Barcelona School of Economics

# Classification of Climate-Change-related claims

Aleksander Smolin

Pablo Fernández

Blanca Jimenez

June 13, 2025

# Contents

# 1 Motivation

The dataset for this task consists on a set of claim-evidence pairs where each claim is linked to a climate change topic[1]. Each claim-evidence pair is labeled according to whether they are related (1) or not (0).

For instance, a related claim-evidence pair is the following:

- **Claim**: "Global sea level rise surged between November 2014 and February 2016, with the El Nino event helping the oceans rise by 15mm."
- **Evidence**: "As El Nino conditions started to develop during early 2014, sea levels in western Micronesia including in waters surrounding the island nations of Palau and Guam dropped by 6–9 feet ( 1.8–2.7 m )."

While an unrelated pair of claim-evidence is:

- **Claim**: "The climate-change agreement between the United States and China "requires the Chinese to do nothing at all for 16 years.""
- **Evidence**: "According to NASA, the most widespread Antarctic surface melting of the past 30 years occurred in 2005, when an area of ice comparable in size to California briefly melted and refroze; this may have resulted from temperatures rising to as high as 5 °C (41 °F)."

Climate change has become an increasingly political issue, with rising demands for action from citizens and rising concerns about its impacts on global production chains and on the lifelihoods of many. Simultaneously, politics has, in the last years, become increasingly sensationalist with the appearance, around the globe, of populist figures and parties that have come to be at the center of the political stage. This, together with social media, has paved a perfect path for fake news and myths void of scientific evidence with climate change being one of its main victims.

Disinformation travels fast so we need a just-as-fast and accessible solution to detect them, as it is a first step to fight it. Meaning tools to detect causal links and find evidence behind widespread claims are urgent.

In this project we explore different natural language processing paradigms - zero-shot learning (ZSL) and few-shot-learning (FSL) - for the task of determining whether a claim is related to evidence. Our goal is to find in LLMs a tool that allows us to fight disinformation, rather than promote it.

To that end, we try out and test the performance of 6 different models:

- Three models for zero-shot learning with natural language inference (NLI):
  - Two models where the claim-evidence pairs are passed as inputs, one where we use a custom NLI template[2] and another where we use the default one.
  - One model where we only use the claim as input for ZSL, to assess whether just by the wording of the claim we are able to detect whether it relies on actual evidence or not.

[1] The dataset description can be checked here: https://huggingface.co/datasets/mwong/climate-claim-related

[2] This claim is label to the evidence.

- Four models for few-shot learning:
  - Two models where we use contrastive learning (SetFit), with 16 and 32 labelled examples and a generic pre-trained language model (`all-MiniLM-L6-v2`).
  - One model with standard fine-tuning, but using a pre-trained language model that was pre-trained on factual data similar to our downstream task: `albert-xlarge-vitaminc-mnli`. We fine-tune by freezing the transformer layers and only fine-tuning the classification head.

**All of the results and performance metrics shown below are calculated on (unseen) test data.**

# 2 Models

Explain with higher detail the models that have been used.

## 2.1 Zero-Shot Learning models

### 2.1.1 NLI with default hypothesis template

Our first ZSL model is a Natural Langugage Inference model with its default template for the hypothesis construction. We set it up by concatenating each claim to its corresponding hypothesis. So a string such as:

- " Claim: A cold day in Chicago in winter has nothing to do with the trend of global warming.
  nEvidence: 'A novel probabilistic forecast system predicting anomalously warm 2018–2022 reinforcing the long-term global warming trend' "

The candidate labels are "related" and "not related". So, internally, the Zero-Shot Classifier constructs as possible hypothesis: "This text is about *label*."

As is expected, this classifier performs very poorly, with evaluation metrics practically equal to those of a random classifier. Below are the confusion matrix, ROC curve (area under the curve = 0.50), and Precision-Recall curve (average precision = 0.52). We can see that all texts get classified as related except two of them which, surprisingly, get correctly classified as unrelated.

