



Article

Compact Models for Structured Argument and Stance Analysis: A Framing-Aware, Retrieval-Augmented Pipeline

Antonios Charalampous  and Constantinos Djouvas * 

Department of Communication and Internet Studies, Faculty of Communication and Media Studies,
Cyprus University of Technology, Limassol 3036, Cyprus; ag.charalampous@edu.cut.ac.cy

* Correspondence: constantinos.djouvas@cut.ac.cy

Abstract

We propose a modular, retrieval-augmented pipeline for computational argumentation that integrates two complementary components: ARGSTANCE, a multi-task model for argument and stance reasoning, and TARGETMATCH, a contrastive retrieval model that treats target identification as a first-class retrieval task. We further formulate stance detection as a framing-aware problem, recognizing that the polarity of a stance toward a target depends on how the proposition is framed. To support broad generalization, we construct a large dataset spanning Kialo discussions, Wikipedia, and curated news articles, and introduce a cross-source injection strategy that mitigates domain and style biases. Our compact models achieve F1 scores of 0.94 for argument detection (ModernBERT-base) and 0.84 for same-side stance detection (ModernBERT-large), while TARGETMATCH attains a top-10 retrieval accuracy of 0.75. Under controlled zero-shot comparisons with large language models, our models remain competitive while offering advantages in reproducibility, deployment cost, and controllable intermediate predictions.

Keywords: computational argumentation; argument mining; stance detection; target identification; contrastive learning

1. Introduction

Argumentation is a fundamental aspect of human communication, employed to defend positions, persuade others, and facilitate reasoning in domains ranging from politics and law to education and science. Computational methods for analyzing argumentative discourse have given rise to the field of Computational Argumentation [1–3]. This interdisciplinary area combines natural language processing (NLP) and artificial intelligence to address various tasks such as argument mining [4], stance detection [5], argument quality [6], summarization [7], generation [8], clustering [9] and many more.

Argument-detection systems have the potential to enable large-scale qualitative analysis of online discussions, providing valuable input for policymakers, researchers, and businesses [10,11]. On the other hand, stance detection specifically can help track social attitudes toward contentious issues like climate change [12], COVID-19 vaccination [13] and abortion [14].

Despite significant advances, several challenges remain. First, most existing approaches to argument detection treat it as part of a broader argument mining pipeline, where the goal is not only to decide whether text is argumentative, but also to identify the precise boundaries of argumentative components and classify them into specific component types such as claims or premises [15–20]; while valuable, this formulation comes at



Academic Editors: Shaheen Khatoon,
Azhar Mahmood and Marek Sikora

Received: 8 June 2026

Revised: 9 July 2026

Accepted: 9 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

a cost: it relies on small, domain-specific and labor-intensive annotated datasets [21–25]. Given the inherent complexity and subjectivity of argumentation [26], models trained on such data lack generalizability and their performance falls short of what is needed in realistic, large-scale applications [4,27]. To address this, we frame argument detection as a sequence classification task and introduce a dataset construction process that blends distant supervision [28] with cross-source injection, yielding models with improved reliability and cross-domain generalizability.

Second, most existing approaches to stance detection train models with respect to predefined targets such as abortion, marijuana legalization, or climate change, exploring settings like in-target, cross-target, or zero-shot transfer [29]. However, these approaches neglect the fact that the meaning of a stance is inherently dependent on the way the debate itself is framed. For example, in a debate framed as “*Abortion should be illegal*”, a *pro* stance indicates opposition to abortion, whereas in a debate framed as “*Abortion is a human right*”, a *pro* stance signals support for abortion. Treating “abortion” as a single target without accounting for its framing leads to brittle models that fail to generalize across real-world debates. To overcome this limitation, we propose a more flexible stance detection pipeline that adapts stance interpretation to the specific proposition under discussion, thereby capturing the debate-dependent nature of stance.

Lastly, approaches that jointly address argument detection and stance detection remain limited [3,30,31]. Yet, integrating these tasks is crucial for advancing the field, as it enables a more complete understanding of opinions expressed around controversial topics. Our work addresses this by designing a transparent, modular pipeline that combines these tasks, supported by open resources for the community.

The contributions of this paper are organized around problem formulation, data, methodology, and resources, rather than around novel algorithmic primitives. Specifically, we make the following contributions:

1. **A framing-aware formulation of stance detection:** We articulate stance detection as a task that is not only target-dependent but also *framing-dependent*: the same target may carry different polarities depending on how it is propositionally framed (e.g., “*abortion should be illegal*” vs. “*abortion is a human right*”). This framing motivates treating target identification as a retrieval problem rather than a closed classification problem.
2. **A large, cross-source dataset for argument and stance detection:** Building on distant supervision over Kialo discussions, we construct a dataset that incorporates Wikipedia and news articles via a cross-source injection strategy designed to dilute source-specific stylistic cues and improve generalization across linguistic registers.
3. **Retrieval-based target identification as a first-class task:** We formulate target identification as a contrastive retrieval problem rather than a fixed-target classification problem, and report retrieval performance directly. To our knowledge, this is the first work to evaluate target identification in this form, providing a transparent account of what the model learns and enabling deployment in settings where the set of possible targets is large or unbounded.
4. **A modular, retrieval-augmented pipeline for structured stance analysis:** We integrate argument detection, target retrieval, and same-side stance classification into a compact, modular pipeline (ARGSTANCE + TARGETMATCH) that yields interpretable intermediate outputs at each stage and remains competitive with substantially larger LLMs under controlled zero-shot evaluation.
5. **Open and reproducible resources:** We release all trained models on Hugging Face (<https://huggingface.co/azza1625/argument-same-side-stance-classification>) (accessed on 8 July 2026); (<https://huggingface.co/azza1625/target-identification>) (accessed on 8 July 2026) and provide an interactive Streamlit dashboard (<https://huggingface.co/azza1625/argument-same-side-stance-classification>) (accessed on 8 July 2026).

[//arg-stance.com/](https://arg-stance.com/)) (accessed on 8 July 2026) that allows users to explore and visualize the pipeline’s predictions.

These contributions reposition the problem of argument and stance analysis around framing-aware formulation, dataset construction, and retrieval-based deployment, while providing a compact and reproducible system that is competitive with much larger models under transparent evaluation.

To establish a consistent vocabulary, we adopt the general terms *argument*, *target*, and *debate* throughout this paper, mapping them to the Kialo concepts of *claims*, *theses*, and *discussions*, respectively. We note that these concepts are not always strictly equivalent outside Kialo, and Table 1 serves as the authoritative reference for how each term is used in the remainder of the paper.

Table 1. Terminology used throughout the paper and its mapping to Kialo concepts.

Paper Term	Kialo Term	Meaning
Debate	Discussion	The full discussion space organized around a top-level question or proposition.
Target	Thesis	The proposition with respect to which a stance is evaluated.
Argument	Claim	A user-contributed statement that supports or opposes its parent node.
Parent	—	The immediate node (target or argument) to which an argument is attached.
Same-side pair	—	Two arguments with the same polarity (pro/pro or con/con) toward a shared parent.
Opposing-side pair	—	Two arguments with opposite polarity (pro/con) toward a shared parent.

2. Related Work

In this section, we review related work in argument detection and stance detection, situating our approach and contributions with respect to prior research in both domains.

2.1. Argument Detection

Argument detection, or argument identification, refers to the task of distinguishing argumentative from non-argumentative text [26]. As noted earlier, most existing approaches consider argument mining at the discourse level, where the goal is to identify the boundaries of argument components and classify them into types such as claims or premises [15,27]; while this line of work is valuable, it often relies on complex annotations and small datasets, which limits robustness and scalability.

In contrast, a smaller body of research simplifies the task by casting it as a binary (or multi-class) sequence classification problem: given a sentence or short passage, the goal is to decide whether it is argumentative, without explicitly modeling its internal components or relations. This formulation trades structural granularity for scalability, making it particularly attractive for large, heterogeneous corpora where fine-grained argument structure annotation is impractical.

Early work explored this setup on social media and political discourse. Ref. [32] framed argument detection on Brexit/Grexit tweets as sentence-level classification using lexical, syntactic, sentiment, and platform-specific features, achieving an F1 score of 0.78 with a logistic regression model. Ref. [33] annotated 300 German tweets about climate change and combined BERT-based document embeddings with XGBoost, reaching an F1 of 0.72 and illustrating the value of contextual representations for short, noisy texts. A step toward large-scale, cross-topic sequence-level argument mining came with ArgumenText [34], which treated argument detection as sentence classification over heterogeneous web sources; an attention-based neural classifier achieved an F1 of 0.738. In follow-up work on cross-topic argument mining over 25,000 sentences across eight controversial issues, ref. [35] reported within-topic F1 of 0.741 and cross-topic F1 of 0.658, highlighting both the potential and the difficulty of achieving robust generalization across

topics. More recently, ref. [36] studied the legal domain, creating a corpus of fiscal state aid decisions and compared tf-idf, SBERT, and Legal-BERT representations with several classifiers; their best performing model (tf-idf + linear SVC) achieved an F1 score of 0.70. Finally, ref. [3] evaluated large language models such as GPT-3.5-Turbo, Flan-T5-XXL, and LLaMA2 for claim detection in zero- and few-shot settings, obtaining statistically significant improvements over random baselines and demonstrating the emerging role of LLMs in computational argumentation, particularly in low-resource setups.

