Argument Detection in Student Essays under Resource Constraints

Omid Kashefi, Sophia Chan, Swapna Somasundaran
Educational Testing Service (ETS)
660 Rosedale Rd, Princeton, NJ, USA
{okashefi,schan,ssomasundaran}@ets.org

Abstract

Learning to make effective arguments is vital for the development of critical-thinking in students and, hence, for their academic and career success. Detecting argument components is crucial for developing systems that assess students' ability to develop arguments. Traditionally, supervised learning has been used for this task, but this requires a large corpus of reliable training examples which are often impractical to obtain for student writing. Large language models have also been shown to be effective few-shot learners, making them suitable for low-resource argument detection. However, concerns such as latency, service reliability, and data privacy might hinder their practical applicability. To address these challenges, we present a low-resource classification approach that combines the intrinsic entailment relationship among the argument elements with a parameter-efficient prompt-tuning strategy. Experimental results demonstrate the effectiveness of our method in reducing the data and computation requirements of training an argument detection model without compromising the prediction accuracy. This suggests the practical applicability of our model across a variety of real-world settings, facilitating broader access to argument classification for researchers spanning various domains and problem scenarios.

1 Introduction

In today's educational landscape, the development of critical thinking and persuasive writing skills holds significant importance. The ability to construct compelling arguments is essential for effective communication and argumentative writing enables students to express ideas clearly, present clear evidence, and address counterarguments effectively. These skills are vital for academic success, professional growth, and civic engagement (Farra et al., 2015; Bertling et al., 2015). Therefore, having a system to analyze and detect argumentation in students' writing would be essential for educators to assess and provide feedback on students' argumentative skills and foster continuous growth in their argumentative writing skills. Furthermore, by using the tool to evaluate their writing, students can identify any weaknesses or gaps in their arguments and make necessary revisions independently. This promotes self-reflection and empowers students to take ownership of their learning, improving their critical thinking and communication skills.

However, the task of detecting arguments within students' essays poses several challenges due to the nuanced nature of argumentation. Constructing an argument involves presenting a *"claim"* and supporting it with "*premises*." However, claims can take various forms, ranging from explicit statements to implicit assertions that require inferential reasoning. Similarly, premises may be stated explicitly or indirectly implied, further complicating the process of argument detection.

Traditional supervised models for argument analysis often rely on large amounts of training data to achieve satisfactory performance. Collecting and annotating such data can be time-consuming and resource-intensive, making it challenging to build large training datasets that cover the diverse range of argumentative patterns and structures present in student essays. Moreover, the practical deployment of large language models such as GPT (Radford et al., 2019; Brown et al., 2020) and PaLM (Chowdhery et al., 2022) can be hindered by cost, latency, and data privacy concerns.

To address these challenges, we introduced an argument classification approach that combines the inherent linguistic characteristics of argumentation with advanced machine learning techniques. We showed the efficacy of exploiting the natural language inference (NLI) relationship between argument components to prime a pre-trained language model for the argument detection task. By merging this with a well-suited prompt-tuning strategy, we

Proceedings of the 10th Workshop on Argument Mining, pages 64–75

established a streamlined architecture that effectively reduces the data and computation requirements of training an argument detection model without compromising the prediction accuracy.

We evaluate the performance and generalizability of our approach across two scenarios: one characterized by availability of reliable training data, and the other representing a resource-constrained noisy domain more akin to real-world settings. In both cases, our approach yielded competitive results, often surpassing the performance of resource-intensive alternatives in classifying argument components. This suggests the practical viability of our model across a variety of real-world settings. We believe that our approach has the potential to make argument classification accessible to a wider range of researchers and problem domains.

2 Argumentation as Entailment

Automated argument detection systems have the potential to help teachers and students by offering a consistent and objective means of evaluating students' work, providing them with timely feedback to enhance their critical thinking and argumentative skills. By automating the process of identifying argument components like claims and premises, educators can redirect their efforts toward other crucial aspects of teaching and providing personalized support to students. However, developing a reliable and accurate automatic system poses certain challenges. Natural language processing algorithms must be sophisticated enough to comprehend the nuances of human language, including various writing styles and levels of proficiency. The system must also recognize context and cultural differences to avoid misinterpretations.

