Evolutionary Self-Supervised Contradiction Detection for Biomedical NLI

Anonymous authors
Paper under double-blind review

Abstract

Identifying conflicting claims in biomedical literature is critical for advancing scientific understanding, yet the scarcity of high-quality training data remains a significant challenge. We introduce EvoNLI, an evolutionary algorithm that learns how to transform entailing sentence pairs into challenging contradictions by mutating words until a frozen teacher model confidently flips its prediction, while preserving topical coherence. EvoNLI, applied to PubMed randomized controlled trials (RCTs), generates SciCon, a dataset of premise–hypothesis pairs whose labels achieve 94.4% agreement across expert judgments in an audit by five domain experts. Fine-tuning large language models on SciCon improves contradiction ROC-AUC consistently across eight biomedical NLI benchmarks. EvoNLI and SciCon are publicly available to support evidence synthesis and robust biomedical natural language inference, and to advance robust domain-specific contradiction detection.[1][2]

1 Introduction

Natural Language Inference (NLI), which determines whether a premise entails, is neutral to, or contradicts a hypothesis, is a long-standing benchmark for measuring whether an NLP system truly understands text. Contradiction detection is especially critical in biomedicine: PubMed alone contains over 39 million citations, with the U.S. National Library of Medicine adding roughly 1.5–1.7 million new records annually (U.S. National Library of Medicine, 2023), far beyond what any human team can continuously reconcile. Moreover, biomedical claims genuinely do reverse. A classic study of highly cited clinical intervention papers found that 16% were later contradicted and another 16% showed initially stronger effects than subsequent evidence supported (Ioannidis, 2005). For U.S. hypertension clinical practice guidelines, 36% of overlapping recommendations were inconsistent in direction (Wright et al., 2011). Together, the exponential scale of the literature and the real frequency of reversals make automated contradiction detection important for evidence synthesis and for flagging conflicting biomedical evidence at scale.

Beyond offline evidence synthesis, contradiction detection is increasingly important for automated reasoning over scientific text: LLM-based assistants that synthesize multiple sources for the same query may present mutually inconsistent claims as jointly true, undermining reliability. Contradiction-aware systems can flag contested findings and support more faithful evidence aggregation (Gokul et al., 2025; Ge et al., 2025).

De Marneffe et al. (2008) distinguish surface contradictions (arising from antonyms, negation, or mismatched numbers) from deep contradictions that rely on modality, presupposition, or world knowledge and remain challenging even for expert annotators. Biomedical literature is rife with deep contradictions: competing randomized controlled trials regularly report opposing outcomes, yet state-of-the-art NLI models require abundant labeled data that is scarce in specialized domains. MedNLI (Romanov & Shivade, 2018), for instance, required four clinicians over six weeks to annotate just 14k sentence pairs. Distant-supervision approaches (Makhervaks et al., 2023) could scale to 1.4M pairs but achieve only 79% precision, introducing noisy or trivial contradictions. In this work, we tackle the open question: how to automatically generate

[1] Code and data: https://anonymous.4open.science/r/183F/

[2] Authors used large language models for writing assistance and limited coding/debugging support. All scientific decisions, experimental design, implementation validation, result interpretation, and conclusions were produced by the authors.

high-precision contradictory pairs without expert annotation while preserving topical coherence and semantic validity?

Standard data augmentation techniques, back-translation, paraphrasing with negation injection, or directly prompting large language models to generate contradictions, fail to guarantee that mutated sentences (i) remain on the same medical topic, (ii) retain a meaningful claim, and (iii) produce contradictions confident enough to train robust classifiers. We introduce EvoNLI, an evolutionary framework that addresses these challenges through guided iterative search. Starting from naturally entailing sentence pairs (e.g., sentences from the same PubMed abstract), EvoNLI applies LLM-guided word substitutions and retains only candidates that (i) flip an LLM classifier's prediction to CONTRADICTION with high confidence, (ii) preserve topical similarity to the original pair, and (iii) maintain claim validity. Over successive generations, the algorithm searches for transformation sequences that reliably produce deep, semantically coherent contradictions, yielding data that is both label-correct and intrinsically challenging.

Applying EvoNLI to PubMed abstracts produces SciCon, a corpus of automatically labeled sentence pairs whose labels achieve 94.4% agreement with five domain experts (96.0% majority-vote precision; Fleiss' $\kappa = 0.842$), substantially exceeding the precision of distant supervision baselines. Fine-tuning six large language models on SciCon yields an average improvement of +12% ROC-AUC over SNOMED-trained models across eight biomedical NLI benchmarks, with gains reaching +31% on challenging datasets like Hard-Cardio. Ablations confirm that evolutionary search, not simply the pool of LLM-generated mutations, drives these improvements, and that topical consistency checks are critical for maintaining data quality.

The contributions of this work are threefold:

- We propose EvoNLI, an evolutionary algorithm that crafts high-quality contradictory sentence pairs through iterative LLM-guided mutation and multi-criteria selection, achieving 94.4% agreement with expert judgments (96.0% majority-vote precision), enabling robust training data for biomedical NLI and evidence synthesis systems.
- We release SciCon, a corpus of scientific contradiction pairs substantially exceeding distant supervision baselines in precision, supporting trustworthy biomedical NLI and evidence aggregation applications.
- We demonstrate that models trained on SciCon improve contradiction detection for evidence aggregation and automated scientific reasoning by an average of +12% ROC AUC over SNOMED across eight biomedical NLI benchmarks, with consistent gains across model architectures. Corpus, code, and trained models are publicly available.[1]

2 Related Work

The contradiction detection task. Recognizing textual entailment (RTE), also known as natural language inference (NLI), is the task of determining whether one text can be logically inferred from another. The first PASCAL RTE challenge was introduced by Dagan et al. (2005), establishing a foundation for computational entailment recognition. The task was later extended to include contradiction detection (Giampiccolo et al., 2007), recognizing that identifying when texts contradict one another is equally important for understanding natural language semantics.

Evolution of general-domain NLI. Large-scale datasets like SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) enabled powerful neural NLI models, but predominantly contain surface contradictions based on simple negation or lexical cues. This limits their utility for specialized domains like biomedicine, where contradictions are subtle and require domain expertise.

Biomedical contradiction detection challenges. In the medical domain, contradiction detection takes on particular significance. The complex and evolving nature of medical research means that contradictory findings across papers are not rare phenomena. Detecting such contradictions is valuable for advancing medical research, as it highlights areas needing further investigation, refinement, or clarification. However,

biomedical contradiction detection presents unique challenges compared to general-domain NLI: it requires deep domain knowledge, subtle reasoning about quantitative findings, and the ability to distinguish between genuine contradictions and differences in experimental design. De Marneffe et al. (2008) distinguish surface contradictions, which arise from explicit cues such as antonymy, negation, or mismatched numbers, from deep contradictions that rely on modality, presupposition, discourse structure, or world knowledge. While surface contradictions are comparatively easy to detect, deep contradictions remain challenging even for human annotators and are particularly prevalent in scientific writing.

Biomedical contradiction resources. The scarcity of expert-annotated training data in biomedicine, where annotation requires weeks per thousands of pairs (Romanov & Shivade, 2018), has motivated research into both manual and automated approaches. Romanov & Shivade (2018) introduced MedNLI, a dataset of 14k clinician-annotated pairs, though hypotheses were clinician-generated rather than authentic clinical text. Alamri & Stevenson (2016) created ManConCorpus from systematic reviews, covering naturally occurring contradictions but limited to narrow topics; their automatically generated AutoConCorpus variant using SemMedDB (Kilicoglu et al., 2012) captures only surface-level predicate clashes with noisy labels. More recently, Makhervaks et al. (2023) introduced SNOMED, which uses distant supervision with the SNOMED CT ontology to automatically extract pairs from PubMed, achieving 79% precision though ontology-based relationships may not capture all forms of scientific contradiction.