These works demonstrate that sequence-level formulations offer simpler and more scalable pipelines than discourse-level models, but also reveal persistent limitations, including reliance on relatively small or domain-specific datasets and limited cross-domain robustness. The present work builds on this line of research by constructing a large-scale, cross-source, distantly supervised dataset and showing that, when combined with multi-task modeling, this setup can yield argument detection models that generalize reliably across topics and domains.

2.2. Stance Detection

Understanding people's attitudes toward topics, entities, or ideas expressed in text has become a central research goal in computational social science and natural language processing. Stance detection refers to the task of automatically identifying the position or attitude (e.g., Favor, Against, or Neutral) expressed in a text toward a given target, typically a controversial issue, entity, or claim [37–39]. From a computational perspective, stance detection is often framed as a supervised classification problem, where the input consists of a (text, target) pair and the output is the author's stance category [29,40].

Stance detection has evolved into several distinct formulations, each addressing different practical challenges and generalization requirements. In its traditional form, often termed in-target stance detection, the target of interest is explicitly known and present in both training and test data [37,41]. Models are trained to classify stance for a fixed set of predefined targets. However, in real-world applications such as social media analysis, new or unseen targets frequently emerge, making this assumption unrealistic. To address this, cross-target stance detection [42–46] explores transferring knowledge from one target to another, typically adapting a classifier trained on a source topic (e.g., “Donald Trump”) to predict stance on a related but unseen one (e.g., “Joe Biden”). A further extension is zero-shot stance detection [47,48], in which the model must infer stance toward completely unseen targets without any target-specific training data, an increasingly realistic yet difficult setup.

Prior work can be grouped into three main directions. Knowledge-enhanced approaches incorporate external resources such as Wikipedia, commonsense, or discourse knowledge to better handle implicit targets and nuanced reasoning [49–52]. Multi-task learning methods leverage auxiliary objectives, such as sentiment analysis, target prediction, or stance agreement detection, to improve generalization across targets and settings [12,39,53]. Finally, emerging paradigms include knowledge distillation [54,55], prompt-based reformulations [56], and the use of large language models in few-shot or zero-shot settings [5,31,55]. Our work utilizes the same-side stance detection approach [57] and extends it with a target identification module based on contrastive learning.

3. Methodology

In this section, we describe the methodology underlying our approach. We also introduce our two core models, detailing key details and architectural design choices for each.

3.1. Pipeline Overview

Our system is designed as a modular pipeline that integrates argument detection, target retrieval, and stance reasoning (Figure 1). The pipeline proceeds in a stepwise fashion. Given an input sequence, the text is first passed through the argument detection head of our multi-task model (ARGSTANCE), which determines whether the input is argumentative or non-argumentative. If the sequence is classified as non-argumentative, the pipeline terminates without further processing.

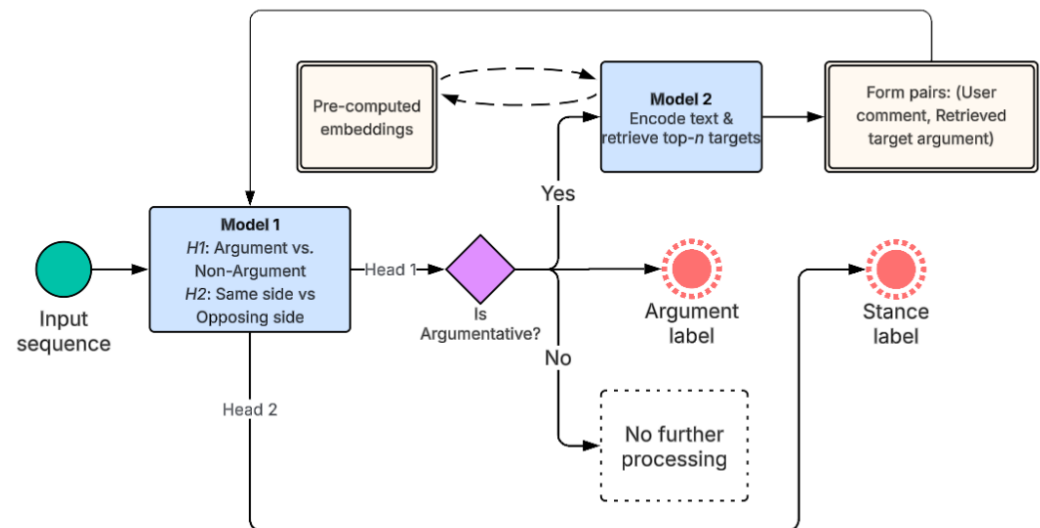


Figure 1. Pipeline overview.

If, however, the text is classified as argumentative, it is then passed to the target identification model (TARGETMATCH). This model generates an embedding representation of the input, which is compared against a stored index of target embeddings. The system retrieves the top-n closest targets that represent potential framings of the debate relevant to the input.

For each retrieved target, we then collect its associated arguments. The input text is paired with each retrieved argument, and the pair is processed by the stance-relation head of ARGSTANCE, which classifies whether the two arguments are on the same side or opposing sides of the debate. Since the stance of the retrieved argument is known relative to its target, this relation can be used to infer the stance of the user's input. In this way, our pipeline provides stance predictions contextualized by the framing of multiple targets.

3.2. ArgStance

The first component of our system is a multi-task learning model based on the ModernBERT [58] encoder. The encoder is shared across tasks and is followed by two classification heads, one for argument detection and one for stance-relation classification. To improve generalization across a wide range of topics, we extend the base encoder with a Mixture-of-Experts (MoE) layer [59]. We set the number of experts to 8 and the number of active experts to 2, following common practice in prior work on sparse Mixture-of-Experts architectures [60,61], which balances model diversity and computational efficiency. Each expert is realized as an independent feed-forward layer.

For the versions of the model that incorporate additional linguistic features (part-of-speech tags, dependency relations, named entities), we include a dedicated embedding layer for each feature type, initialized with the embedding dimension set to 8. These embeddings are concatenated and projected to 768 dimensions through a linear layer, after which they are added to the encoder's last hidden state.

Each classification head follows the same architecture: a linear projection of the encoder output into a 768-dimensional hidden space, followed by a GELU activation, a LayerNorm layer, and a final linear projection to the task-specific number of classes.

We adopt a multi-task setup since prior work has shown that jointly training related objectives in computational argumentation can improve robustness and generalization [62–64]. Argument detection and stance detection are closely connected, and sharing representations encourages the model to capture argumentative cues that benefit both tasks.

The two supervised tasks are defined as follows:

1. *Argument detection*: binary classification of a sequence as argumentative or non-argumentative.
2. *Stance detection*: binary classification of an argument pair as taking the same side or opposing sides.

A single model is trained jointly with a loss function defined as the sum of the cross-entropy losses for the two tasks.

3.3. TargetMatch

Stance is inherently relative: the label pro in one framing (e.g., “Abortion should be illegal”) corresponds to the opposite position in another framing (e.g., “Abortion is a human right”). Modeling this directly as a classification problem with thousands of possible targets quickly becomes impractical due to memory constraints, sparse supervision, and the difficulty of distinguishing semantically similar targets.

To address these challenges, we formulate target identification as a semantic retrieval problem. Specifically, we adopt a dual-encoder architecture trained with contrastive learning [65]. For each candidate target, the model is fed one positive example, which consists of an argument correctly matched with its target, along with n negative examples, or mismatched pairs. Both arguments and targets are encoded using ModernBERT encoders, and their embeddings are projected into a shared vector space. Correct argument–target pairs are optimized to have high similarity, while mismatched pairs are pushed apart. Cosine similarity is used as the matching function.

We employ an improved cosine-based contrastive loss that is robust to batches containing only positive or negative pairs. For a pair with label $y \in \{0, 1\}$ and cosine similarity s , the loss is defined as:

$$\mathcal{L}(s, y) = y \cdot (1 - s) + (1 - y) \cdot \max(0, \text{margin} - (1 - s)), \quad (1)$$

where the margin is set to 0.5 by default. This formulation penalizes positive pairs with low similarity and negative pairs that are too close, encouraging well-separated embeddings.

4. Data

The effectiveness of our approach relies on carefully curated datasets that capture both argumentative and non-argumentative discourse across a wide range of topics. Since our pipeline requires (i) distinguishing whether a piece of text constitutes an argument and (ii) identifying the stance and corresponding target of an argument, we constructed datasets tailored to these tasks. A key goal in this process was to ensure broad topical coverage, as it has been shown to play a crucial role in developing argumentative models that generalize well across domains [66]. Our primary source of argumentative material is the *Kialo* (<https://www.kialo.com>) (accessed on 8 July 2026) debate platform, which provides a rich collection of structured, high-quality arguments. For non-argumentative

material, we relied on factual text from *Wikipedia* and curated *news articles*. Below, we describe in detail the dataset construction process for each component of our system.

4.1. Datasets for ArgStance

ArgStance is trained jointly on two tasks: argument detection (argumentative vs. non-argumentative) and stance detection (same-side vs. opposing-side). Accordingly, we constructed datasets that provide supervision for both.

