths on minority classes, and parameter-efficient fine-tuning (Dettmers et al., 2023) may close the gap. The dataset and code are publicly available.

9 Limitations

Argument components are identified at the sentence level, meaning that claims or premises embedded within longer sentences rather than ex-

pressed as standalone sentences are not captured by the pipeline. The evaluation of instruction-tuned LLMs relies on a sample of 200 test instances per task rather than the full test set, a choice motivated by inference cost, which may introduce sampling variance relative to results obtained on the complete data. Additionally, the instruction-tuned LLMs (Qwen2.5-7B, Mistral-7B, and Nemotron-4B) are evaluated only under prompting-based approaches, with no fine-tuned variants reported. Hence it remains an open question whether parameter-efficient fine-tuning of these models could narrow the performance gap with the classical and fine-tuned transformer baselines. Healthcare LLM applications also require safety guardrails (Arun et al., 2025), which are beyond the scope of this evaluation.

References

- Aetna. 2024. [Medical clinical policy bulletins](#). Accessed: 2024.
- Govind Arun, Sadhvik Bathini, and Balu Bhasuran. 2026. [Temporal epistemic graph of thought: Uncertainty over latent reasoning structure for selective abstention in clinical prediction](#). In *5th Workshop on NLP for Positive Impact*.
- Govind Arun, Rohith Syam, Aiswarya Anil Nair, and Sahaj Vaidya. 2025. [An integrated framework for ethical healthcare chatbots using langchain and nemo guardrails](#). *AI Ethics*, 5(4):3981–3992.
- Balu Bhasuran. 2022. Biobert and similar approaches for relation extraction. In *Biomedical text mining*, pages 221–235. Springer.
- Centers for Medicare & Medicaid Services. 2024. [Medicare coverage database](#). Accessed: 2024.
- Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized language models. In *Advances in Neural Information Processing Systems*, volume 36.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Peng. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Aric Hagberg, Pieter Swart, and Daniel Schult. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pages 11–15.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7B. Technical report, Mistral AI.
- Qiao Jin, Nicholas Wan, Robert Leaman, Shubo Tian, Zhizheng Wang, Yifan Yang, Zifeng Wang, Guangzhi Xiong, Po-Ting Lai, Qingqing Zhu, and 1 others. 2026. Tutorial: guidance on the use of large language models for medical research. *Nature Protocols*, pages 1–12.
- John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.
- Aejin Lee, Rajesh Das, Ashlee Huang, Balu Bhasuran, and Zhe He. 2026. A comparative study of machine learning and language model approaches for classifying mental health disorders. In *2026 IEEE 14th International Conference on Healthcare Informatics (ICHI)*, pages 631–640. IEEE.
- Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):1–25.
- Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, pages 2108–2115. IOS Press.
- Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314. Association for Computational Linguistics.
- NVIDIA. 2024. Nemotron-mini-4b-instruct. Technical report, NVIDIA Corporation.
- Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.
- Prakash Poudyal, Teresa Gonçalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument

- mining. In *Proceedings of the 7th Workshop on Argument Mining (ArgMining 2020)*, pages 67–75. Association for Computational Linguistics.
- Qwen Team. 2024. Qwen2.5: A party of foundation models. Technical report, Alibaba Group.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 3982–3992.
- Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. In *Computational Linguistics*, volume 43, pages 619–659. MIT Press.
- UnitedHealthcare. 2024. [Unitedhealthcare provider policies and protocols](#). Accessed: 2024.
- Kevin Wang, Anna Thöni, Benjamin Kempinski, Bobby Cheng, Jianzhu Yao, Benjamin Finch, Leon Guertler, Viraj Nadkarni, Yihan Jiang, Aliaksei Korshuk, and 1 others. 2026. Mindgames: A live arena for evaluating social and strategic reasoning in multi-agent llms. *arXiv preprint arXiv:2605.29512*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in NeurIPS 2022*.