To address these challenges, we propose leveraging semantic relationships between argument elements by framing the argument detection task as natural language inference (NLI). NLI involves discerning the semantic connection between two sentences, where one sentence logically follows (entails) from the other (van Benthem, 2008; MacCartney and Manning, 2009). This notion of entailment and contradiction serves as a foundation for enhancing the semantic representation of various natural language understanding (NLU) problems, including parsing, coreference resolution, and reasoning tasks (Bowman et al., 2015). Similarly, we argue that the NLI framework can be effectively extended to capture the semantic relationships between different components within argumentation. For instance, a counter-claim may contradict the main claim of an argument, or a supportive premise might entail the corresponding claim (Cabrio and Villata, 2013).

We believe that this formulation allows NLP models to leverage their inherent understanding of semantic relationships between logical elements to recognize whether a sentence provides the necessary support or context for a given argument component, and facilitate the development of argument component classification systems, even with a limited volume of training examples. However, employing the entailment paradigm for argument classification requires (a small set of) reliable labeled training data and careful consideration of complex structure of argumentation to ensure accurate and robust results.

3 Proposed Approach

Given that a primary emphasis of this research lies in addressing the challenges posed by resource-limited and noisy conditions in student essay argument detection, we naturally lean towards the utilization of zero-shot/few-shot classification methodologies. In Section 3.1, we discuss how to leverage the inherent structure of zero-shot classification to improve the performance of argument-detection models, and in Section 3.2, we discuss an approach based on efficiently tuning prompts for argument component classification using a small set of training examples.

3.1 Entailment Tuning (ARG-NLI)

Zero-shot classification is a machine learning approach that allows a model to classify instances belonging to classes it has never seen during training. Zero-shot classification in NLP is often approached as an NLI problem, where the goal is to determine the relationship between two sentences: a *premise* (not to be confused with the premise in argumentation) and a *hypothesis*, categorized as "entailment," "contradiction," or "neutral". This framework can be extended to zero-shot classification by casting the classification task as an entailment problem, where the input serves as the premise, and the hypothesis corresponds to a descriptive representation of the target class (Yin et al., 2019).

As we mentioned in Section 2, the relation between argument components can be represented as entailment relations:

- a *premise* **"entails"** the corresponding *claim* ($premise \rightarrow claim$)
- a *claim* **"entails"** the *stance* of the essay ($claim \rightarrow stance$)
- a *counter-claim* **"contradicts"** the *stance* and *claims* of the essay ($counter\text{-}claim \perp stance$)
- *unrelated* argument components are **"neutral"** to each other

We believe further fine-tuning a zero-shot classifier (i.e., a pre-trained transformer-based model trained for NLI task (Bowman et al., 2015; Williams et al., 2018)) on a small set of argumentative training data orchestrated as the entailment task (we refer to this as **ARG-NLI**) would help the model better understand the semantic relationship between different argument components (i.e., between premise and claim, between claim and stance, and between counter-claim and claims/stances). By fine-tuning zero-shot models through ARG-NLI, we anticipate improvements in performance of such models on the task of argument component classification.

3.2 Prompt-Based Tuning (Bart-PEPT)

Large pre-trained language models like GPT (Radford Alec et al., 2018) and BERT (Devlin et al., 2019) have achieved impressive results in NLP benchmarks. However, fine-tuning these models on downstream tasks requires a large dataset of labeled data, which may be a barrier for many NLP tasks. In-context learning is an alternative approach that allows large language models (**LLMs**) to learn new tasks from a few examples, where a single pre-trained model with fixed parameters is shared across all downstream tasks (Radford et al., 2019). This approach works by providing the model with a prompt design for a given task. A prompt is a hand-crafted piece of text that describes or provides examples of the task, usually in natural language. For example, to condition a model for sentiment analysis, one could attach the prompt, "Is the following sentence positive or negative" before the input sequence, "No reason to watch."

Le Scao and Rush (2021) show that a prompt may be worth 100 conventional data points, suggesting that prompts can bring a giant leap in sample efficiency; sharing the same frozen model across tasks also greatly simplifies serving and allows for efficient mixed-task inference. However, task performance can be highly dependent on the prompt design; seemingly trivial changes to the prompt may affect the results. Prompt tuning is an emerging research area that aims to address the limitations posed by manually crafted prompts. Instead of relying on fixed prompts, this approach leverages tunable prompts that are dynamically generated from a small set of training examples. Prompt tuning can improve sample efficiency and enable the seamless integration of mixed tasks, facilitating more effective and versatile inference processes (Schick and Schütze, 2020; Gao et al., 2020; Qin and Eisner, 2021; Zhong et al., 2021; Li and Liang, 2021; Liu et al., 2021; Zhao et al., 2021).