Data augmentation and adversarial collection for NLI. Standard augmentation techniques like EDA (Wei & Zou, 2019) or back-translation (Sennrich et al., 2015) are designed to preserve labels through paraphrasing. Our work differs fundamentally: we generate label-flipping transformations that convert entailments into contradictions while maintaining semantic relatedness. Adversarial collection frameworks like ANLI (Nie et al., 2020) and adversarial filtering (Zellers et al., 2018; Sakaguchi et al., 2020) surface challenging examples through iterative human annotation or filtering, but require substantial human-in-the-loop effort and are general-domain. EvoNLI targets fully automatic supervision generation for biomedical contradictions through guided mutation under explicit topicality and claim-validity constraints.

Models for medical contradiction detection. Early contradiction detection systems relied on hand-crafted features and traditional machine learning. Tawfik & Spruit (2018) used a linear support vector machine (SVM) classifier with manually engineered features for detecting contradictions. With the advent of deep learning, Yazi et al. (2021) compared various neural architectures including GloVe embeddings with LSTMs and BERT-based models, demonstrating that BERT substantially outperformed both traditional and earlier neural approaches. Current state-of-the-art systems fine-tune large language models on available biomedical contradiction corpora (Makhervaks et al., 2023; Romanov & Shivade, 2018), suggesting that data quality and scale may be key limiting factors rather than model architecture.

Our contribution. EvoNLI departs from both distant supervision approaches and standard augmentation techniques by using evolutionary search to discover targeted mutations that flip entailment relationships into contradictions. Rather than relying on predefined ontological relationships or label-preserving paraphrases, our method learns to identify which word-level changes create valid contradictions while maintaining topical coherence and claim structure. Applied to PubMed abstracts, this yields SciCon, a corpus whose labels achieve 94.4% agreement with expert judgments (96.0% majority-vote precision) and that substantially improves biomedical and cross-domain contradiction detection when used for fine-tuning state-of-the-art transformers. This represents a scalable, label-efficient alternative to costly expert annotation and noisy distant supervision.

3 Methods

We present a two-stage approach for generating high-quality contradiction pairs. First, we identify candidate sentence pairs that discuss the same medical topic from scientific abstracts. Second, we apply targeted mutations to transform these pairs into contradictions while preserving their semantic focus. The following subsections detail each stage.

3.1 Data Collection

In this work, we focus on sentences from the abstracts of biomedical scientific papers. Our primary data source is the PubMed 200K RCT Dernoncourt & Lee (2017) dataset, which contains abstracts from randomized controlled trial (RCT) articles. Each sentence is labeled with one of five rhetorical roles: BACKGROUND, OBJECTIVE, METHODS, RESULTS, or CONCLUSIONS. We specifically leverage these rhetorical labels to validate the predictions of our foundation models during the sample creation process, ensuring that generated pairs are grounded in sentences with appropriate discourse roles. Full dataset statistics are provided in Appendix A.

When creating a dataset of potentially contradictory sentence pairs, the class of the sentences plays an important role. Sentences that do not contain a clear claim typically cannot serve as contradictions. For example, consider the following sentence (taken from Scholfield et al. (2014)):

> *This study analyzed liver function abnormalities in heart failure patients admitted with severe acute decompensated heart failure (ADHF).*

This sentence is marked as a background sentence in the abstract, which provides context rather than stating a claim. Because it does not include a claim, it is challenging to derive a meaningful contradiction from it. In contrast, sentences labeled as result or conclusion usually contain clear claims. Therefore, our focus is on these types of sentences for generating contradiction pairs.

Another critical factor is topical alignment. For a pair of sentences to be contradictory, they must address the same specific topic. Sentences discussing different topics are, by definition, not contradictory. To address these challenges, we focus on two main types of sentence pairs: conclusion–conclusion and result–conclusion. Crucially, for both pair types, we restrict sentences to the same abstract, substantially increasing the likelihood that they refer to the same underlying medical context, patient population, and intervention, reducing the need for a separate topic classifier.

Conclusion–Conclusion. While the PubMed 200K RCT dataset provides role labels, abstracts often contain multiple sentences labeled as CONCLUSION (averaging 1.78 per abstract).

Result–Conclusion. In medical abstracts, a result sentence typically reports specific quantitative findings, whereas a conclusion sentence summarizes the overall interpretation of those findings. Naturally, these two sentences align, with the result supporting the conclusion.

Since an abstract may contain multiple result sentences (see Appendix A), we must extract the specific one that supports the particular conclusion we intend to use. To do so, we prompt a foundation model to identify the supporting sentence.[3]

For example, consider the following pair from Spector et al. (1996):

- **Result sentence:** "*The incidence of CMV retinitis after 12 months was 24 percent in the placebo group and 12 percent in the ganciclovir group (P<0.0001).*"
- **Conclusion sentence:** "*In persons with advanced AIDS, prophylactic oral ganciclovir significantly reduces the risk of CMV disease.*"

Changing the word "reduces" to "increases" creates a contradiction, while changing "CMV" to another disease name would break topical alignment, the sentences would no longer discuss the same subject.

Crucially, the result sentence typically contains quantitative data (e.g., percentages, p-values) and the conclusion contains a qualitative interpretation; the model cannot rely on simple surface-level pattern matching. Instead, it must infer the logical relationship between the numerical findings and the mutated claim, bridging the semantic gap between evidence and interpretation.

[3]All prompts, hyperparameters, and implementation details are provided in E

Algorithm 1 EvoNLI Generation Algorithm. In our default instantiation, all foundation-model calls (ReplaceWords, Classify, similarity and claim checks) use Llama 3.1 70B; see Section 3.2.1.

function EvoNLI_Generate((P_1, P_2), L_{target})
  $Population \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** N_{pop} **do**
    $P_2' \leftarrow$ SampleMutation(P_1, P_2, L_{target})
    Add (P_1, P_2') to $Population$
  end for
  for $iter \leftarrow 1$ **to** $MAX\_ITERATIONS$ **do**
    $P'_{2,best} \leftarrow$ BestCand$(Population, L_{target})$
    $conf_{P'_{2,best}} \leftarrow$ Classify$(P_1, P'_{2,best}, L_{target})$
    if $conf_{P'_{2,best}} \geq T_{conf}$ **then**
      return $(P_1, P'_{2,best})$
    end if
    $NextPop \leftarrow \emptyset$
    for $(P_1, P_2) \in Population$ **do**
      $P_2' \leftarrow$ SampleMutation(P_2, P_1, L_{target})
      Add (P_1, P_2') to $NextPop$
    end for
    $Population \leftarrow NextPop$
  end for
  return *failure*
end function

function SampleMutation(P_1, P_2, L_{target})
  $Candidates \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** $N_{candidates}$ **do**
    $P_2'' \leftarrow$ ReplaceWords(P_2, P_1, L_{target})
    Add (P_2'') to $Candidates$
  end for
  $P'_{2,best} \leftarrow$ BestCand$(Candidates, L_{target})$
  return $P'_{2,best}$
end function

function BestCand$(Candidates, L_{target})$
  $valid \leftarrow \{P_2'' \in Candidates \mid$ SimScore$(P_2'') \geq T_{sim} \wedge$ ClaimScore$(P_2'') \geq T_{claim}\}$
  return $\arg\max_{P_2'' \in valid}$ Classify(P_1, P_2'', L_{target})
end function

This approach enables us to pair conclusions with corresponding result sentences that share the same topic and contain claims, which can later be mutated to create contradictions. To prevent positional bias, sentence order is randomized in the final dataset.