4.1.1. Task 1: Argument Detection

To construct a robust dataset for argument detection, we combined multiple sources and explicitly addressed potential confounding between argumentative and non-argumentative classes. Our goal was to ensure that the model learns to recognize argumentative content based on its semantic and rhetorical properties, rather than exploiting stylistic or domain-specific cues.

Argumentative class: To populate the argumentative class, we collected debates from Kialo, a large-scale platform where users engage in structured discussions around one or more theses. Debate platforms such as Kialo have been widely used in prior research as sources of argumentative content [28,67,68]. In Kialo, users can create *discussions* on specific topics, propose one or more child *theses* (central standpoints), and contribute *claims* either supporting or opposing them. Each discussion may also include optional *background information*, offering introductory context about the topic under debate. Discussions are organized hierarchically, with each claim attached either to a thesis or to another claim, forming a tree structure (Figure 2). Importantly, each contribution is related specifically to its immediate parent rather than necessarily to the root thesis. A key strength of Kialo is the quality of its content. Because user claims must first be accepted by discussion administrators, and because contributors are encouraged to provide sources and references to back up their claims, the resulting text is articulate, substantiated, and of consistently high quality.

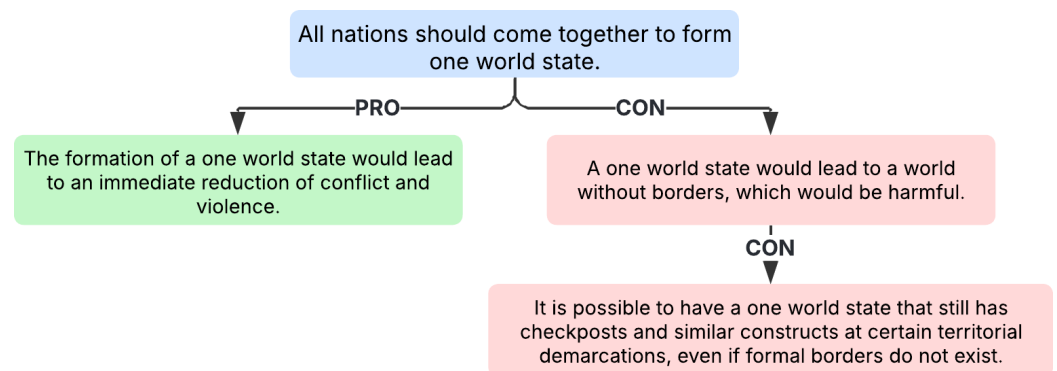


Figure 2. Tree-structured example of Kialo discussion with a single thesis.

We gathered a diverse set of discussions in two phases. First, we used the platform’s *tag* feature to ensure topical breadth, sampling the top discussions from 306 different tags according to their popularity. This yielded 2115 discussions. Each discussion was then exported as structured JSON objects containing metadata, theses, and claims annotated with stance and parent–child relations. Finally, we also retrieved the background information for each discussion where available.

We decided to retain only claims up to a depth of two, relative to the root thesis (i.e., top-level claims directly under a thesis and second-level claims that respond to those). The rationale behind this choice was that, while top-level and second-level claims are very likely to remain directly tied to the root thesis, deeper levels of discussion often

diverge into tangential aspects. Other preprocessing steps included removing empty and duplicate discussions, non-English claims and inline references (e.g., source links). Lastly, all instances were enriched with part-of-speech, dependency and entity tags using *spaCy* (<https://spacy.io>) (accessed on 8 July 2026).

The resulting pool of argumentative content contains 67,621 *unique claims* in total, drawn from 2022 *unique discussions*, covering 3310 *distinct theses*.

Non-argumentative class: To create a balanced non-argumentative counterpart, we collected factual and descriptive text from two complementary sources: *Wikipedia* and curated *news outlets*. For each Kialo discussion, we assigned a topic label using a hierarchical topic classifier and retrieved the top five Wikipedia pages associated with that topic. For news articles, we first created a curated list of URL domains grouped by topical categories such as general news, technology, health, and finance. Within each category, we selected outlets known for factual reporting, including domains such as Reuters, Associated Press, and Forbes. Using the Event Registry API, we then collected up to 5000 articles per category.

Each example in our argument detection dataset corresponds to a single sequence. For Kialo, a sequence is defined as an entire user claim, while for Wikipedia pages, news articles, and Kialo background information, sequences correspond to individual sentences. To ensure consistency in input length across all sources, we retained only sequences between 50 and 400 characters, reflecting the typical span of a short argumentative statement or factual assertion. We verified that length distributions were consistent across sources to eliminate length as a confounding variable for classification.

Mitigating source confounding: To address the risk of models learning source-specific cues rather than intrinsic argumentativeness, we implement a *cross-source injection strategy* that semantically mixes examples across domains.

We first established the argumentative semantic space by computing the centroid of all Kialo claim embeddings. Using cosine similarity, we injected semantically similar samples from Kialo background descriptions, Wikipedia, and news articles into the argumentative class. Similarly, we established the non-argumentative centroid using Wikipedia embeddings and retrieved comparable samples from Kialo background and news (excluding Kialo claims). Following a 100-sample manual audit per source, we confirmed the semantic consistency of the injected samples.

The final dataset (Table 2) integrates all four distinct sources across both classes. This design prevents the model from relying on domain or stylistic cues, thereby encouraging robust generalization based on the underlying argumentative nature of the text.

Table 2. Dataset composition for Task 1.

	Arg	Non-Arg
Kialo	67,621	-
Wiki	4962	69,030
News articles	2941	2886
Kialo bkg. information	1994	1997
Total	77,518	73,913

4.1.2. Task 2: Stance Detection

Since Kialo explicitly labels each claim as *pro* or *con*, it provides a natural supervision signal for stance. To construct training data, we retained theses with at least two supporting claims and generated all possible pairs of arguments under the same *parent* (in this case, a root thesis or top-level claim). Each pair was then labeled as *same-side* or *opposing-side*, depending on whether the stances between the two claims matched. For model input, pairs

were encoded in the form [CLS] claim 1 [SEP] claim 2 [SEP], enabling the model to jointly attend to both arguments when predicting the stance relationship.

The resulting stance dataset contained 90,328 *same-side* and 95,956 *opposing-side* pairs.

4.2. Datasets for TargetMatch

TARGETMATCH addresses the task of target identification, which we formulate as a contrastive retrieval problem. For each root thesis, we construct paired examples consisting of one *claim* and one *thesis* (concatenated with the corresponding discussion title to provide additional context). The dataset includes two types of pairs:

- *Positive pairs*: a claim and its true thesis (from the same discussion).
- *Negative pairs*: a claim and a thesis drawn from a different discussion.

Each text is encoded separately using a dual-encoder architecture based on Modern-BERT. Specifically:

- *Claim encoder input*: [CLS] claim [SEP].
- *Thesis encoder input*: [CLS] thesis [SEP] discussion_title [SEP].

To explore how different dataset configurations might affect training, we constructed multiple dataset variants. Along one dimension, we controlled the number of negatives per positive (1, 5, or 10). Along the other, we varied the type of negatives: either randomly sampled or “hard” negatives, the latter selected by embedding all theses with a pretrained *SentenceTransformer* model [69] and retrieving semantically close, but incorrect, theses.

5. Experiments

In this section, we present the experimental design used to assess our proposed models. We first outline the general setup and then describe the specific experimental configurations for each model.

5.1. Experimental Setup

All experiments were conducted using the PyTorch v2.7.1 deep learning framework [70], with pretrained models accessed via the Hugging Face Transformers library (<https://github.com/huggingface/transformers>) (accessed on 8 July 2026). Training was carried out on Lambda Labs cloud infrastructure using GH200 instances.

For all experiments, we used the AdamW optimizer [71] with a learning rate of 5×10^{-5} , batch size of 128, and training was performed for three epochs. Model checkpoints were saved based on the lowest validation loss. For ARGSTANCE, total loss was computed as the average of the two task-specific cross-entropy losses. For TARGETMATCH, we used a cosine-based contrastive loss with $\text{margin} = 0.5$. In the multi-task learning (MTL) setting, the total loss was computed as the average of the losses for the two tasks.

For argument detection and stance detection, we report accuracy and F_1 scores. For target identification, we evaluate retrieval performance using top- k accuracy, a standard metric in information retrieval tasks that measures whether the correct thesis appears among the top- k retrieved candidates. Across all datasets, we reserved 10% of the data for testing.

Train/Test Splitting and Topical Overlap

For the stance-detection task, training and evaluation data consist of claim pairs constructed under shared parents (a thesis or a top-level claim) within Kialo discussions. We performed the train/test split at the level of individual claim pairs, rather than at the parent, thesis, or discussion level. This choice was deliberate: by maintaining a substantial overlap of theses between the two splits, we sought to expose the model to the broadest possible range of debate framings during training, in line with our broader goal of maximizing

topical breadth. We hypothesized that such exposure would benefit generalization to unseen claims at inference time, given the diversity of argumentative styles that even a single thesis may attract.