In addition to natural language prompts, LLMs can also be primed by *soft prompts*. These soft prompts are learnable vectors rather than pre-existing vocabulary items (Qin and Eisner, 2021; Zhong et al., 2021; Han et al., 2022). This mechanism allows for end-to-end optimization over a training dataset, and for the prompt to serve as a mechanism for condensing information from large datasets (Lester et al., 2021).

Parameter efficient prompt tuning (**PEPT**) (Lester et al., 2021) is a (soft) prompt tuning approach that focuses on optimizing only a small subset of the model's parameters, specifically the prompt, while keeping the rest of the parameters fixed. PEPT was initially introduced in the context of the T5 model (Raffel et al., 2019) for text-to-text problems. Lester et al. (2021) show that by just tuning the prompt rather than fine-tuning the entire model, T5 can achieve comparable performance on generation and NLU tasks.

Inspired by this, we adapted a version of PEPT to utilize Bart (Lewis et al., 2019) as the core transformer model and made slight modification by incorporating a linear classification head. This model serves as our approach for few-shot classification using smaller language models (**SLMs**). The overarching architecture of PEPT is illustrated in Figure 1. PEPT operates by attaching a tunable vector of numbers to the beginning of the (encoded) input, which functions as the prompt. During the training process, the model parameters are frozen, and gradient updates are only applied to this (soft) prompt vector. Subsequently, the trained prompt is concatenated to the beginning of each input during inference to generate predictions.

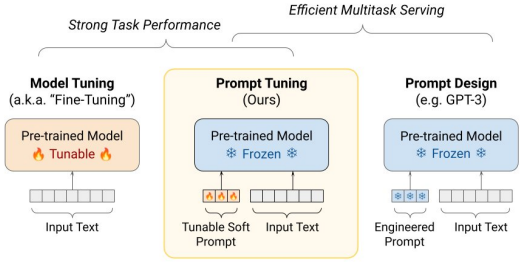

Figure 1: PEPT Model Structure (Lester et al., 2021)

4 Evaluation Methodology

In order to examine the practical viability of our proposed low-resource argument classification approach in a variety of real-world settings, we evaluate the performance and generalizability of our methods in two different scenarios: (i) a problem domain characterized by an abundance of reliable training data (Section 5.1), wherein the availability of the data allows for training traditional *supervised* models, and (ii) a resource-constrained noisy domain more akin to real-world conditions (Section 5.2), wherein *LLMs* as an effective low-resource alternative to supervised training, may seem a more suitable option to approach the problem. While we were able to carry out a small annotation project to collect data for the middle school domain, such annotations may not be feasible, especially if we wish to adapt the system to multiple new domains.

Further information about these problem domains can be found in Section 4.1. The detailed overview of the baseline models we established for both the supervised and zero-shot/few-shot LLM training approaches, as well as the details of our proposed low-resource argument classification methods, are discussed in Section 4.2.

4.1 Problem Domains

4.1.1 Abundance of Reliable Data

In our first set of experiments we use the dataset from Stab and Gurevych (2017), which we refer to as **SG17** in this work. This is a well-known, reliable dataset of argumentation annotations containing essays from "essayforum.com", a site where users submit their academic essays for feedback.

By leveraging this dataset, we can train traditional supervised models as benchmarks for topline performance for the argument classification problem and allows us to assess the comparative effectiveness of our proposed approach against these traditional methods in an ideal scenario where reliable training data is available.

	SG17		ARG	
	train	test	train	test
Claim	1,800	457	64	202
Premise	3,023	809	64	799

Table 1: Total number of samples in each class for sentence-level datasets for Experiment 1 (SG17) and Experiment 2 (ARG).

The statistics of class distribution of examples in SG17 dataset is shown in Table 1. For simplicity, we project the label of the clauses onto sentences and use the dataset at sentence level in all of our experiments. It's important to highlight that there are sentences that contain multiple clauses with different labels (e.g., "*CLAIM because PREMISE*"). However, these cases are comprising only about 2% of the dataset, wherein we assign the label of the minority class to theses sentences to enhance the diversity of the class distribution within our sentence-level dataset.

4.1.2 Limited Noisy Data

A second set of experiments was conducted on an in-house dataset of students' essays, which we refer to as **ARG**. We consider this our low-resource and noisy domain and use these experiments to demonstrate that our approach is suitable for such real-world settings.

This dataset comprises of essays written by students in grades 5 through 9 who reside in the United States. These essays, along with the prompt, were presented to eight annotators as part of the annotation project. Annotators were asked to provide a score along four different persuasive dimensions (*claim*, *counter-claim*, *premise*, and *persuasive strategy*), and to select a text span as the rationale for that score. We consider these *rationales* as our argument components, and used a remapping heuristic to project them to the binary $\{Claim, Premise\}$ classes (see Appendix A for more details).