3.2 Labeling Samples

Starting from entailing sentence pairs, we create contradictions through targeted word mutations. The key challenge is identifying which words to modify: changing the wrong words can break topical coherence, while insufficient changes may fail to create a true contradiction. For instance, in the sentence pair from Spector et al. (1996) shown in Section 3.1, simply replacing "reduces" with "increases" creates a contradiction, but changing "CMV" to another disease would break topical alignment.

To address this, we introduce EvoNLI, an evolutionary algorithm that iteratively mutates the second sentence until a foundation model classifier confidently predicts the target label (contradiction or non-contradiction) while preserving topical similarity and claim validity. Algorithm 1 formalizes this process. The algorithm takes as input an entailing pair (P_1, P_2) and target label L_{target}, outputting a mutated pair (P_1, P_2') with confidence $\geq T_{conf}$ or returning failure after $MAX\_ITERATIONS$. The corresponding generation hyperparameters are summarized in Table 6.

3.2.1 Mutation Pipeline

Each mutation iteration generates $N_{candidates}$ modified versions of the second sentence through foundation model-guided word replacement. In our main instantiation, we use Llama 3.1 70B (Meta, 2024), which receives the sentence pair and target label and suggests specific word substitutions designed to flip the label. However, EvoNLI is not tied to this specific model: any foundation model with comparable reasoning and instruction-following capabilities can serve as the backbone, as supported by the ablation results in Section 6. After experimentation with various approaches, including random replacement and masked language model predictions, we found that explicit prompting of a strong large model yields the most effective and semantically coherent mutations, and we therefore use Llama 3.1 70B as our default backbone.

Each candidate is evaluated on three criteria, all implemented via Llama 3.1 70B (Meta, 2024) prompting in our default setup:

Classification Confidence: In our default setup, the backbone model outputs a probability distribution over contradiction/non-contradiction. We extract the confidence score for L_{target}.

Topical Consistency: While initial pairs come from the same abstract, mutations may inadvertently shift topics (e.g., changing disease names). In our default setup, we verify topical alignment with the original pair using the backbone model as a similarity filter, which returns a similarity score. This serves as a filter that rejects off-topic mutations rather than a retrieval system, mirroring a retrieval-time validation step: we ensure generated contradictions remain within the same query/topic neighborhood. Our ablation study (Section 6) confirms this filter is critical for maintaining data quality.

Claim Validity: Mutations may transform claim-bearing sentences into non-assertions. In our default setup, we verify claim preservation using the backbone model as a claim validator, which returns a claim score. Specifically, the claim score is a probability $\in [0, 1]$ that the sentence expresses a falsifiable scientific assertion about an outcome or intervention (e.g., "Drug X reduces risk of Y") rather than uninformative background or methods text; candidates below T_{claim} are discarded.

The BestCand function selects the candidate with highest classification confidence among those exceeding similarity threshold T_{sim} and claim threshold T_{claim}. If the best candidate's confidence reaches T_{conf}, the algorithm terminates. Otherwise, the population undergoes another mutation iteration, continuing until success or reaching $MAX\_ITERATIONS$. A per-stage accuracy analysis of each filtering component on manually reviewed pairs is provided in Appendix B.

A sensitivity analysis of these filtering thresholds is reported in Appendix C.

4 Empirical Evaluation

In this section, we evaluate our approach empirically, comparing it directly with the results reported by Makhervaks et al. (2023). We use the same benchmark datasets, train/dev/test splits, and ROC-AUC evaluation metric, while departing from their training protocol: we fine-tune solely on the generated corpus (SciCon or SNOMED) without additional in-domain samples, directly isolating the contribution of each data generation method. Additionally, we report the results of a manual evaluation conducted by five domain experts on a sample of our generated data.

4.1 Evaluation Datasets

We use the same evaluation datasets as Makhervaks et al. (2023), which include a mix of naturally occurring and clinician-written sentence pairs across various medical domains.

- **Cardiology Dataset:** This dataset consists of naturally occurring contradictory and non-contradictory sentence pairs in the cardiology domain, derived from ManConCorpus (Alamri & Stevenson, 2016). It serves as a benchmark for evaluating contradiction detection in a specialized medical field.

- **Hard Cardiology Dataset:** A more challenging variant of the Cardiology Dataset, created by removing negation words from the sentences. This modification reduces the reliance on simple negation cues, making it a robust test for model generalization and reasoning.
- **MedNLI Datasets:** MedNLI is a clinical natural language inference dataset tailored for the medical domain Romanov & Shivade (2018). It contains sentence pairs labeled as contradiction, entailment, or neutral. For binary contradiction detection, we follow Makhervaks et al. (2023) and merge the entailment and neutral classes into a single non-contradiction category.

4.2 Baselines

Makhervaks et al. (2023) achieved state-of-the-art performance by fine-tuning large language models on SNOMED, a distant-supervision dataset constructed from PubMed abstracts using the SNOMED CT clinical ontology. Sentence pairs are automatically extracted and weakly labeled using ontological relations, yielding PubMed-scale supervision for biomedical contradiction detection without manual annotation. We use SNOMED as our primary baseline because it is the most directly comparable prior approach in both scale and objective: like EvoNLI, it aims to generate training supervision at scale to address annotation scarcity in biomedical contradiction detection, and it provides an established benchmark suite and experimental protocol. Recent contradiction resources outside SNOMED (e.g., COVID-19 NLI datasets, ContraDoc) are typically narrower in domain coverage and serve primarily as evaluation resources rather than large-scale training supervision. Since our goal is to compare data generation methodologies, holding the downstream fine-tuning and evaluation setup fixed, we follow the experimental setup of Makhervaks et al. (2023) to ensure a direct comparison; alternative approaches such as curriculum learning modify the optimization strategy rather than the supervision-generation pipeline, and are therefore orthogonal to this comparison.

As additional context, we also compare against two prior supervised baselines: Yazi et al. (2021), whose ManConCorpus model concatenates BERT embeddings for the question and claim in a Siamese-like feed-forward network, and Romanov & Shivade (2018), whose best-performing model in the original MedNLI study is based on InferSent, with small additional gains from knowledge-directed attention. Following Makhervaks et al. (2023), we report these baselines fine-tuned on the full in-domain training split without any generated corpus. We report these results in Table 1 as an in-domain supervised reference point.

4.3 Evaluation Setup

All models are evaluated on the same benchmark datasets: Cardiology, Hard Cardiology, MedNLI, and MedNLI subsets (gynecology, endocrinology, obstetrics, and surgery), using identical train/dev/test splits. Each model is fine-tuned solely on the generated corpus (SciCon or SNOMED) and evaluated on the benchmark test sets without any additional in-domain training data. We compare against the strongest reported SNOMED instantiation from Makhervaks et al. (2023). This setup directly compares the two data generation methodologies as the sole source of supervision, isolating the effect of data quality from any in-domain fine-tuning signal. The consistent directionality of improvements across six model architectures and eight benchmarks mitigates seed-sensitivity concerns.