We acknowledge, however, that this design entails a trade-off that should be made explicit. Because related claims sharing the same parent may appear across both splits, the reported stance detection metrics reflect not only generalization to previously unseen claims but also a degree of familiarity with the underlying debate context. Stricter splitting regimes (e.g., at the parent, thesis, or discussion level), would more directly measure generalization to previously unseen framings, and we expect performance under such regimes to be somewhat lower than the numbers reported here. The reported stance detection scores should therefore be interpreted as an upper bound on the model's performance under the current training configuration, and a systematic comparison across split regimes remains an important direction for future work.

5.2. Training and Evaluation of ArgStance

ARGSTANCE was assessed on two tasks: (i) detecting whether a text is argumentative, and (ii) classifying a pair of arguments as being on the same or opposing side of a thesis. We explored two main experimental dimensions:

1. *Model size*: We compared the *base* and *large* variants of the ModernBERT architecture.
2. *Feature augmentation*: We tested whether including additional linguistic features (part-of-speech tags, dependency relations, and named entities) improved performance.

For clarity, the variant incorporating additional linguistic features is denoted as *-ef* in all figures and tables.

5.2.1. Baselines

As baselines for argument detection, we refer to the results reported in [28,30], which follow a similar framing of the task. To complement these encoder-based baselines, we additionally include a recent LLM-based reference point from [3], who evaluate a 20B-parameter Flan-UL2 model in a zero-shot setting. We stress, however, that due to the highly contextual and situational nature of argumentation, and because the underlying datasets differ across these works, cross-model comparisons are inherently difficult and should be read as indicative rather than as direct head-to-head evaluations.

For same-side stance detection, we use the results from [57] as a baseline. Note that only accuracy scores are available in that work.

5.2.2. Ablation Studies

To quantify the contribution of different components of ARGSTANCE, we performed ablations along the following dimensions:

1. *Multi-task learning (MTL)*: comparing the full MTL setup against single-task variants (argument detection only; stance detection only).
2. *Mixture-of-Experts (MoE) layer*: evaluating performance with and without the MoE component.

For brevity, experiments were performed using the ModernBERT-large variant of our architecture.

5.3. Training and Evaluation of TargetMatch

TARGETMATCH addresses target identification through a dual-encoder retrieval setup, implemented with the ModernBERT -base variant of our architecture. Experimental conditions were varied along two axes:

1. Number of negatives: For each positive pair, we sampled 1, 5, or 10 negative examples.
2. Type of negatives: We compared random negatives to hard negatives (as defined in Section 4.2).

Baseline

As a baseline, we use the results from [39], where target identification is cast as a classification task with a large output space, while their setup differs from ours, it provides a useful point of comparison.

5.4. Full-Pipeline Stance Detection Experiments

In the context of this study, full-pipeline stance detection refers to the complete inference path that performs target retrieval followed by stance inference. Specifically, for a given claim, we first retrieve the most relevant theses using TARGETMATCH. Then, for each retrieved thesis, we take its associated arguments and use the stance-relation head (Head 2) of ARGSTANCE to classify each pair, composed of the input argument and a retrieved target argument, as either same-side or opposing-side. The final stance label of the given argument is inferred via majority voting over the aggregated predictions across these pairs.

Evaluating this full-pipeline process presents a framing challenge: model outputs are relative to retrieved targets, whereas gold labels are defined with respect to the argument's original target. To address this, we adopt a two-pronged evaluation strategy: (i) a strict conditional evaluation that measures performance only when the original (gold) target is among the top- k retrieved, and (ii) a framing-aware evaluation that automatically aligns retrieved targets to the original target using an NLI-based (Natural Language Inference) alignment model [72]; mapped predictions are flipped when the mapping indicates *contradiction* with *confidence* ≥ 0.9 . For the first evaluation strategy, we set top- k to 3.

6. Results and Discussion

In this section, we present the outcomes of our experiments and discuss their implications. Results are organized by model and evaluation setting, beginning with argument detection and stance detection (ARGSTANCE), followed by target retrieval (TARGETMATCH), and finally the full-pipeline stance detection results.

6.1. Performance of ArgStance

Tables 3 and 4 present the performance of our multi-task ModernBERT models across the two tasks. The best results are achieved with ModernBERT-base for argument detection (Task 1) and ModernBERT-large for stance detection (Task 2), both without additional linguistic features, yielding F1 scores of 0.94 and 0.84, respectively, and outperforming all of the corresponding baselines. Examining model size, no notable difference is observed for Task 1, but the large variant provides a clear advantage in the more challenging stance-detection task. Adding linguistic features (POS, dependency, and named entities) did not lead to measurable gains, suggesting that ModernBERT already captures sufficient syntactic and semantic information.

Beyond the encoder-based baselines, Table 3 also reports a recent LLM reference point from [3], who evaluate a 20B-parameter Flan-UL2 model in a zero-shot setting and achieve a Macro- F_1 of 0.868 on argument detection. As emphasized earlier, the underlying datasets differ across these works and a direct numerical comparison is therefore not meaningful. Nonetheless, this baseline provides a useful indication that, under zero-shot conditions, even a generative LLM more than fifty times larger than our model does not necessarily surpass a compact, task-specialized architecture, an observation that aligns with the broader LLM comparison reported later in this section.

Table 3. Benchmark results for the argument detection task. Best-performing scores are highlighted in bold.

Method	Encoder	Argument Detection	
		Accuracy	F ₁
[28]	Flan-UL2 (20B)	0.918	0.922
[30]		-	0.925
[3]		-	0.868
ArgStance	MBERT-base	0.941	0.940
	MBERT-large	0.935	0.935
	MBERT-ef	0.914	0.913

Table 4. Benchmark results for same-side stance detection task. Best-performing scores are highlighted in bold.

Method	Encoder	Stance Detection	
		Accuracy	F ₁
[57]		0.770	0.743
ArgStance	MBERT-base	0.809	0.804
	MBERT-large	0.840	0.837
	MBERT-ef	0.820	0.815

The ablation studies provide further insight (Table 5). First, examining the effect of multi-task learning, we find no impact on Task 1 but a notable improvement on Task 2: for ModernBERT-large, removing MTL reduces stance F1 by 1.1%. Second, regarding the mixture-of-experts (MoE) layer, its removal causes a 1.5% drop in stance F1, again underscoring that Task 2 benefits more from these design choices.

Table 5. Ablation results for ArgStance. Best-performing scores for each ablation and task are highlighted in bold.

Variant	MTL		MOE	
	Task 1	Task 2	Task 1	Task 2
Without	0.945	0.826	0.940	0.822
With	0.935	0.837	0.935	0.837

Overall, the best-performing configuration is ModernBERT-large with MTL and MoE, which balances robustness across both tasks and achieves the strongest stance detection performance.

6.1.1. Evaluating Our Cross-Source Injection Strategy

To assess the impact of our cross-source injection strategy, we trained a baseline single-source model (Kialo claims vs. Wikipedia/news sentences) and evaluated it against a rigorous, unseen cross-source gold dataset (700 balanced, manually annotated examples).

As shown in Table 6, the single-source model achieved high in-distribution performance (0.96 Macro-F₁), but its score dropped markedly on the cross-source set (0.659 Macro-F₁). In contrast, the cross-source model demonstrated robust generalization, maintaining strong performance both in-domain (0.94 Macro-F₁) and cross-domain (0.817 Macro-F₁).

This cross-domain evaluation complements standard test results by rigorously quantifying domain robustness and generalization to new writing styles.

Table 6. Evaluation of cross-source versus single-source argument detection models.

Variant	Single-Source		Cross-Source	
	Test Set	Gold Set	Test Set	Gold Set
Macro-F ₁	0.96	0.659	0.94	0.817

6.1.2. Source-Level Error Analysis

While Section 6.1.1 established the aggregate effectiveness of the cross-source injection strategy on a gold benchmark, that evaluation does not reveal where source-specific patterns may still influence model behavior within the cross-source model. To probe this directly, we conducted a source-level error analysis of the argument detection model, examining performance variation across the four sources used during training (Kialo claims, Kialo background information, Wikipedia, and news articles), and inspecting representative misclassifications. Because cross-source injection is designed precisely to dilute source-specific cues, this analysis serves as a diagnostic for whether and where the strategy is fully effective.

Per-Source Classification Accuracy

On the argumentative class (Table 7), the model performs best on Kialo claims (97.9%), which constitute the bulk of the argumentative training data and are typically concise, assertive, and explicitly pro or con relative to a thesis. Performance moderately decreases for Kialo background information (90.5%) and news articles (90.0%), and drops further for Wikipedia (84.8%). The latter sources frequently express arguments through embedded justifications, implicit inferential relations, or descriptive exposition rather than overt stance-taking, yielding a less stereotypical argumentative surface form.

Table 7. Classification performance per source for the argumentative class.

Source	Correct	Total	Percentage
Kialo	6596	6736	0.979
Wikipedia	419	494	0.848
News articles	253	281	0.900
Kialo bkg. info	202	223	0.905

On the non-argumentative class (Table 8), the pattern differs. The model performs best on news articles (98.1%) and Wikipedia (90.3%), but is weakest on Kialo background information (77.3%). The drop on Kialo background is the most informative observation here: although these sections are intended to be descriptive, they often contain evaluative language, implicit causal claims, or preparatory context that resembles argumentative phrasing. Because Kialo also dominates the argumentative class, stylistic similarity within the same platform appears to confound the model. This suggests that while cross-source injection successfully dilutes source identity as a global cue, a residual stylistic imprint persists when a single platform contributes text to both classes.