4.2 Argument Classification Models

In this section, we present the technical details of our proposed low-resource argument classification approaches. Furthermore, we outline the supervised, SLM, and LLM-based baselines that we have established as alternative methods.

4.2.1 Supervised Models

We establish two supervised argument classifier baselines as follows:

Bert-Sequence We use the HuggingFace (Wolf et al., 2019) `bert-base-uncased` (Devlin et al., 2018) model with a classification head to predict whether a sentence is either a `Claim`, or a `Premise`.

Bert-BIO We adopt the model architecture introduced by Alhindi and Ghosh (2021), which employs a BIO classification scheme to identify and classify argument components. We use `bert-base-uncased` as the base transformer model and train a token-level classifier head on top. This baseline aims to label each token as `B-claim`, `I-claim`, `B-premise`, `I-premise`, or `O`. For consistency in our evaluation, we incorporate a label projection heuristic to map BIO prediction to sentence-level labels, as discussed in Appendix B.

4.2.2 Large Language Models

To establish our LLM-based baselines, we utilize the OpenAI GPT-3 models in zero-shot and few-shot settings. In the zero-shot configuration, the model relies solely on its pretrained knowledge without any task-specific fine-tuning. In the latter setup, we provide the model with a limited amount of task-specific examples to adapt it to our argument detection task. It's also important to note that at the time of conducting this study, the newer GPT-4 model was not publicly accessible, restricting our experiments to the utilization of the GPT-3 version.

GPT3:Zero-shot We use `text-davinci-001` via the OpenAI Completion endpoint[1] with the following prompt:

```
Classify the text as
{claim_label} or {premise_label}.
Text: {sentence}
Label:
```

For each sentence in the test set, we replace the placeholder in the prompt with that sentence and feed it to the completion endpoint. We experiment with a couple different values for `claim_label` ($\{Claim|Idea\}$) and `premise_label` ($\{Premise|Support\}$) due to a trait of generative models that "causes probability to be rationed between different valid strings, even ones that differ trivially" (Holtzman et al., 2021). We then pick and report the result of the combination that performs best within each experiment and problem domain.

GPT3:Fine-tuned The extensive pretrained knowledge of LLMS enables them to adapt efficiently to specific tasks or domains, even with a relatively small number of training examples, making them a potentially suitable low-resource baseline for argument detection tasks. Accordingly, we fine-tuned a GPT3-DaVinci model via the OpenAI endpoint using 64 randomly sampled sentences from each class and obtained predictions from the completions endpoint.

4.2.3 Smaller Language Models

Bart-MNLI:Zero-shot We use the HuggingFace port of `facebook/bart-large-mnli` out of the box as our zero-shot baseline. This is a checkpoint for the Bart-large model (Lewis et al., 2019) after training on the MultiNLI (MNLI) dataset (Williams et al., 2018). Similar to the `GPT3:Zero-shot` baseline, we used a simple prompt template of:

```
This sentence is {label}
```

Again, we experiment with a couple different values for `claim_label` ($\{Claim|Idea\}$) and `premise_label` ($\{Premise|Support\}$). Our experiments revealed that employing the labels $\{Idea|Support\}$ yielded the the most promising and robust results, so we present and discuss the results of this label configuration in this study.

Adjustment for Bias. The language models, including our `Bart-MNLI:Zero-shot` baseline, may exhibit biases towards certain values within the answer space. For example, there could be an imbalance in the training data, resulting in a higher likelihood of predicting certain answers, such as "positive", over others like "negative".

To address this issue of prompting bias, we implemented a threshold adjustment strategy as suggested by Sun et al. (2022). We initiated this process by determining the probability of an empty input ($x =$ "") being classified as "claim" by querying the model with the prompt:

$$[x] \texttt{ is an idea}$$

This probability value serves as the basis for establishing the threshold used to categorize inputs as claims. For instance, if the probability of being claim for the empty input be 0.63, any input with

[1] https://openai.com/blog/openai-api

a probability of lower than 0.63 would no longer be classified as a claim, whereas any value above 0.5 would have been categorized as such prior to the bias adjustment. This strategy has the potential to enhance the fairness, accuracy, and reliability of our zero-shot baselines, making them more equitable and dependable classifiers.