We evaluate a range of large language models (LLMs), including both general-purpose and biomedical-specialized architectures: Llama 3.1 8B and Llama 3.1 70B (Meta, 2024), Qwen2.5-72B (Team, 2024), Phi-4 (Microsoft Research, 2024), Mistral 7B (AI, 2023), and BioMistral 7B (Labrak et al., 2024). Each model is fine-tuned solely on either the SNOMED dataset or our proposed dataset SciCon for supervision. The exact fine-tuning hyperparameters are summarized in Table 7.

In addition to fine-tuned models, we also compare the zero-shot performance of these LLMs. For zero-shot evaluation, models are prompted to classify sentence pairs as contradictory or non-contradictory without additional training. This setup assesses the out-of-the-box clinical reasoning and generalization ability of each model.

Performance is measured using the area under the ROC curve (ROC-AUC), consistent with prior work. This ensures a direct and meaningful comparison between our method and the SNOMED-based distant supervision approach. To verify experimental integrity, we performed n-gram overlap analysis between

training and test sets; minimal 5-gram overlap (<2% for MedNLI, <3% for Cardio) confirms the absence of data contamination, while higher 3-gram overlap (29-57%) reflects shared biomedical terminology inherent to the domain.

5 Empirical Results

5.1 Domain Expert Evaluation

To audit label quality, we randomly sampled 100 SciCon pairs for expert audit. Each pair was independently labeled by five domain experts who were blinded to the algorithmic label. Across expert judgments, 94.4% agreed with the algorithmic label (95% CI: [92.0%, 96.6%]); using majority vote as ground truth, the algorithmic labels reached 96.0% precision. Inter-rater reliability was high (Fleiss' $\kappa = 0.842$; Gwet's $AC_1 = 0.89$), with unanimous consensus on 78% of pairs. The resulting 96.0% majority-vote precision substantially exceeds the 79% precision reported for SNOMED (Makhervaks et al., 2023) and validates our automated labeling methodology.

A detailed taxonomy and distribution of the contradiction types generated by EvoNLI in SciCon, along with a manual audit of mutation quality, are provided in Appendix D.

5.2 Main Result

Table 1: Performance Comparison: SciCon vs Baselines (ROC-AUC scores). Bold indicates best performance per dataset-model pair. * indicates the better method is statistically significant ($p < 0.05$). Yazi et al. 2021 and Romanov & Shivade 2018 columns report results following Makhervaks et al. 2023 without any generated corpus. All other models are fine-tuned solely on SciCon or SNOMED.

				Model Architecture					
Dataset	Yazi et al. 2021	Romanov & Shivade 2018	Method	Llama 3.1 8B	Phi-4	BioMistral 7B	Mistral 7B	Llama 3.1 70B	Qwen2.5 72B
Cardio	0.858	0.824	SciCon	**0.904***	**0.867***	**0.862***	**0.883***	**0.845***	**0.837**
			SNOMED	0.729	0.732	0.543	0.697	0.768	0.832
			Zero-Shot	0.470	0.557	0.473	0.602	0.530	0.586
Hard Cardio	0.687	0.688	SciCon	**0.881***	**0.861***	**0.847***	**0.873***	**0.842***	**0.831**
			SNOMED	0.706	0.691	0.538	0.672	0.746	0.812
			Zero-Shot	0.475	0.614	0.497	0.521	0.574	0.560
MedNLI	0.529	0.643	SciCon	**0.768***	**0.749***	**0.697***	**0.734***	**0.790***	**0.822***
			SNOMED	0.715	0.654	0.582	0.617	0.697	0.736
			Zero-Shot	0.539	0.622	0.487	0.567	0.651	0.670
MedNLI Cardio	0.557	0.738	SciCon	**0.789***	**0.776***	**0.696***	**0.767***	**0.796***	**0.822***
			SNOMED	0.760	0.668	0.614	0.628	0.710	0.742
			Zero-Shot	0.519	0.635	0.506	0.589	0.687	0.693
MedNLI Gynecology	0.508	0.708	SciCon	**0.788***	**0.795***	**0.732***	**0.774***	**0.853***	**0.870***
			SNOMED	0.691	0.657	0.599	0.618	0.707	0.743
			Zero-Shot	0.508	0.591	0.493	0.519	0.742	0.667
MedNLI Endocrinology	0.560	0.707	SciCon	**0.795***	**0.752***	**0.691***	**0.783***	**0.806***	**0.827**
			SNOMED	0.746	0.676	0.603	0.638	0.714	0.800
			Zero-Shot	0.503	0.557	0.521	0.523	0.681	0.576
MedNLI Obstetrics	0.505	0.538	SciCon	**0.769***	**0.742***	**0.708***	**0.719***	**0.782***	**0.800***
			SNOMED	0.673	0.614	0.554	0.604	0.659	0.680
			Zero-Shot	0.529	0.621	0.475	0.548	0.722	0.675
MedNLI Surgery	0.602	0.898	SciCon	**0.755***	**0.739***	**0.686***	**0.775***	**0.804***	**0.834***
			SNOMED	0.717	0.646	0.610	0.637	0.733	0.731
			Zero-Shot	0.537	0.590	0.490	0.557	0.707	0.595

Table 1 compares models fine-tuned on SciCon against two baselines: SNOMED (state-of-the-art distant supervision) and Zero-Shot (foundation models without fine-tuning). The first two columns show architecture baselines: Yazi et al. (2021) and Romanov & Shivade (2018).

The two leftmost columns show Yazi et al. (2021) and Romanov & Shivade (2018) fine-tuned on full in-domain data (following Makhervaks et al. (2023)) as a supervised reference. Despite this advantage, SciCon-tuned

models match or exceed them on seven of eight benchmarks without any task-specific labels. The exception is MedNLI-Surgery, where Romanov & Shivade (2018) reaches 0.898 versus SciCon's best of 0.834, likely reflecting a strong fit between InferSent and this particular test set.

SciCon's key strength is its consistency on the settings where surface cues are unavailable. On Hard-Cardio, where negation words are removed, SciCon outperforms SNOMED on all six models by up to +31 ROC-AUC points. These are precisely the settings that demand semantic generalization rather than negation-spotting. On the standard Cardiology dataset, where negation cues remain present, SciCon leads on all six models; fine-tuning on SciCon provides gains of up to 41 percentage points over zero-shot baselines on Hard-Cardio. Across the MedNLI subsets, SciCon leads on all six model architectures, with statistically significant advantages on the majority of model-dataset pairs.

SciCon outperforms SNOMED across all six models on all eight benchmarks, with no architecture-level exception. The advantage is statistically significant on 45 of 48 model-dataset pairs. The three non-significant cases are all for Qwen2.5 72B: Cardio (0.837 vs. 0.832), Hard-Cardio (0.831 vs. 0.812), and MedNLI-Endocrinology (0.827 vs. 0.800), where SNOMED's broader training scale yields competitive but ultimately lower scores on this specific model. All other 45 comparisons favour SciCon significantly, confirming that the quality and semantic precision of the generated supervision signal drives consistent gains across cardiology and all five MedNLI clinical sub-domains.

We validated results using bootstrap significance testing (1,000 samples), which provides reliable performance variability estimates without costly repeated fine-tuning. Asterisks (*) mark cases where the better-performing method is statistically significant ($p < 0.05$).

6 Ablation Studies

To better understand the contribution of different components and settings of our method, we report several types of ablation studies in this section. These studies systematically evaluate the impact of individual elements, such as the word selection strategy, the necessity of similarity and claim detection, and the choice of classification model. Additionally, we compare EvoNLI against traditional data augmentation baselines adapted for contradiction generation.