Table 8. Classification performance per source for the non-argumentative class.

Source	Correct	Total	Percentage
Wikipedia	6255	6920	0.903
News articles	272	277	0.981
Kialo bkg. info	164	212	0.773

Inspection of Misclassified Instances

A closer look at the misclassifications themselves clarifies the nature of the residual errors. Among non-argumentative instances incorrectly predicted as argumentative (Table 9), many contain implicitly normative or causal content that resembles argumentative structure such as “inequity in education is hard to avoid because of inequities in socioeconomic standing...” or “vaccinating animals is important not only to the animals’ health but also to human health...” and the model’s predictions arguably align more closely with the semantic structure of these sequences than the distant labels do. Similar borderline cases appear in Kialo background and news. Such examples expose a structural limitation of distant supervision: label noise is inevitable when argumentativeness is inferred heuristically from document structure rather than annotated manually.

Table 9. Examples of misclassified non-argumentative texts.

Text	Source	True Label	Predicted
However, inequity in education is hard to avoid because of inequities in socioeconomic standing, race, gender, and disability.	Wikipedia	Non-arg.	Arg.
In many cases vaccinating animals is important not only to the animals’ health but also to human health and prosperity.	Wikipedia	Non-arg.	Arg.
Chevy vehicles outperform every other brand in areas of durability, user-friendly interface, affordability and vehicle performance.	Kialo bkg. info	Non-arg.	Arg.
He is solid just about everywhere, but he’s spectacular nowhere.	News articles	Non-arg.	Arg.

Conversely, argumentative instances misclassified as non-argumentative (Table 10) reveal a more systematic pattern. Many errors occur when the argumentative force is embedded within reported speech or institutional attribution (“In a statement, the body said the measures aim to mitigate excessive risk-taking...”), or when strong evaluative claims are wrapped in encyclopedic phrasing (“Tedros therefore concluded that a strategy of achieving herd immunity... is ‘morally unconscionable and unfeasible’”). More broadly, third-person and indirectly attributed argumentation consistently challenges the model: while Kialo claims are typically written in first person or directly assertive form, sequences such as “He writes that it would be morally wrong...” embed argumentation within indirect attribution and diverge from the canonical surface form the model has internalized. A small additional subset of errors involves genuinely non-argumentative Kialo claims that nevertheless inherit the argumentative label via distant supervision (e.g., references to fictional characters or trivia), which challenges the underlying assumption that all Kialo claims are argumentative.

The preceding residual error patterns reflect three intertwined factors: (1) unavoidable label noise introduced by distant supervision, particularly for sources where descriptive content frequently overlaps with argumentative phrasing; (2) genuine ambiguity in what constitutes an argument, especially in third-person, reported, or implicitly inferential constructions; and (3) lingering stylistic influence of source domains within a single platform, most visibly Kialo, which contributes to both classes. For the cross-source injection strategy specifically, the analysis indicates that the approach is effective at neutralizing source identity as a coarse global cue, but does not fully eliminate fine-grained, within-platform stylistic confounds. Addressing these residual effects would likely require complementary mechanisms such as supplementary handling of reported arguments, expanded negative sampling within the dominant source, or explicit modeling of argument frames beyond canonical pro–con structures.

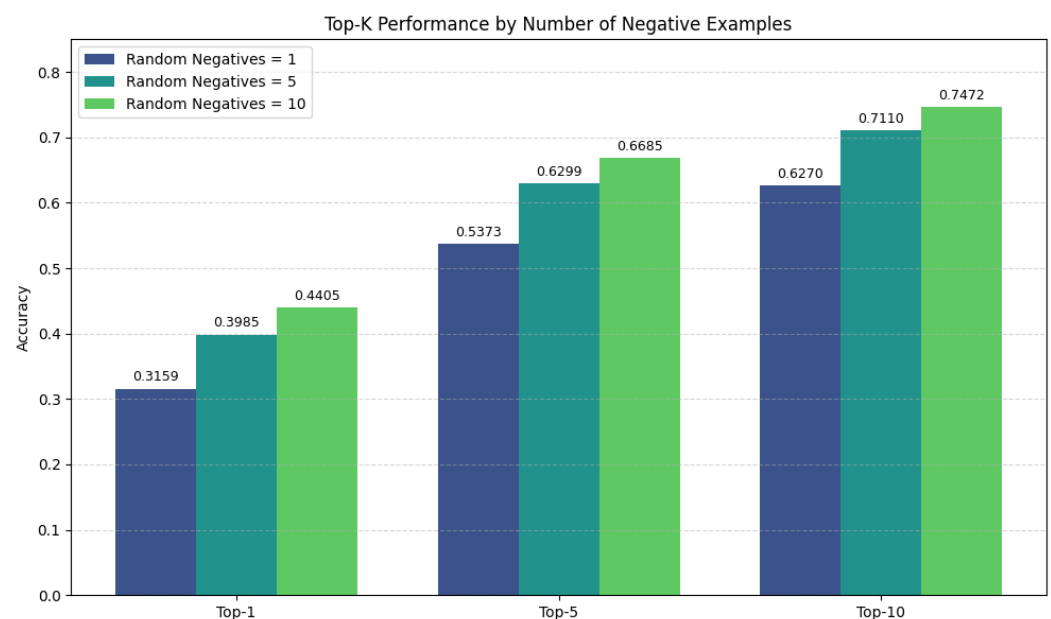
Table 10. Examples of misclassified argumentative texts.

Text	Source	True Label	Predicted
An example of this is Zartan, a member of the Ninja Force who can shapeshift.	Kialo	Arg.	Non-arg.
In a statement, the body said the measures aim to mitigate excessive risk-taking by the institutions.	News articles	Arg.	Non-arg.
Tedros therefore concluded that a strategy of achieving herd immunity by infection, rather than vaccination, is “morally unconscionable and unfeasible”.	News articles	Arg.	Non-arg.
He writes that it would be morally wrong for humans to tamper with fundamental aspects of themselves (or their children) in an attempt to overcome universal human limitations...	Wikipedia	Arg.	Non-arg.
Similarly, the worry about agriculture and food consumption is also a disputed fact: some consider there is plenty of food (more arable land used, better technology, more efficiency, etc.)	Kialo bkg. info	Arg.	Non-arg.

6.2. Performance of TargetMatch

To our knowledge, TARGETMATCH is the first model to treat target identification as a dedicated retrieval task with explicitly reported retrieval performance under a contrastive learning setup; while contrastive learning has been employed in prior stance detection work [48,73,74], it has typically been used as an intermediate representation-learning step supporting zero-shot or cross-target transfer, with reported results referring to downstream stance classification accuracy rather than target retrieval accuracy. In TARGETMATCH, by contrast, target retrieval is a first-class output of the model, and its performance is measured and reported directly, providing a more transparent account of what the model actually learns to do.

Figures 3 and 4 show retrieval performance for TARGETMATCH under different configurations. Two trends stand out. First, increasing the number of negatives during training consistently improves performance for both random and hard negative sampling. Second, while hard negatives lead to stronger top-1 accuracy, random negatives perform better overall at larger values of k, particularly for top-5 and top-10 accuracy.

**Figure 3.** Top-k accuracy for TargetMatch with random negatives.

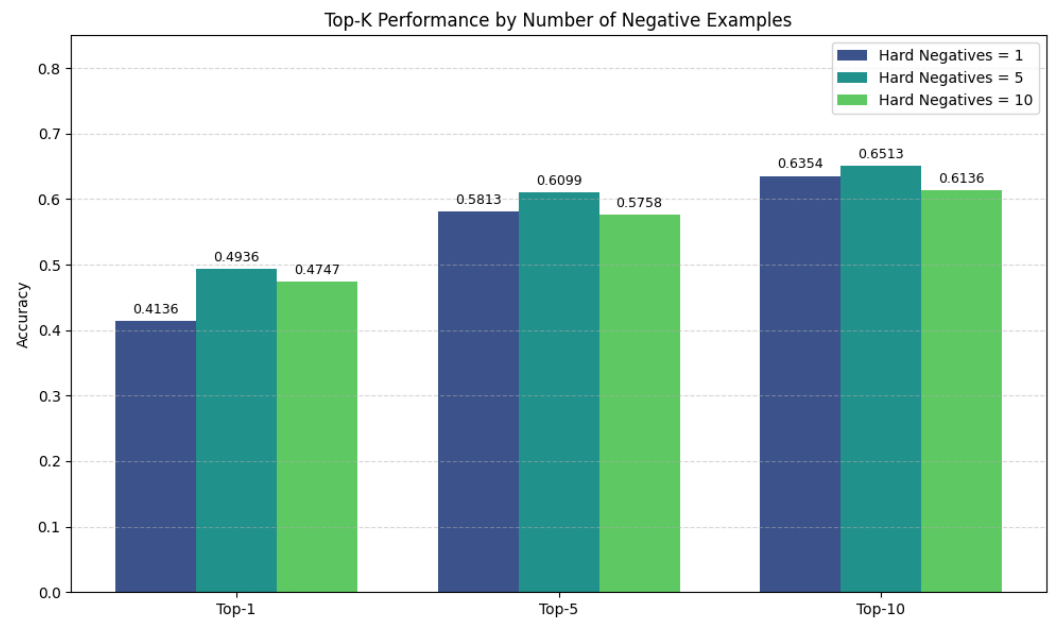


Figure 4. Top-k accuracy for TargetMatch with hard negatives.