4.2.4 Our Proposed Models

ARG-NLI In order to investigate the effectiveness of using the entailment formulation of argument classification problem as we proposed in Section 3.1, we randomly picked a few essays from the training datasets and created the entailment pairs for the premises and related claims, and claims and major claims. The SG17 dataset (Section 4.1.1) contains relation annotations in the form of (source, target) tuples, where the source claim/premise either supports or attacks the target claim/premise. An attacking claim is also known as a counter-claim. In addition to claims and premises, major claims that express the writer's stance towards the prompt are also annotated. We used this information to create the NLI representation of argumentative annotation of claim and premises in SG17, as follows:

- claims **entail** major claims in the same essay
- premises **entail** their related claim
- counter-claims **contradict** their related major claim
- premises of an essay are **neutral** towards the claims of other essays

Our in-house ARG dataset (Section 4.1.2) does not have the relation annotations so we used a simple heuristic to relate the argument components:

- claims within an essay **entail** one another
- premises **entail** claims within the same paragraph
- counter-claims **contradict** all the claims in the same essay
- premises of an essay are **neutral** towards the claims of other essays

After creating the NLI representation of argumentation datasets (pair of sentences with appropriate entailment label), we use them to fine-tune the same `Bart:MNLI:Zero-shot` we used in Section 4.2.3. We then used the fine-tuned model in zero-shot classification fashion– feed in a *single* sentence and prompt the model to determine whether the input sentence is a `claim`, or a `premise`?

Bart-PEPT As mentioned in Section 3.2, we developed a modified version of the model introduced by Lester et al. (2021) to operate on `facebook/bart-large-mnli` of HuggingFace for "classification" tasks as our approach for few-shot classification using SLMs.

ARG-NLI + Bart-PEPT This variation of Bart-PEPT uses the argument-NLI finetuned version of the Bart we developed (a.k.a, `ARG-NLI`) as the core transformer model; a prompt is then tuned on top of this base model.

5 Experiments

5.1 Exp. 1: Large Reliable Training Data

In this experiment, we use the SG17 dataset described in Section 4.1.1 to evaluate our model in a scenario where a large corpus of reliable training data with argument annotation is available.

The anticipation is that supervised models will excel in the task of *distinguishing between "claim" and "premise" sentences* within this context. Therefore, our main objective of this is *to explore the comparative capabilities of our proposed low-resource alternative models in relation to the well-established supervised training paradigm.*

We trained all argument classifier models on the SG17 train set described in Table 1. The `Bert-Sequence`, and `Bert-BIO` baselines are trained on the entire training set of the SG17, which consists of 4.8K sentences with 115K tokens. The zero-shot baselines (`GPT3:Zero-shot` and `Bart-MNLI:Zero-shot`) are not exposed to any training examples. The `GPT3:Finetuned` and `Bart-PEPT` models are trained with 64 claim examples and 64 premise examples from the training set. For entailment tuning for `ARG-NLI` model, we randomly picked 20 essays from the train set and created the argument component pairs of "entailment" and "contradiction" examples.

Overall, we fine-tune the Bart-MNLI model with 700 argumentative entailment examples and evaluated that as a zero-shot classifier on the test set of SG17. For more details on our argumentative entailment dataset please refer to Appendix D.

Model	SG17	ARG
Supervised		
Bert-Sequence	72%	66%
Bert-BIO	69%	62%
L(arge)LM		
GPT3:Zero-shot	61%	56%
GPT3:Finetuned	66%	62%
S(mall)LM		
Bart-MNLI:Zero-shot	52%	51%
Our approach		
ARG-NLI	61%	59%
Bart-PEPT	70%	72%
ARG-NLI + Bart-PEPT	**73%**	**77%**

Table 2: Macro-F1 scores for argument classification across various models and training paradigms for Experiment 1 (SG17) and Experiment 2 (ARG). The bold-faced numbers indicate the best performing models.

5.1.1 Results

Table 2 shows the macro-F1 score of our models in classifying 1.3K argument-related sentences of the SG17 test set as either "claim", or "premise". As expected, both supervised baselines are capable of reliably predicting the correct label for the argument components within this dataset, with the sequence classifier baseline (F1 = 72%) performs better than the token classifier baseline (F1 = 69%).

Both zero-shot baselines yield sub-par performance compared to their counterparts. Also in line with our expectations, the LLM-based baseline outperformed the SLM-based baseline (61% versus 52%). These results highlight the challenging nature of argument classification, indicating that distinguishing between claims and premises involves subtleties beyond what can be achieved through simply prompting pre-trained transformers. Incorporating argument entailment tuning (`ARG-NLI`) leads to a substantive 9% enhancement over the SLM zero-shot baseline (61% vs. 52%), indicating that priming models with the entailment relationship between argument components can make them better zero-shot learners for the task.