This poor performance can partly be explained by the absence of logical coherence of the hypothesis when using our labels with the default template. However, we will see below that the ZS-NLI classifier also performs very poorly, which leads us to attribute this to either the internal knowledge of the language model regarding, on the one hand, climate change, and on the other hand, claim-evidence relations – as this goes beyond NLI-trained models' "expertise" of establishing connections across topics and logical paths; or either the quality of our data, as these statements are very short and often very out of context.

### 2.1.2 NLI with custom hypothesis template

This model works just as the one above, except that this time we customize the hypothesis template text, which becomes: "This claim is *label* to the evidence.". It also returns performance close to that of a random classifier with 50% accuracy, as it classifies **all** pairs as related.

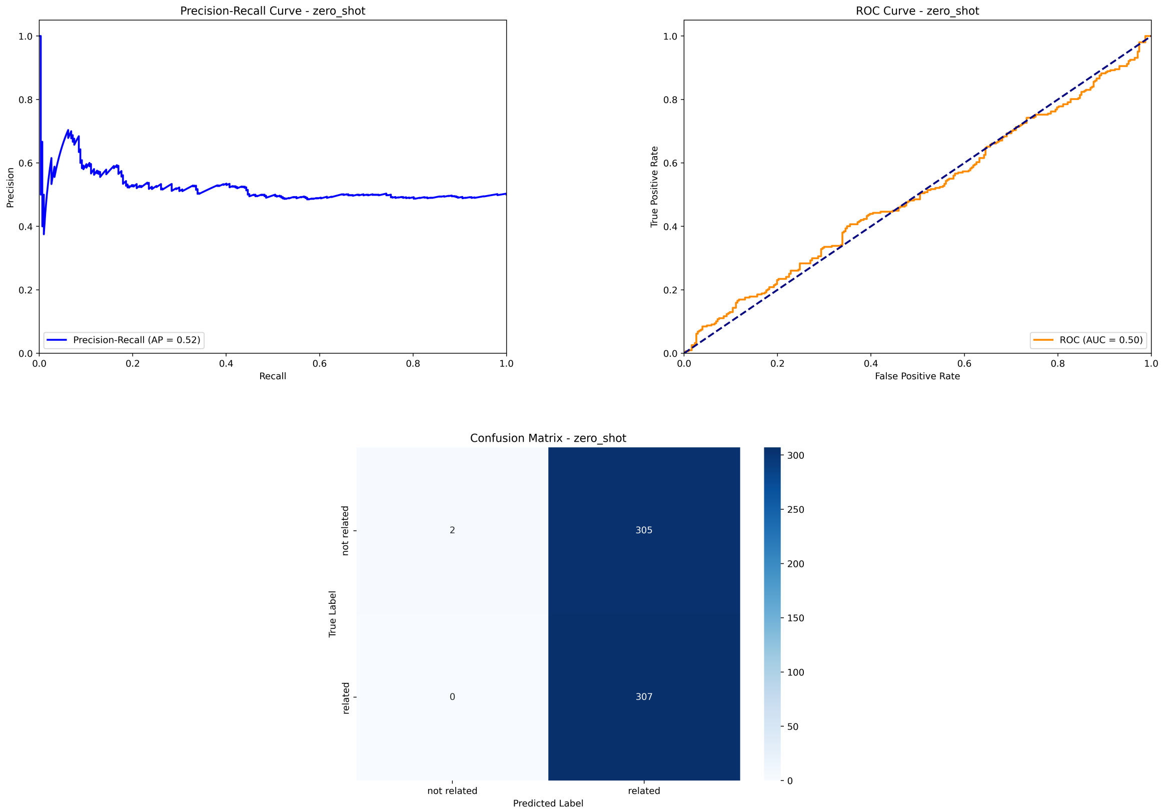

**Figure 1:** Metrics for zero-shot learning with the default hypothesis-entailment template.

As the model above, we can attribute bad performance to lack of task and domain-specific training of the language model and/or bad quality of the data. Below are the evaluation visualizations.