Table 2: Ablation Studies and Foundation Model Variant Baselines: Evaluating the Contribution of Similarity Checking, Claim Detection, Classification Methods, Word Selection Strategies, Data Augmentation Baselines, and Alternative Backbone Models (ROC-AUC scores)

		Cardio Dataset			MedNLI-General Dataset		
Ablation	Method	Qwen2.5 72B	BioMistral 7B	Llama 3.1 8B	Qwen2.5 72B	BioMistral 7B	Llama 3.1 8B
Quality Checks	Ours	**0.837**	**0.862**	**0.904**	**0.822**	**0.697**	**0.768**
	No Claim Checking	0.822	**0.862**	0.902	0.799	0.652	0.766
	No Similarity Checking	0.642	0.497	0.567	0.557	0.488	0.474
Classification Method	Ours	**0.837**	**0.862**	**0.904**	**0.822**	**0.697**	**0.768**
	Cross Encoder	0.827	0.687	0.739	0.742	0.526	0.767
Word Selection Strategy	Ours	**0.837**	**0.862**	**0.904**	**0.822**	**0.697**	**0.768**
	Random Replacement	0.67	0.807	0.874	0.500	0.504	0.573
	MLM Replacements	0.606	0.840	0.870	0.635	0.506	0.643
Data Augmentation	Ours (EvoNLI)	**0.837**	**0.862**	**0.904**	**0.822**	**0.697**	**0.768**
	Modified EDA	0.574	0.524	0.509	0.485	0.502	0.496
	Modified Back-Translation	0.583	0.547	0.473	0.574	0.540	0.501
Foundation Model Variants	Ours (EvoNLI, Llama 3.1 70B)	0.837	**0.862**	**0.904**	**0.822**	**0.697**	0.768
	EvoNLI (Qwen2.5 72B backbone)	**0.859**	0.802	0.889	0.801	0.528	**0.820**
	EvoNLI (Mistral 7B backbone)	0.540	0.477	0.569	0.507	0.484	0.470
	EvoNLI (Mistral 7B gen+filter, Qwen2.5 7B cls)	0.724	0.660	0.803	0.570	0.522	0.514
	Direct LLM Prompt (Llama 3.1 70B, no algo)	0.495	0.500	0.546	0.577	0.533	0.493

6.1 Necessity of Similarity and Claim Detection

Table 2 shows that removing the similarity check causes severe performance degradation across all models, confirming that topical consistency is essential, as off-topic mutations produce meaningless contradictions.

Removing the claim check has a smaller effect, likely because mutations rarely eliminate claims entirely and effective prompting reduces non-claim generation. This suggests topical drift is more common than claim elimination during mutation.

6.2 Classification Model Options

This study evaluates the impact of using a cross-encoder for classification instead of the foundation model, we generate data using the cross-encoder that has been fine-tuned on SNOMED and fine-tune models to observe the effects on performance. The results are presented in Table 2. It can be seen that using a cross-encoder for classification attains good performance across models and datasets. However, it generally reaches lower performance than achieved with our original method. This indicates that while the cross-encoder can be effective for classification, it may not be as powerful as the foundation model in generating high-quality data. This offers a trade-off between performance and computational efficiency, as cross-encoders are typically faster and require less computational resources than foundation models.

6.3 Word Selection Strategy

We compare our LLM-guided word selection against two alternatives: (1) random word replacement and (2) masked language model (MLM) predictions. Table 2 shows our approach outperforms both. The reason for this is that the original strategy is more targeted and focused on specific words that are likely to contribute to the contradiction, while random selection may introduce noise and irrelevant changes. The MLM selection also shows a drop in performance, and that might be because an MLM would suggest similar words to the original word, which will also make it hard to create a contradiction, as can be seen in Table 3.

6.4 Comparison with Data Augmentation Baselines

We compare EvoNLI against traditional data augmentation techniques: EDA (Wei & Zou, 2019) and Back-Translation (Sennrich et al., 2016). Since these methods are typically designed to preserve the original label (entailment), we adapted them to generate contradictions, modifying EDA to prioritize antonym substitution and negation insertion, and introducing negation markers into Back-Translation's intermediate representation. Table 2 shows that EvoNLI consistently outperforms these approaches, demonstrating that generating valid scientific contradictions requires the semantic guidance provided by our evolutionary framework rather than surface-level transformations.

6.5 Foundation Model Variant Baselines

To assess whether EvoNLI's gains are specific to the Llama 3.1 70B backbone, we evaluate four additional data generation variants under the same downstream fine-tuning and evaluation protocol:

- **EvoNLI (Qwen2.5 72B backbone)**: all pipeline stages (mutation, classification, similarity, claim checking) use Qwen2.5 72B instead of Llama 3.1 70B.
- **EvoNLI (Mistral 7B backbone)**: a computationally lighter variant using Mistral 7B for all stages, testing the cost–performance trade-off.
- **EvoNLI (Mistral 7B gen+filter, Qwen2.5 7B cls)**: a decoupled pipeline where Mistral 7B handles mutation generation, similarity filtering, and claim validation, while Qwen2.5 7B performs classification, directly testing whether separating generation from classification improves data quality and mitigates single-model bias.
- **Direct LLM Prompt (Llama 3.1 70B, no algorithm)**: contradictions generated via a single prompt with no iterative evolutionary search, similarity filtering, or claim validation, serving as an empirical baseline for direct prompting without algorithmic scaffolding.

All four variants are evaluated on Cardiology and MedNLI-General with the same three downstream models as the other ablations (Table 2). Results reveal three consistent patterns.

Backbone scale dominates. The Mistral 7B-only backbone performs near chance across all cells (range 0.477–0.569 on Cardio; 0.470–0.507 across all three downstream models on MedNLI), confirming that a 7B-parameter model generally lacks the capacity to generate coherent, semantically valid biomedical contradictions without stronger scaffolding. The decoupled pipeline (Mistral 7B gen+filter, Qwen2.5 7B cls) partially recovers, most notably yielding 0.803 ROC-AUC for Llama 3.1 8B fine-tuned on Cardio, suggesting that classification signal quality matters more than generation capacity when data volume is held constant.

72B-scale backbones generalise across model families. EvoNLI with the Qwen2.5 72B backbone performs competitively on MedNLI across downstream models (0.801 vs. 0.822 for Qwen2.5 72B fine-tuning; 0.820 vs. 0.768 for Llama 3.1 8B), indicating that EvoNLI's gains are not artefacts of a particular model family and that any capable 70B+ backbone produces high-quality supervision for general NLI.

The evolutionary algorithm matters most for domain-specific data. Comparing Ours (0.837 on Cardio) against the Direct LLM Prompt variant (0.495/0.500/0.546 on Cardio), the algorithm contributes roughly 0.34 ROC-AUC on Cardio for the Qwen2.5 72B downstream model, a substantially larger margin than on MedNLI (0.822 vs. 0.577 for Qwen2.5 72B). Direct prompting surfaces common negation-style contradictions that transfer well to MedNLI's clinician-written pairs, but the iterative search and topicality constraints are essential for generating the precise, domain-grounded contradictions needed to improve cardiology-specific detection.

Table 3: Comparison of contradiction ratios for datasets generated using EvoNLI and its ablations. The 'Ratio' indicates the proportion of generated pairs successfully labeled as contradictions.

Dataset	Ratio
SciCon (Ours)	0.482
Random Replacement	0.132
No Similarity Check	0.500
No Claim Check	0.448
MLM Replacement	0.171
Cross-Encoder	0.271

7 Qualitative Application on Real Abstracts

To bridge the gap between our sentence-level evaluation and a more realistic clinical application, we conducted an extrinsic experiment. This preliminary study assesses our method's ability to detect contradictions between entire abstracts, moving beyond isolated sentences. The results of this abstract-level analysis were evaluated manually by a domain expert.