Comparing the two regimes reveals a trade-off between local discriminability and global calibration. Hard negatives train the model to resolve fine-grained distinctions between the correct thesis and its nearest competitors, optimizing for the hardest part of the ranking problem. Random negatives, by contrast, train the model to place the correct thesis consistently above the bulk of unrelated candidates, yielding a more uniformly structured embedding space. Concretely, the ten-random-negative configuration achieves the best top-5 and top-10 scores (0.67 and 0.75, respectively), while the five-hard-negative configuration achieves the best top-1 score (0.49). For our downstream pipeline, global calibration is what matters: ARGSTANCE receives a shortlist of k candidate theses and performs stance–relation reasoning over all of them, so a model that is locally sharp but globally miscalibrated would degrade full-pipeline performance. The ten-random-negative model is therefore the most suitable configuration, as approximately three out of four claims have their gold thesis among the top-10 retrieved candidates, providing ARGSTANCE with a high-recall shortlist from which to reason.

6.3. Performance of Full-Pipeline Stance Detection

Finally, we evaluate the full pipeline by combining TARGETMATCH (retrieval) with ARGSTANCE (stance detection). Two complementary approaches are considered, as described in Section 5.4.

For the first approach, which measures full-pipeline performance only when the gold target is among the top- k retrieved, the best model achieves an F1 score of **0.92** when 5 or more arguments are taken from the gold target (Table 11). Generally, performance improves as more arguments are sampled, although even with only one example, F1 remains quite high. Notably, the fraction of test cases where the gold target is within the top-3 retrieved is **0.429**.

Table 11. Full-pipeline stance detection performance with gold target. Best-performing score is highlighted in bold.

Metric	# of Claims			
	1	3	5	10
Macro-F ₁	0.880	0.919	0.920	0.922

For the second approach, which incorporates contradiction handling via NLI-based label flipping and evaluates stance predictions against all retrieved theses, the best model achieves F_1 scores above **0.84** for top-10 retrieval (Table 12). Although performance is slightly lower compared to the first approach, it improves by more than 2% as the number of retrieved targets increases, highlighting the value of leveraging multiple perspectives.

Table 12. Full-pipeline stance detection performance with all targets. Best-performing scores are highlighted in bold.

Metric	Top-n	# of Claims			
		1	3	5	10
Macro- F_1	Top-3	0.812	0.832	0.830	0.820
	Top-10	0.846	0.848	0.850	0.848

While stance detection remains a challenging task, we believe that the proposed pipeline achieves strong performance and demonstrates practical feasibility for large-scale argument mining

6.4. Comparison with Closed- and Open-Source LLMs

Recent work employs LLMs for argumentation tasks, ranging from zero-shot baselines to auxiliary support for fine-tuning [3,75,76]. To position our approach, we evaluated closed-source (gpt-5-mini) and open-source (Llama3-7B, Mistral-7B, Qwen2.5-7B/0.5B) models on argument and stance detection using zero-shot inference (Macro- F_1).

As shown in Table 13, gpt-5-mini yields the strongest zero-shot performance (0.84 argument detection, 0.94 stance), marginally outperforming our ARGSTANCE model on both tasks. In the open-source setting, 7B-parameter models achieve moderate results (0.62–0.78 on argument detection, 0.82–0.91 on stance) and consistently fall short of our fine-tuned 0.35B backbone, despite being roughly twenty times larger. Notably, the size-comparable Qwen2.5-0.5B performs poorly on both tasks (0.41/0.33), confirming that parameter count alone is insufficient without task-specific adaptation. Figures 5 and 6 visualize the parameter–performance trade-off across all evaluated systems, illustrating that the proposed model attains performance competitive with much larger systems while remaining substantially smaller.

Table 13. LLM comparison experiment results.

Model	Argument	Stance
Ours-0.35B	0.817	0.922
gpt-5-mini	0.840	0.944
Llama3-7B	0.615	0.824
Mistral-7B	0.781	0.854
Qwen 2.5-7B	0.729	0.907
Qwen 2.5-0.5B	0.406	0.334

These results indicate that while closed-source LLMs achieve the strongest zero-shot results, the proposed compact models remain competitive, outperform size-comparable open-source baselines, and offer advantages in reproducibility, deployment cost, and controllable intermediate predictions.

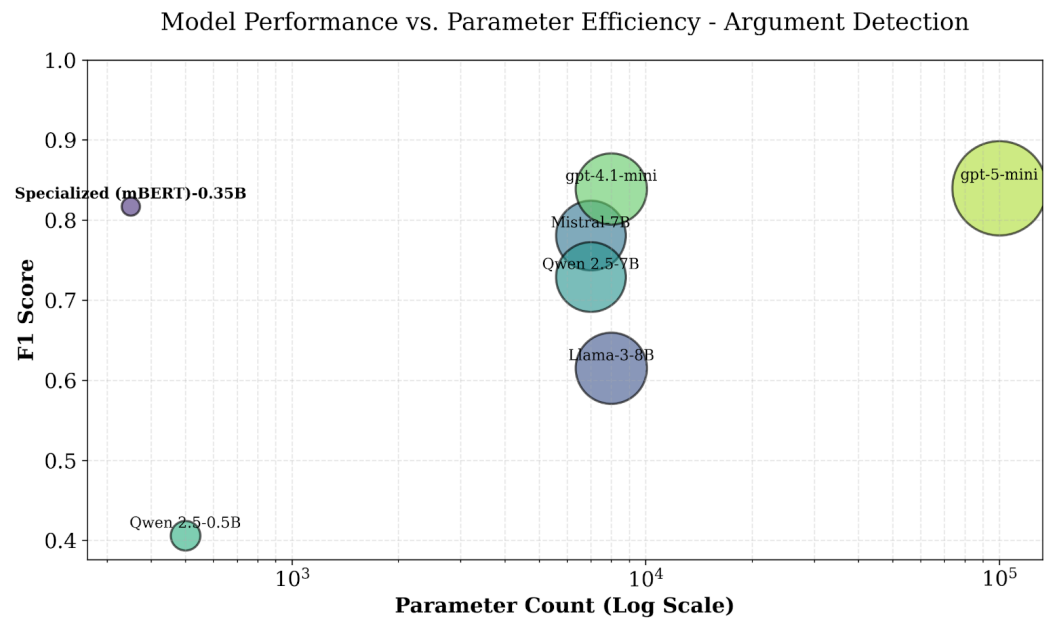


Figure 5. Parameter count vs. performance comparison between open- and closed-source LLMs and our approach for the argument detection task.

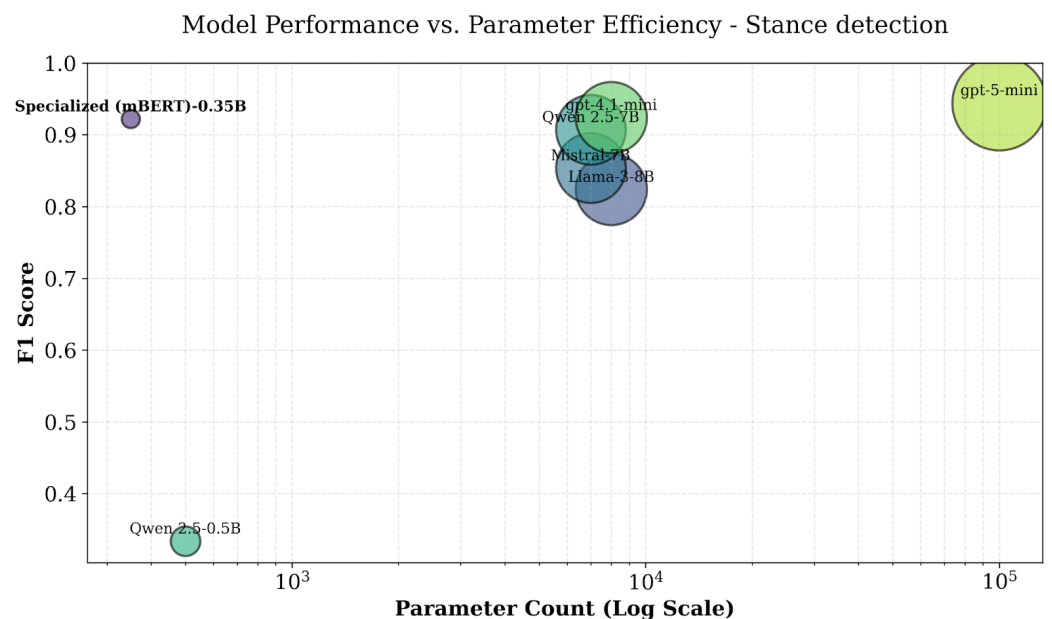


Figure 6. Parameter count vs. performance comparison between open- and closed-source LLMs and our approach for the same-side stance-detection task.