Fine-tuning LLM on the task with 128 training examples led to a 6% performance increase compared to the baseline achieved by the zero-shot LLM. However, with the same number of training examples, our `Bart-PEPT` approach achieved a remarkable F1 performance of 70%, trails the best-performing supervised alternative by only 2%, even though the latter is trained on a corpus over 35 times larger. Furthermore, once we combined our argument NLI fine-tuned model with PEPT (`ARG-NLI + Bart-PEPT`), we achieve a substantial 21% improvement over the SLM zero-shot baseline and 3% over our Bart-PEPT model. This model surpasses the top-performing supervised model in terms of F1 performance, despite using only a fraction of the training data.

5.2 Exp. 2: Limited Noisy Training Data

In this experiment, we evaluate our model on the ARG dataset (described in Section 4.1.2), a scenario more akin to real-world conditions, wherein a large corpus of reliable training data is not available. In this setting, we annotate about 1.2K argumentative sentences of student's essays. We used 128 of these examples for training the models and held-out the remainders for testing.

Since there is not enough data to train a robust supervised model, we anticipate that traditional supervised models will fail to accurately distinguish between "claim" and "premise" sentences in this experiment. Therefore, this experiment would help us to assess the applicability of our proposed low-resource argument classifier approaches as an alternative to data and resource intensive supervised and LLM baselines.

We trained the `Bert-Sequence`, `GPT3:Finetuned` and `Bart-PEPT` models on the 64 claim examples and 64 premise examples of our in-house argumentative student writing dataset ARG. The `Bert-BIO` baseline a variation of ARG dataset with token-level annotation, containing 68 claim and 96 premise entities. Appendix C presents the BIO statistics of ARG dataset. The zero-shot baselines (`GPT3:Zero-shot` and `Bart-MNLI:Zero-shot`) are not exposed to any training examples. In addition, we used the ARG training essays to create 700 argument component pairs with entailment labels. We then leverage this dataset to finetune our proposed `ARG-NLI` model.

5.2.1 Results

The numbers under the ARG column of Table 2 are showing the macro-F1 score of different models in classifying 1K argument-related sentences of our ARG test set as either "claim" or "premise".

Although both the sequence and BIO supervised classifier baselines are still performing in a reasonable range (62% and 66%, respectively), we observe a noticeable drop (5% on average) in performance compared to the previous experiment,

which was conducted on a larger training dataset. These outcomes corroborate that supervised approaches rely heavily on access to high-quality training data, a requirement that does not consistently align with resources available for various real-world NLP problems.

Consistent with previous experiments, zero-shot baselines continue to show relatively poor performance on this dataset (51% and 56% for SLM and LLM zero-shot baselines respectively). This outcome, however, is inline with expectations, as these baselines are not trained with examples from the target domain. Our proposed argumentative entailment fine-tuning approach (`ARG-NLI`) exhibits an 8% improvement over the SLM zero-shot baseline (59% vs. 51%). These consistent observations from both of our experiments demonstrates the effectiveness of pre-training (smaller) foundational models with the inherent entailment structure of argument elements. This approach helps models comprehend the semantic structure of argumentation more thoroughly, leading to improved performance as zero-shot learners for the task.

Fine-tuning LLMs with domain-specific training data shows certain performance enhancements compared to their zero-shot counterparts (62% vs. 56%). However, similar to the previous experiment, these improvements remain limited. Despite the relatively high costs associated with using LLMs, their performance as a low-resource solution still falls short of being viable for production deployment in the argument classification task.

As shown in Table 2, our prompt-based tuning approach `Bart-PEPT` outperformed all other methods in this low-resource setting (F1=72%). Moreover, once it uses our `ARG-NLI` model as the core foundation models, we observe an additional 5% performance boost. These outcomes underscore the suitability of our proposed approach as a reliable and accurate method for argument classification in low-resource domains. Our approach achieves results on par with data and resource-intensive supervised and LLM alternatives within resource-abundant contexts, while outperforming them in problem domains lacking such extensive training corpora. This positions our approach as a versatile choice for a broader range of problems.

5.3 Latency Analysis

While LLMs can yield reasonable results with a small number of training examples, fine-tuning them demands extensive parameter updates, consuming substantial time and computation. For instance, the fine-tuning of the GPT-3 "davinci" model entails updating over 170B parameters, whereas our Bart-PEPT model requires modifying only 40K parameters within the prompt (the model parameters frozen). This raises practical concerns regarding the latency when working with these models. Therefore, we conducted a comparison of inference latency among the methods discussed in this study, as shown in Table 3.