7.1 Similar Articles

To find similar articles, we use the PubMed Related Articles metric Lin & Wilbur (2007). After manually examining numerous abstract pairs and tuning threshold values across multiple similarity features, we found that shared MeSH terms, Jaccard similarity between titles and abstracts, and the number of shared chemicals provide the most reliable indicators for determining whether two abstracts discuss the same topic.

7.2 Finding Contradictions

We sampled 1667 abstracts and used a Mistral 7B model fine-tuned on our SciCon dataset to flag potentially contradictory pairs. The model identified 24 pairs, which were then manually reviewed by a domain expert. Of these, 19 were confirmed as true contradictions. The remaining 5 pairs were false positives, primarily due to nuanced differences in study design that our sentence-level model was not trained to capture. Common reasons for misclassification included different drug interventions or dosages, distinct patient populations, or disparate clinical endpoints. These results highlight the challenge of abstract-level contradiction detection

and suggest that future work should incorporate a more detailed analysis of experimental design to improve robustness.

8 Conclusions

We introduced EvoNLI, an evolutionary algorithm that generates high-quality contradictory sentence pairs by iteratively mutating entailing biomedical sentences while preserving topical coherence and claim validity. The resulting SciCon dataset achieves substantially higher precision than distant supervision baselines and consistently improves contradiction detection across diverse biomedical benchmarks and model architectures.

Beyond offline evidence synthesis, this work addresses a critical challenge in modern biomedical information access. RAG pipelines and LLM-based systems are increasingly sensitive to conflicting context (Wang et al., 2025; Niu et al., 2024; Zhang et al., 2026): presenting contradictory evidence without detection can degrade answer quality, amplify hallucinations, and undermine user trust. The high-precision training data provided by SciCon enables deployment of contradiction detection models in practical retrieval pipelines. A typical integration would involve: (1) retrieving candidate abstracts or snippets for a user query via standard retrieval methods, (2) applying pairwise contradiction scoring between retrieved snippets or between snippets and a generated summary, and (3) clustering results by stance, re-ranking toward internally consistent result sets, or explicitly flagging conflicting claims in the search interface. This enables transparent presentation of contested findings rather than treating inconsistent evidence as uniformly authoritative, improving trust and transparency in biomedical information access.

While our approach currently operates at the sentence-pair level, the EvoNLI framework is domain-agnostic. Future work could explore more sophisticated mutation operators that capture deeper semantic contradictions, extend to document-level or multi-hop reasoning, investigate whether similar evolutionary approaches benefit other NLI tasks, and apply the methodology to other specialized domains where expert annotation is costly.

9 Ethics and Broader Impact

Our pipeline uses publicly available scientific corpora and PubMed abstracts; to the best of our knowledge, it does not include personally identifiable patient information. However, publication-driven data may over-represent common diseases and outcomes while under-representing rare conditions.

The main risk is contradiction-detection error in biomedical search/RAG settings: false positives may overstate disagreement, and false negatives may hide real conflict. Therefore, outputs should be used only as decision-support signals for literature triage, not as autonomous clinical recommendations. We mitigate this risk by keeping humans in the loop, exposing evidence and uncertainty scores, monitoring subgroup performance, and periodically auditing generated labels and model predictions.

10 Limitations

Our methodology relies on LLMs for mutation and verification, which may introduce noise through hallucinations or errors (Zhang et al., 2023), though expert audit (94.4% agreement; 96.0% majority-vote precision) suggests such cases are relatively rare. This issue is especially salient for LLM-mediated search and summarization, where noisy or auto-generated snippets can introduce conflicts that must be detected or surfaced.

EvoNLI operates at the sentence level and does not explicitly model three factors that are central to biomedical contradiction detection: (i) population mismatches, where sentences use opposite outcome language but refer to different patient subgroups; (ii)study design differences, where divergent interventions, dosages, or endpoints produce superficially contradictory sentences that are not logically incompatible, as also reflected by false positives in the abstract-level experiment (Section 7); and (iii)epistemic hedging, where modal verbs obscure genuine opposition, contributing to the classifier's 12% error rate (Section B). Addressing these would require extending the filtering pipeline with participant-level metadata matching and assertion-strength scoring, which we leave to future work.

References

Mistral AI. Mistral 7b, 2023. URL https://mistral.ai/news/announcing-mistral-7b/.

Abdulaziz Alamri and Mark Stevenson. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics*, 7(1):1–9, 2016.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.

Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pp. 1039–1047, 2008.

Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*, 2017.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented llms. *arXiv preprint arXiv:2505.17762*, 2025.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pp. 1–9, 2007.

Vignesh Gokul, Srikanth Tenneti, and Alwarappan Nakkiran. Contradiction detection in rag systems: Evaluating llms as context validators for improved information consistency. *arXiv preprint arXiv:2504.00180*, 2025.

John PA Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *Jama*, 294 (2):218–228, 2005.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch. Semmeddb: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, 2012.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

Jimmy Lin and W John Wilbur. Pubmed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):1–14, 2007.

Dave Makhervaks, Plia Gillis, and Kira Radinsky. Clinical contradiction detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1248–1263, 2023.

Meta. Llama 3.1 model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct.

Microsoft Research. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of ACL*, 2020.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Aflite: Adversarial filtering for natural language understanding. In *Proceedings of ACL*, 2020.

Michael Scholfield, Matthew B Schabath, and Maya Guglin. Longitudinal trends, hemodynamic profiles, and prognostic value of abnormal liver function tests in patients with acute decompensated heart failure: an analysis of the escape trial. *Journal of cardiac failure*, 20(7):476–484, 2014.

R Sennrich, B Haddow, and A Birch. Improving neural machine translation models with monolingual data. corr abs/1511.06709 (2015). *arXiv preprint arxiv:1511.06709*, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 86–96, 2016.

Stephen A Spector, George F McKinley, Jacob P Lalezari, Tobias Samo, Robert Andruczk, Stephen Follansbee, Paula D Sparti, Diane V Havlir, Gail Simpson, William Buhles, et al. Oral ganciclovir for the prevention of cytomegalovirus disease in persons with aids. *New England Journal of Medicine*, 334(23): 1491–1497, 1996.

Noha S Tawfik and Marco R Spruit. Automated contradiction detection in biomedical literature. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 138–148. Springer, 2018.

Qwen Team. Qwen2.5: A party of foundation models, 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

U.S. National Library of Medicine. Pubmed overview. https://pubmed.ncbi.nlm.nih.gov/about/, 2023. Accessed 2024.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*, 2025.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

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

James M. Wright, Krista Bassett, Vijaya M. Musini, Andrew Wareham, and Finlay A. McAlister. Conflicting clinical practice guidelines for hypertension. *Annals of Internal Medicine*, 156(5):355–363, 2011. doi: 10.7326/0003-4819-156-5-201203060-00007.

Fatin Syafiqah Yazi, Wan-Tze Vong, Valliappan Raman, Patrick Hang Hui Then, and Mukulraj J Lunia. Towards automated detection of contradictory research claims in medical literature using deep learning approach. In *2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP)*, pp. 116–121. IEEE, 2021.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of EMNLP*, 2018.

Boya Zhang, Alban Bornet, Rui Yang, Nan Liu, and Douglas Teodoro. Healthcontradict: Evaluating biomedical knowledge conflicts in language models. *npj Digital Medicine*, 9(152), 2026. doi: 10.1038/s41746-025-02336-0. URL https://doi.org/10.1038/s41746-025-02336-0.