It is important to note that this comparison is inherently contextual and should not be interpreted as a definitive assessment of LLM capabilities on these tasks. The LLMs were evaluated in a zero-shot setting with a standardized prompt, without any task-specific fine-tuning, few-shot examples, or prompt optimization. It is plausible that carefully engineered prompting strategies or fine-tuned LLM variants would yield substantially stronger results. Within these constraints, however, the results support the view that compact, task-specialized architectures remain highly competitive for computational argumentation and are often preferable when efficiency, controllability, and reproducibility are critical.

7. Conclusions and Future Work

In this work, we introduced a modular, retrieval-augmented pipeline for computational argumentation that integrates argument detection, target identification, and same-side stance classification. The proposed ARGSTANCE model, built on a ModernBERT backbone, achieves strong performance on both tasks, while our cross-source injection strategy mitigates confounding domain effects and supports broader generalization across linguistic registers. In parallel, the contrastive retrieval model TARGETMATCH treats target identification as a first-class retrieval task, providing a transparent account of what the component learns. Under controlled zero-shot comparisons with closed- and open-source LLMs, our compact models remain competitive with substantially larger systems and outperform size-comparable open-source baselines, while offering advantages in reproducibility, deployment cost, and interpretable intermediate predictions. Together, these components form a system that supports structured argument and stance analysis while remaining compact, reproducible, and inspectable at each stage of the pipeline.

Several limitations of the present study also point to natural directions for future work. First, the train/test split for stance detection was performed at the level of individual claim pairs (Section 5), allowing substantial overlap of theses across splits; a systematic evaluation under stricter parent-, thesis-, or discussion-level splits would more directly measure generalization to previously unseen debate framings. Second, the dataset construction process inherits the distantly supervised structure of Kialo, including the assumption that all claims are argumentative, an assumption our error analysis (Section 6.1.2) shows does not always hold; refining the data creation process through enhanced filtering and validation heuristics could further reduce supervision noise. Third, the present qualitative error analysis focuses on the argument detection component, and extending it to target retrieval, stance relation classification, and full-pipeline error propagation would strengthen the evaluation. Fourth, the LLM comparison is conducted under zero-shot, standardized-prompt conditions and should be read as indicative rather than definitive; few-shot, retrieval-augmented, or fine-tuned LLM variants would likely yield stronger results. Beyond these, further investigation of expert routing and efficiency trade-offs within the Mixture-of-Experts layer, and extension of the cross-source injection strategy to additional domains or multilingual settings, remain promising directions.

Author Contributions: Conceptualization, A.C.; Writing—Original draft, A.C.; Data curation, A.C.; Methodology, A.C. and C.D.; Software, A.C. and C.D.; Writing—Review and Editing, A.C. and C.D.; Supervision, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study were derived from Kialo. The authors do not have permission to publicly release these data, as redistribution is not permitted by Kialo's terms of use.

Acknowledgments: During the preparation of this manuscript/study, the author(s) used Claude Sonnet 4.6 for the purposes of improving the language and readability of the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gleize, M.; Shnarch, E.; Choshen, L.; Dankin, L.; Moshkowich, G.; Aharonov, R.; Slonim, N. Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*; Korhonen, A., Traum, D., Màrquez, L., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2019; pp. 967–976. <https://doi.org/10.18653/v1/P19-1093>.
2. Gretz, S.; Friedman, R.; Cohen-Karlik, E.; Toledo, A.; Lahav, D.; Aharonov, R.; Slonim, N. A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis. *Proc. Aaai Conf. Artif. Intell.* **2020**, *34*, 7805–7813. <https://doi.org/10.1609/aaai.v34i05.6285>.
3. Chen, G.; Cheng, L.; Luu, A.T.; Bing, L. Exploring the Potential of Large Language Models in Computational Argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand*; Ku, L.W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2024; Volume 1, pp. 2309–2330. <https://doi.org/10.18653/v1/2024.acl-long.126>.
4. Bao, J.; He, Y.; Sun, Y.; Liang, B.; Du, J.; Qin, B.; Yang, M.; Xu, R. A Generative Model for End-to-End Argument Mining with Reconstructed Positional Encoding and Constrained Pointer Mechanism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2022; pp. 10437–10449. <https://doi.org/10.18653/v1/2022.emnlp-main.713>.
5. Sharma, A.; Gupta, A.; Bilalpur, M. Argumentative Stance Prediction: An Exploratory Study on Multimodality and Few-Shot Learning. In *Proceedings of the 10th Workshop on Argument Mining, Singapore*; Alshomary, M., Chen, C.C., Muresan, S., Park, J., Romberg, J., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 167–174. <https://doi.org/10.18653/v1/2023.argmining-1.18>.
6. Mirzakhmedova, N.; Gohsen, M.; Chang, C.H.; Stein, B. Are Large Language Models Reliable Argument Quality Annotators? In *Robust Argumentation Machines*; Cimiano, P., Frank, A., Kohlhase, M., Stein, B., Eds.; Springer: Cham, The Netherlands, 2024; pp. 129–146.
7. Bar-Haim, R.; Eden, L.; Friedman, R.; Kantor, Y.; Lahav, D.; Slonim, N. From Arguments to Key Points: Towards Automatic Argument Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 4029–4039. <https://doi.org/10.18653/v1/2020.acl-main.371>.
8. Gurcke, T.; Alshomary, M.; Wachsmuth, H. Assessing the Sufficiency of Arguments through Conclusion Generation. In *Proceedings of the 8th Workshop on Argument Mining, Punta Cana, Dominican Republic*; Al-Khatib, K., Hou, Y., Stede, M., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 67–77. <https://doi.org/10.18653/v1/2021.argmining-1.7>.
9. Reimers, N.; Schiller, B.; Beck, T.; Daxenberger, J.; Stab, C.; Gurevych, I. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*; Korhonen, A., Traum, D., Màrquez, L., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2019; pp. 567–578. <https://doi.org/10.18653/v1/P19-1054>.
10. Lippi, M.; Torroni, P. Argumentation Mining: State of the Art and Emerging Trends. *ACM Trans. Internet Technol.* **2016**, *16*, 1–25. <https://doi.org/10.1145/2850417>.
11. Ouertatani, A.; Gasmi, G.; Latiri, C. Parsing argued opinion structure in Twitter content. *J. Intell. Inf. Syst.* **2021**, *56*, 327–353.
12. Upadhyaya, A.; Fisichella, M.; Nejd, W. A multi-task model for sentiment aided stance detection of climate change tweets. In *Proceedings of the International AAAI Conference on Web and Social Media, Austin, TX, USA, 30 April–4 May 2023*; Volume 17, pp. 854–865.
13. Cotfas, L.A.; Delcea, C.; Roxin, I.; Ioanăș, C.; Gherai, D.S.; Tajariol, F. The longest month: Analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access* **2021**, *9*, 33203–33223.
14. Hasan, K.S.; Ng, V. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Kerrville, TX, USA, 2014; pp. 751–762.
15. Sazid, M.T.; Mercer, R.E. A unified representation and a decoupled deep learning architecture for argumentation mining of students’ persuasive essays. In *Proceedings of the 9th Workshop on Argument Mining*; International Conference on Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 74–83.
16. Cao, L. Autoam: An end-to-end neural model for automatic and universal argument mining. In *Proceedings of the International Conference on Advanced Data Mining and Applications*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 517–531.
17. Bao, J.; Fan, C.; Wu, J.; Dang, Y.; Du, J.; Xu, R. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*; Association for Computational Linguistics: Kerrville, TX, USA, 2021; Volume 1, pp. 6354–6364.