Models	Latency (ms)
Bert-Sequence	0.66
Bert-BIO	22.46
GPT3:Finetuned	19.74
Bart-PEPT	7.6

Table 3: Average inference time of selected argument classifier models.

Latency measurements were conducted on the ARG test set, comprising 1K sentences with an average of 17 tokens per sentence. For transformer-based models, we use a single Tesla K20Xm GPU with 22.5 GiB of RAM and a batch size of 32. For GPT-3 we batched up to 20 requests, the current maximum allowed by the completion endpoint.

6 Conclusion

In this work, we introduced an argument classification strategy that effectively leverages the logical entailment relationship within argument components, along with a parameter efficient prompt-tuning technique. Our approach demonstrates remarkable efficiency in reducing data and computational requirements for training while maintaining high prediction accuracy. Its robust performance across diverse scenarios highlights its practical applicability in real-world settings, making argument classification more accessible to researchers across various domains. Notably, the model's ability to achieve competence with a minimal number of examples per class sets it apart from traditional data-intensive supervised alternatives.

Additionally, unlike expensive and time-intensive LLM-based solutions, our proposed approach can reliably operate on smaller foundation models such as Bart, offering expedited training and inference, making it a cost-effective and efficient solution suitable for in-house deployment and enjoying the added benefits of data privacy.

Limitation

The focus of this study lies in argument component classification. A more practical application would entail a pipeline system that initially distinguishes argumentative sentences from non-arguments—potentially through a separate predictive model. Then, our approach in this study could offer fine-grained insights into the usage and developmental stages of argumentation within student writing at the claim and premise levels. It is also important to note that while our method streamlines requirements, it still requires a small amount of data for model tuning.

As future work, we intend to expand our efforts towards multi-class prediction, incorporating the "*none-argument*" category as a potential label. This expansion necessitates re-annotating our in-house dataset using an argumentative annotation scheme, as we suspect that rationale-based annotation schemes tend to classify argumentative elements as non-arguments, inviting the need for a more specific annotation guideline.

Furthermore, our company's data privacy policy prohibits us from publicly releasing student-written essays. Unfortunately, we are unable to make our in-house argument dataset (ARG) mentioned in this work available to the public.

Ethics Statement

While we strive to contribute positively to the field of argument detection, we are fully aware of the ethical dimensions and potential challenges associated with deployment of AI models, particularly in education domain. We recognize the potential for representational harm (Suresh and Guttag, 2021), which is complex and often challenging to quantify. Biases can emerge from multiple sources, including annotators, system designers, and the data itself, and it can shape how claims, premises, and arguments are defined and interpreted (Gaskins, 2023). Despite our efforts to source a diverse range of student essays and annotators, biases within the data are possible. We are also aware of well-documented biases in language models like Bert and GPT (Monarch and Morrison, 2020). These biases could inadvertently manifest in our system's output, potentially perpetuating and amplifying inequalities.

To mitigate these risks, we have taken several steps. Our primary intention is to assist students in becoming better writers and reduce the burden on teachers, fostering formative assessment. We require teacher approval before presenting feedback to students, thereby minimizing representational harm by ensuring that feedback aligns with educational objectives. Additionally, we commit to avoiding the use of our system in high-stakes testing or consequential decisions, thereby reducing allocational harm. We remain committed to continuous evaluation, refinement, and transparent communication of the ethical considerations in our work, with the ultimate goal of fostering responsible and equitable AI adoption in education.

References

Tariq Alhindi and Debanjan Ghosh. 2021. "Sharks are not the threat humans are": Argument Component Segmentation in School Student Essays. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 210–222, Online. Association for Computational Linguistics.

Maria Bertling, G. Tanner Jackson, Andreas Oranje, and V. Elizabeth Owen. 2015. Measuring argumentation skills with game-based assessments: Evidence for incremental validity and learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9112:545–549.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mccandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Elena Cabrio and Serena Villata. 2013. Detecting Bipolar Semantic Relations among Natural Language Arguments with Textual Entailment: a Study. In *Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 24–32.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton,

Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *Computing Research Repository, ArXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Computing Research Repository*.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring Persuasive Essays Using Opinions and their Targets. *10th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2015 at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015*, pages 64–74.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making Pre-trained Language Models Better Few-shot Learners. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3816–3830.