Yue Zhang, Yafu Li, Jaebong David Choi, and Sang-goo Lee. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Table 4: PubMed 200K RCT dataset statistics and average rhetorical role distribution per abstract.

	Train	**Dev**	**Test**
Abstracts	205,654	5,000	5,000
Sentences	2,391,901	59,144	59,628
Avg. sent./abstract	11.63	11.83	11.93
Avg. sentences per role:			
RESULTS	4.01	3.96	4.00
METHODS	3.80	3.90	3.93
CONCLUSIONS	1.78	1.80	1.80
BACKGROUND	1.06	1.20	1.26
OBJECTIVE	0.97	0.96	0.94

A PubMed 200K RCT Dataset Statistics

B Qualitative Analysis of Prompt Performance

To validate the reliability of our pipeline components, we manually analyzed randomly sampled pairs at each stage.

Word Replacement achieved 58% success by cleanly inverting key claims (e.g., "favorable" → "poor", "decrease" → "increase", "reduces" → "increases"). The 42% failures included: no replacement (24%, sentences remained unchanged), wrong target (12%, altering disease stage or population rather than inverting the claim), incomplete negation (e.g., one case modified a sentence to describe "unbearable toxicity" and "less survival time" while still characterizing the therapy as a "reasonable palliative option"), non-contradictory modifications despite word changes:

> *Original:* "Future interventions that **reduce** maternal viremia and **minimize** pediatric HIV infection will have a **positive** impact on infant health outcomes."
>
> *Mutated:* "Future interventions that **increase** maternal viremia and **maximize** pediatric HIV infection will have a **negative** impact on infant health outcomes."

stating that increases in viremia lead to negative outcomes, which actually aligns with rather than contradicts the premise. Finally, unnatural wording introduced double negatives like "not offering an un-promising alternative." This 58% immediate success confirms the necessity of iterative refinement, as failed mutations are discarded.

Classification demonstrated 88% accuracy. The 12% errors revealed: (1) Indirect inversions (7%), where the model failed to recognize opposition through paraphrases rather than explicit negation. For example:

> *Sentence 1:* "Oral ganciclovir significantly **reduces** the risk of CMV disease."
>
> *Sentence 2:* "Prophylactic oral ganciclovir may **not be effective** in reducing CMV disease risk."

The classifier failed to recognize that "not be effective" contradicts "reduces." Similar patterns included "attenuated correlation" versus "may strengthen the link," and directional opposites like "increase" versus "decrease." (2) Hedged contradictions (4%), where modal verbs ("may", "might", "could") obscured opposition when one sentence made a definitive claim and another hedged with epistemic modality. False positives arose from scope mismatches, comparing warfarin tolerance in "elderly patients with atrial fibrillation" versus "young patients with normal heart rhythm" (different populations that cannot contradict each other).

Claim Detection achieved perfect 100% accuracy, with all sampled mutations retaining valid declarative statements about outcomes, relationships, or interventions. This filter reliably prevents non-assertive or irrelevant mutations.

Similarity Detection achieved 96% accuracy in maintaining topical alignment. The single false negative (1%) involved synonym recognition failure when outcomes were negatively phrased. False positives (3%) stemmed from subtle aspect shifts, comparing different independent variables (stent types vs. patient age, temporary vs. permanent anesthesia, vitamin D vs. calcium) within the same broad medical domain.

Implications for Data Quality These multi-stage filters create a safety net where no single component must be perfect. The classification model's 88% accuracy represents the critical bottleneck for final data quality, while near-perfect claim detection (100%) and high similarity filtering (96%) prevent non-assertive sentences and topic drift. The evolutionary algorithm's iterative refinement recovers from individual prompt failures, ultimately producing the 96.0% majority-vote precision and 94.4% agreement across expert judgments reported in Section 5.1. Three recurring failure patterns identified here motivate the limitations discussed in Section 10.

C Threshold Sensitivity Analysis

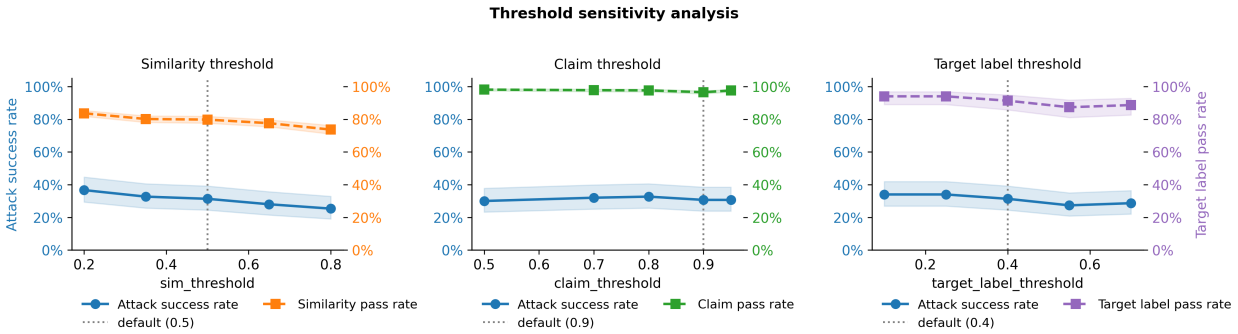

Figure 1: Threshold sensitivity sweep over n=150 result–conclusion and conclusion–conclusion sentence pairs drawn from the same abstract. Each panel varies one threshold across five values while holding the other two at their defaults (`sim`= 0.5, `claim`= 0.9, `target`= 0.4). Left axis (blue, solid): attack success rate with 95% Wilson CI. Right axis (dashed): relevant pass rate for that threshold. Vertical dotted line marks the production default in each panel.

The EvoNLI algorithm is governed by three filtering thresholds whose defaults were initially set by intuition. To ground these choices empirically, we perform a one-at-a-time sensitivity sweep, varying each threshold across five values while holding the other two fixed at their defaults (`sim`= 0.5, `claim`= 0.9, `target`= 0.4). Figure 1 reports attack success rate (ASR) and the pass rate of the filter controlled by each threshold.

Similarity threshold. The similarity threshold has the strongest influence on both metrics. Tightening it from 0.2 to 0.8 reduces ASR monotonically from 36.7% to 25.3%, with the similarity pass rate falling in parallel from 83.6% to 73.6%: stricter topic-coherence screening rejects a larger fraction of candidate mutations and leaves the genetic search with fewer viable paths. The production default of 0.5 sits near the inflection of the curve, yielding 31.3% ASR and a $\approx$80% pass rate, a reasonable balance between coverage and topical fidelity.

Claim threshold. The claim threshold is effectively inert. ASR varies by fewer than three percentage points across the full range tested (30.0–32.7%), and the claim pass rate remains above 96% at every operating point. This indicates that the biomedical claim detector rarely rejects a candidate that the similarity filter has already approved, so tightening or relaxing this threshold has negligible downstream impact. The default of 0.9 is therefore retained for its conservative semantics without sacrificing attack coverage.

Target label threshold. The target label threshold mediates a mild coverage–quality trade-off. At lenient values (0.1–0.25), the genetic search accepts candidates with weaker contradiction scores, achieving the highest ASR (34.0%) and a target label pass rate of 94%. As the threshold rises toward 0.55, ASR falls to

27.3% and the pass rate drops to 87%, reflecting the increased difficulty of finding mutations that clear a stricter contradiction criterion within the allowed iterations. The default of 0.4 achieves 31.3% ASR and 91% pass rate, providing an adequate contradiction signal while remaining well within the flat region of the curve before the steeper drop-off.