18. Mushtaq, U.; Cabessa, J. Argument classification with BERT plus contextual, structural and syntactic features as text. In *Proceedings of the International Conference on Neural Information Processing*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 622–633.
19. Fu, Y.; Wang, S.; Li, X.; Li, D.; Li, Y.; Liao, J.; Zheng, J. Hierarchical neural network: Integrate divide-and-conquer and unified approach for argument unit recognition and classification. *Inf. Sci.* **2023**, *624*, 796–810.
20. Schmidt, F.M.; Cohen, A.; Gottifredi, S.; García, A.J. Improving natural language arguments' identification by leveraging semantic similarity. *Inf. Sci.* **2025**, *733*, 122954.
21. Stab, C.; Gurevych, I. Parsing argumentation structures in persuasive essays. *Comput. Linguist.* **2017**, *43*, 619–659.
22. Habernal, I.; Gurevych, I. Argumentation mining in user-generated web discourse. *Comput. Linguist.* **2017**, *43*, 125–179.
23. Bosc, T.; Cabrio, E.; Villata, S. DART: A dataset of arguments and their relations on Twitter. In *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*; European Language Resources Association (ELRA): Paris, France, 2016; pp. 1258–1263.
24. Poudyal, P.; Šavelka, J.; Ieven, A.; Moens, M.F.; Goncalves, T.; Quaresma, P. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 67–75.
25. Mayer, T.; Cabrio, E.; Villata, S. Transformer-based argument mining for healthcare applications. In *Proceedings of the ECAI 2020–24th European Conference on Artificial Intelligence*, Online, 29 August–8 September 2020.
26. Lawrence, J.; Reed, C. Argument mining: A survey. *Comput. Linguist.* **2020**, *45*, 765–818.
27. Morio, G.; Ozaki, H.; Morishita, T.; Yanai, K. End-to-end argument mining with cross-corpora multi-task learning. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 639–658.
28. Al Khatib, K.; Wachsmuth, H.; Hagen, M.; Köhler, J.; Stein, B. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Kerrville, TX, USA, 2016; pp. 1395–1404.
29. Alturayef, N.; Luqman, H.; Ahmed, M. A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput. Appl.* **2023**, *35*, 5113–5144.
30. Cheng, L.; Bing, L.; He, R.; Yu, Q.; Zhang, Y.; Si, L. IAM: A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Kerrville, TX, USA, 2022.
31. Irani, A.; Park, J.Y.; Esterling, K.; Faloutsos, M. WIBA: What Is Being Argued? A Comprehensive Approach to Argument Mining. In *Social Networks Analysis and Mining*; Aiello, L.M., Chakraborty, T., Gaito, S., Eds.; Springer: Cham, The Netherlands, 2025; pp. 337–354.
32. Dusmanu, M.; Cabrio, E.; Villata, S. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Kerrville, TX, USA, 2017; pp. 2317–2322.
33. Schaefer, R.; Stede, M. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 53–58.
34. Stab, C.; Daxenberger, J.; Stahlhut, C.; Miller, T.; Schiller, B.; Tauchmann, C.; Eger, S.; Gurevych, I. Argumenttext: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*; Association for Computational Linguistics: Kerrville, TX, USA, 2018; pp. 21–25.
35. Stab, C.; Miller, T.; Schiller, B.; Rai, P.; Gurevych, I. Cross-topic Argument Mining from Heterogeneous Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium*; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2018; pp. 3664–3674. <https://doi.org/10.18653/v1/D18-1402>.
36. Grundler, G.; Santin, P.; Galassi, A.; Galli, F.; Godano, F.; Lagioia, F.; Palmieri, E.; Ruggeri, F.; Sartor, G.; Torroni, P. Detecting arguments in cjeu decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*; International Conference on Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 143–157.
37. Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; Cherry, C. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*; Association for Computational Linguistics: Kerrville, TX, USA, 2016; pp. 3945–3952.
38. Küçük, D.; Can, F. Stance detection: A survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–37.
39. Li, Y.; Garg, K.; Caragea, C. A New Direction in Stance Detection: Target-Stance Extraction in the Wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; Volume 1, pp. 10071–10085. <https://doi.org/10.18653/v1/2023.acl-long.560>.

40. Alkhalifa, R.; Zubiaga, A. Capturing stance dynamics in social media: Open challenges and research directions. *Int. J. Digit. Humanit.* **2022**, *3*, 115–135.
41. Li, Y.; Caragea, C. A multi-task learning framework for multi-target stance detection. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 2320–2326.
42. Augenstein, I.; Rocktäschel, T.; Vlachos, A.; Bontcheva, K. Stance detection with bidirectional conditional encoding. *arXiv* **2016**, arXiv:1606.05464.
43. Xu, C.; Paris, C.; Nepal, S.; Sparks, R. Cross-target stance classification with self-attention networks. *arXiv* **2018**, arXiv:1805.06593.
44. Zhang, B.; Yang, M.; Li, X.; Ye, Y.; Xu, X.; Dai, K. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 3188–3197.
45. Liang, B.; Fu, Y.; Gui, L.; Yang, M.; Du, J.; He, Y.; Xu, R. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, Ljubljana, Slovenia, 19–23 April 2021; pp. 3453–3464.
46. Khiabani, P.J.; Zubiaga, A. Few-shot learning for cross-target stance detection by aggregating multimodal embeddings. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 2081–2090.
47. Allaway, E.; McKeown, K. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv* **2020**, arXiv:2010.03640.
48. Liang, B.; Zhu, Q.; Li, X.; Yang, M.; Gui, L.; He, Y.; Xu, R. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Kerrville, TX, USA, 2022; Volume 1, pp. 81–91.
49. He, Z.; Mokherberian, N.; Lerman, K. Infusing Knowledge from Wikipedia to Enhance Stance Detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Dublin, Ireland*; Barnes, J., De Clercq, O., Barriere, V., Tafreshi, S., Alqahtani, S., Sedoc, J., Klinger, R., Balahur, A., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2022; pp. 71–77. <https://doi.org/10.18653/v1/2022.wassa-1.7>.
50. Luo, Y.; Liu, Z.; Shi, Y.; Li, S.Z.; Zhang, Y. Exploiting Sentiment and Common Sense for Zero-shot Stance Detection. In *Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea*; Calzolari, N., Huang, C.R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.S., Ryu, P.M., Chen, H.H., Donatelli, L., Ji, H., et al., Eds.; International Committee on Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 7112–7123.
51. Beck, T.; Waldis, A.; Gurevych, I. Robust Integration of Contextual Information for Cross-Target Stance Detection. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023), Toronto, ON, Canada*; Palmer, A., Camacho-collados, J., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 494–511. <https://doi.org/10.18653/v1/2023.starsem-1.43>.
52. Li, A.; Liang, B.; Zhao, J.; Zhang, B.; Yang, M.; Xu, R. Stance Detection on Social Media with Background Knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 15703–15717. <https://doi.org/10.18653/v1/2023.emnlp-main.972>.
53. Sun, Q.; Xi, X.; Sun, J.; Wang, Z.; Xu, H. Stance detection with a multi-target adversarial attention network. *Acm Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *22*, 40.
54. Li, Y.; Zhao, C.; Caragea, C. Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic*; Moens, M.F., Huang, X., Specia, L., Yih, S.W.t., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 6332–6345. <https://doi.org/10.18653/v1/2021.emnlp-main.511>.
55. Jiang, Y.; Gao, J.; Shen, H.; Cheng, X. Few-shot stance detection via target-aware prompt distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022*; pp. 837–847.
56. Huang, H.; Zhang, B.; Li, Y.; Zhang, B.; Sun, Y.; Luo, C.; Peng, C. Knowledge-enhanced prompt-tuning for stance detection. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–20.
57. Ollinger, S.; Dumani, L.; Sahitaj, P.; Bergmann, R.; Schenkel, R. Same side stance classification task: Facilitating argument stance classification by fine-tuning a BERT model. *arXiv* **2020**, arXiv:2004.11163.
58. Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv* **2024**, arXiv:2412.13663.
59. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87.
60. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv* **2017**, arXiv:1701.06538.

61. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.* **2022**, *23*, 1–39.
62. Farzam, A.; Shekhar, S.; Mehlhaff, I.; Morucci, M. Multi-Task Learning Improves Performance in Deep Argument Mining Models. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, Bangkok, Thailand; Ajjour, Y., Bar-Haim, R., El Baff, R., Liu, Z., Skitalinskaya, G., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 46–58. <https://doi.org/10.18653/v1/2024.argmining-1.5>.
63. Eger, S.; Daxenberger, J.; Gurevych, I. Neural End-to-End Learning for Computational Argumentation Mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada*; Barzilay, R., Kan, M.Y., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2017; Volume 1, pp. 11–22. <https://doi.org/10.18653/v1/P17-1002>.
64. Tran, N.; Litman, D. Multi-task learning in argument mining for persuasive online discussions. In *Proceedings of the 8th Workshop on Argument Mining*; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 148–153.
65. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning, Virtual*, 13–18 July 2020; pp. 1597–1607.
66. Schiller, B.; Daxenberger, J.; Waldis, A.; Gurevych, I. Diversity Over Size: On the Effect of Sample and Topic Sizes for Topic-Dependent Argument Mining Datasets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 10870–10887.
67. Ajjour, Y.; Wachsmuth, H.; Kiesel, J.; Potthast, M.; Hagen, M.; Stein, B. Data acquisition for argument search: The args. me corpus. In *Proceedings of the Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 48–59.
68. Habernal, I.; Gurevych, I. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Kerrville, TX, USA, 2016; Volume 1, pp. 1589–1599.
69. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Kerrville, TX, USA, 2019; pp. 3982–3992. <https://doi.org/10.18653/v1/D19-1410>.
70. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037.
71. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
72. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the International Conference on Learning Representations, Virtual*, 3–7 May 2021.
73. Jiang, Y.; Gao, J.; Shen, H.; Cheng, X. Zero-shot stance detection via multi-perspective contrastive learning with unlabeled data. *Inf. Process. Manag.* **2023**, *60*, 103361.
74. Yao, Z.; Yang, W.; Wei, F. Enhancing zero-shot stance detection with contrastive and prompt learning. *Entropy* **2024**, *26*, 325.
75. Rocha, G.; Lopes Cardoso, H.; Belouadi, J.; Eger, S. Cross-genre argument mining: Can language models automatically fill in missing discourse markers? *Argum. Comput.* **2025**, *16*, 3–35.
76. Djouvas, C.; Charalampous, A.; Christodoulou, C.J.; Tsapatsoulis, N. Llms are not for everything: A dataset and comparative study on argument strength classification. In *Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics, Athens, Greece*, 13–15 December 2024; pp. 437–443.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.