Nettrice Gaskins. 2023. Interrogating Algorithmic Bias: From Speculative Fiction to Liberatory Design. *TechTrends : for leaders in education & training*, 67(3):417–425.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2022. PTR: Prompt Tuning with Rules for Text Classification. *AI Open*, 3:182–192.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Teven Le Scao and Alexander M. Rush. 2021. How Many Data Points is a Prompt Worth? *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 2627–2636.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 4582–4597.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT Understands, Too. *Computing Research Repository*.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Eight International Conference on Computational Semantics*, pages 140–156.

Robert Monarch and Alex Morrison. 2020. Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017.

Guanghui Qin and Jason Eisner. 2021. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5203–5212.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Radford Alec, Narasimhan Karthik, Salimans Tim, and Sutskever Ilya. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Timo Schick and Hinrich Schütze. 2020. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2339–2352.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Yi Sun, Yu Zheng, Chao Hao, and Hangping Qiu. 2022. NSP-BERT: A Prompt-based Few-Shot Learner through an Original Pre-training Task —— Next Sentence Prediction. In *COLING*, pages 3233–3250.

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *ACM International Conference Proceeding Series*.

Johan van Benthem. 2008. A Brief History of Natural Logic. Technical report, ILLC Amsterdam.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3914–3923.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *ICML*.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5017–5033.

A ARG Annotation

We conducted the annotation study on the Inception platform (Klie et al., 2018). In total, eight annotators double-annotated 300 essays after completing a calibration exercise that involved annotating 30 essays. Annotators gave each essay a score along four persuasive dimensions (*claim*, *counter-claim*, *premise*, and *persuasive strategy*). For each dimension, the annotators selected text spans that served as the rationale or explanation for their score, and we take these spans to be our argument components. A span was counted if it was selected by any annotator, and spans were combined when more than 10% tokens overlap.

After the annotation was completed, one of the authors examined ten essays and created rules to map rationale labels to the binary $\{Claim, Premise\}$ classes. In addition, based on our review of the data, we decided to only count double-annotated premise spans and remove any essays that contain no claims. The rules for remapping are as follows:

- *claim* → *claim*
- *counter-claim* → *premise*
- *claim*, *premise* → *claim*
- *persuasive strategy* → discard

B BIO Label Projection

In the `Bert-Sequence` baseline every sentence receives only one label (either claim or premise), while the BIO baseline can predict different segments of the sentence as different argument components. To ensure a uniform sentence-level prediction scheme across baselines, we incorporate a label projection policy as follows:

- when all predicted argument components within a sentence are classified as the same class, we project that prediction to the entire sentence

- if a sentence contains argument components with different classes, we label the sentence with the label of the minority class (in our experiments, the "claim" class)

C Token-Level Annotation

Table 4 shows the token-level class distribution of the SG17 and ARG examples, used to train the supervised token classifier baseline.

To make the token-level dataset for our low-resource ARG examples comparable to the sentence-level dataset described in Table 1, we included a similar amount of claims and premises. The sentence-level dataset contains 64 claims and 64 premises, while the BIO dataset contains 68 claim and 96 premises entities. For both SG17 and ARG, we excluded `O` spans from the test set, as only claims and premises are included in the sentence-level experiments.

Label	**SG17**		**ARG**	
	train	test	train	test
B-claim	1.8k	573	62	527
I-claim	25k	6.2k	1.1k	3.8k
B-premise	3k	833	87	585
I-premise	50k	10.6k	1.8k	7.5k
O	35k	-	1.4k	-

Table 4: Total number of samples in each class for BIO datasets for Experiment 1 (SG17) and Experiment 2 (ARG).

D Entailment Argument Dataset

Table 5 shows the class distribution of the 700 NLI examples we created from SG17 and ARG datasets, used to train our ARG-NLI fine-tuned zero-shot model.

Label	**SG17**		**ARG**	
	train	dev	train	dev
Entails	263	56	225	56
Contradicts	17	4	27	7
Neutral	280	60	308	77
Total	560	140	560	140
	700		700	

Table 5: Total number of argument entailment samples in each class for Experiment 1 (SG17) and Experiment 2 (ARG).

E Hyperparameters

We used the default settings of HuggingFace transformers and OpenAI for most of the parameters except the following:

- Bert-Sequence
  - eps=1e-8
  - lr = 2e-5
  - max_length = 256
- Bert-BIO
  - lr = 5e-5
  - max_seq_length = 512
- LLM zero-shot
  - temperature = 0
  - top_p = 1
  - max_tokens = 16
- LLM fine-tuned
  - temperature = 0
  - top_p = 1
  - max_tokens = 2
- Our approach (Bart-PEPT)
  - model_max_length = 1024
  - prompt length = 20 tokens
  - lr = 2e-5 (significantly different from the value Lester et al. (2021) used)