Taken together, the sweep confirms that the production defaults represent principled operating points: the similarity threshold is the only lever with substantial leverage over ASR, the claim threshold imposes quality constraints at negligible cost, and the target label threshold can be tuned modestly to trade attack breadth against contradiction strength.

D Analysis of Mutation Types

To better understand the nature of the contradictions generated by EvoNLI, we analyzed the SciCon dataset by categorizing the specific types of mutations introduced during the evolutionary process. We developed a hierarchical taxonomy of mutation types and employed a foundation model to classify each generated pair based on the modification made to the hypothesis.

The mutation taxonomy is organized into four high-level categories:

- **Logical:** Contradictions caused by changes in logical operators or truth conditions. Logical operators take priority when present.
  - *Negation/Polarity*: Introduction or removal of negation ("is" → "is not", "exists" → "does not exist")
  - *Quantifier*: Changes in quantity expressions ("all" → "some", "many" → "few")
  - *Numerical*: Changes in explicit numbers or ordinals ("three" → "five", "first" → "second")
  - *Modality*: Changes in certainty, possibility, or obligation ("must" → "might", "will" → "could")
  - *Existence*: Assertion vs denial of existence ("there is" → "there is not")
  - *Conditional*: Changes in conditional structure ("if" → "unless", adding or removing conditions)
- **Temporal/Spatial:** Contradictions involving time or location.
  - *Temporal*: Changes in time reference or ordering ("yesterday" → "tomorrow", "past" → "future")
  - *Spatial*: Changes in location or spatial relations ("here" → "there", "inside" → "outside")
- **Event-based:** Contradictions involving actions or events.
  - *Action*: Changes in the action or process performed ("run" → "walk", "buy" → "sell")
  - *Causation*: Changes in cause-effect relations ("cause" → "prevent", "lead to" → "stop")
- **Property-based:** Contradictions involving attributes, states, or values of entities when not better explained by the above categories.
  - *Scalar Property*: Opposing values on an ordered or gradable scale ("large" → "small", "hot" → "cold")
  - *Categorical Property*: Mutually exclusive categories with no inherent ordering ("red" → "blue", "male" → "female")
  - *Relational Property*: Changes to relations between entities (e.g., switching the direction or type of association)
  - *State/Status*: Binary or reversible states ("open" → "closed", "alive" → "dead")
  - *Evaluative Property*: Value judgments or normative assessments ("good" → "bad", "safe" → "dangerous")

D.1 Distribution of Mutation Types

Table 5 shows the distribution of mutation types in SciCon. *Logical* edits account for 50% of mutations (dominated by *Negation/Polarity* at 42%), followed by *Causation* (30%) and *Evaluative Property* (8%). This indicates EvoNLI generates contradictions beyond simple negation, frequently requiring causal and property-based flips that better match scientific claim structure.

Table 5: Distribution of mutation types in the SciCon dataset.

Mutation Type	**Percentage**
Logical	
Negation/Polarity	42%
Modality	4%
Quantifier	2%
Conditional	2%
Event-based	
Causation	30%
Action	8%
Property-based	
Evaluative Property	8%
Scalar Property	2%
Relational Property	1%
Temporal/Spatial	
Spatial	1%

D.2 Manual Validation of Mutation Quality

To validate both the quality of generated contradictions and the accuracy of our mutation taxonomy, we conducted a detailed manual audit of randomly sampled sentence pairs from the SciCon dataset. Each pair was reviewed to verify whether the modified sentence truly contradicts the original.

The manual audit found 97% agreement. The disagreements stemmed from (i) labeling "no significant difference" statements as contradictory, (ii) topic drift from an incorrect entity replacement (anastomotic leakage → acute leukemia), and (iii) an Action edit that did not change the underlying claim. Overall, this supports that EvoNLI produces high-quality contradictions across diverse mutation types.

E Implementation Details and Hyperparameters

E.1 EvoNLI Generation Hyperparameters

Table 6 reports the hyperparameters used for all SciCon generation experiments.

Table 6: EvoNLI generation hyperparameters.

Hyperparameter	Value
Backbone model	Llama 3.1 70B Instruct
Population size N_{pop}	4
Candidates per mutation $N_{\text{candidates}}$	5
Maximum iterations	4
Target confidence threshold T_{conf}	0.4
Similarity threshold T_{sim}	0.5
Claim threshold T_{claim}	0.9

Table 7: Fine-tuning hyperparameters for SciCon and SNOMED-trained models.

Hyperparameter	Value
Training dataset	SciCon or SNOMED
Evaluation datasets	Cardio, Hard-Cardio, MedNLI subsets
Optimizer	AdamW (8-bit)
Learning rate	2e-5
Per-device batch size	4
Gradient accum. steps	8
Effective batch size	32
Epochs	15
LR scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
LoRA rank r	16
LoRA alpha	32
LoRA dropout	0.05
LoRA target modules	All attn. + MLP proj.

E.2 Fine-tuning Hyperparameters

Unless otherwise stated, all reported SciCon results use the hyperparameters in Table 6. The threshold values were selected before downstream evaluation and are analyzed in the sensitivity study in Appendix C. The same settings were used for SciCon and SNOMED to isolate the effect of the supervision source.

E.3 Prompts Used in EvoNLI

In this section, we provide the prompts used for different steps in our EvoNLI method. These prompts were designed to guide the foundation models in generating and classifying sentences effectively. The prompts were iteratively refined based on initial experiments to ensure effectiveness.

E.3.1 Conclusion Generation Prompt

```
Write a concise, single-sentence conclusion for the following
sentences, summarizing the overall findings or implications of the
research without providing an explanation or elaboration.
Sentences: {abstract_without_conclusion}
Answer:
```

E.3.2 Result Extraction Prompt

```
From the given conclusion, extract a single supporting
sentence from the abstract, which is not the provided
conclusion, without providing any explanation or additional information:
Abstract: {abstract}
Conclusion: {conclusion}
Answer:
```

E.3.3 Word Replacement Prompt

```
Our goal is to get {target_label} sentences.
Choose words from sentence2 and replace them in the sentence.
Write only sentence2 as an answer.
Provide no additional explanation or commentary.
sentence1: {sentence1}
```

```
sentence2: {sentence2}
Answer:
```

E.3.4 Samples Classification Prompt

```
You are a helpful assistant expert in Biomedical domain.
We want to detect contradictions between sentences.
Determine if the following two sentences are contradictory.
Answer with only yes or no.
Provide no additional explanation
or commentary.
sentence1:{sentence1}
sentence2: {sentence2}
Answer:
```

E.3.5 Similarity Detection Prompt

```
You are a helpful assistant expert in Biomedical domain.
We want to check if sentences discuss the same topic even if
they are contradicting. Analyze the following two
sentences and determine if they discuss the same exact topic.
If yes, answer 'yes.'
If no, answer 'no.'
Provide no other explanation.
sentence1: {sentences[0]}
sentence2: {sentences[1]}
Answer:
```

E.3.6 Claim Detection Prompt

```
You are a system that evaluates sentences. When given a sentence,
you must determine if it contains a claim or provides any new
information. If it does, reply with 'yes' and nothing else.
If it does not, reply with 'no' and nothing else.
Do not provide explanations.
Determine whether the following sentence contains a claim, finding,
hypothesis. Respond with 'Yes' if it does, otherwise respond with 'No'.
Do not provide any explanation, only 'Yes' or 'No'.
Here is the sentence:
sentence: {sentence}
Answer:
```