### 2.1.3 NLI with only the claim (and default template)

Finally, we decide to try to pass as text only the claim, without the evidence part. Our reasoning is that we hope it will be able to detect whether claims, by themselves, are evidence-based or not.

Once more, the performance is close to random. We can expect that as, again, this is not a model that is fine-tuned on knowledge on this area and performing well in this would require to have substantial background knowledge about climate change. Moreover, we use the default template once again, so internally the model is determining whether the claim "is about ’related’ / ’not related’.

## 2.2 Few-Shot Learning

### 2.2.1 Contrastive learning with 16 examples

Here we use the SetFit methodology (and library) to few-shot train a model using contrastive learning. SetFit takes *all-MiniLM-L6-v2*, a pre-trained sentence transformer, and fine-tunes it using a small set of examples. The way it is trained is by creating pairs among these examples which are labeled as belonging to the same class or not. The transformer is then fine-tuned to make embeddings of same-class pairs closer to each other and vice-versa.

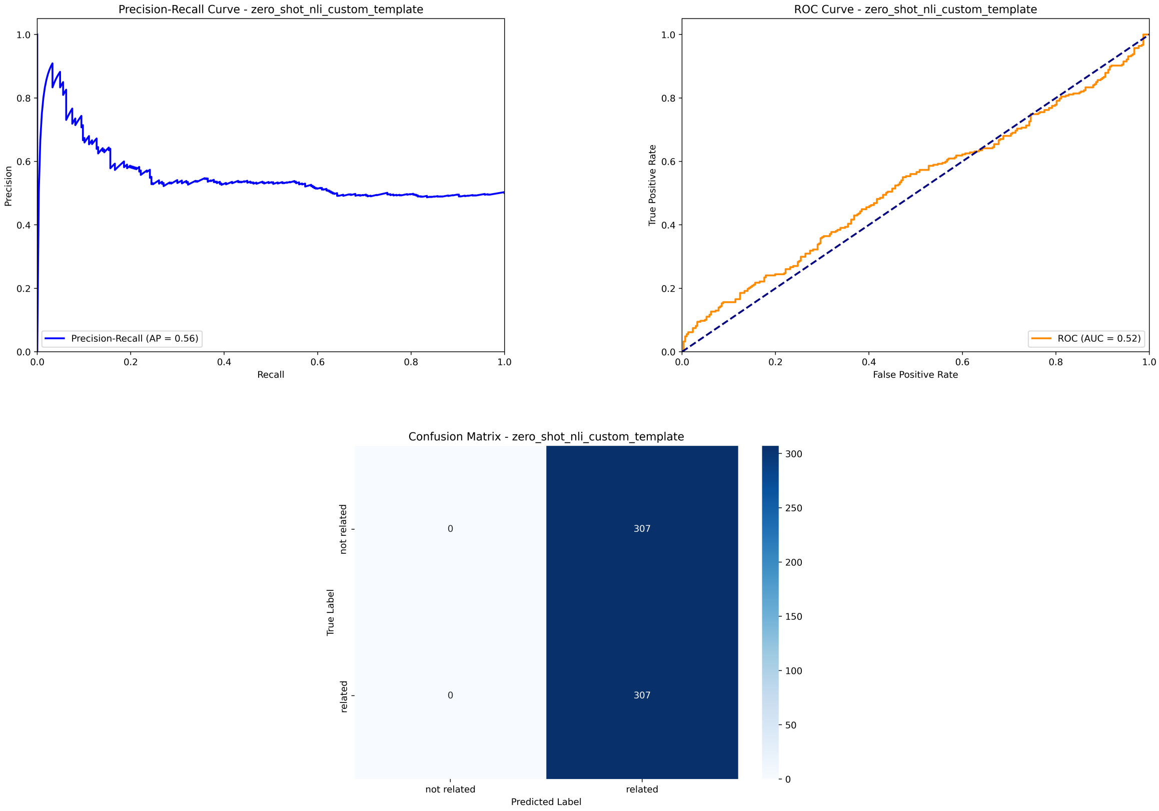

**Figure 2:** Metrics for zero-shot learning with the custom hypothesis-entailment template.

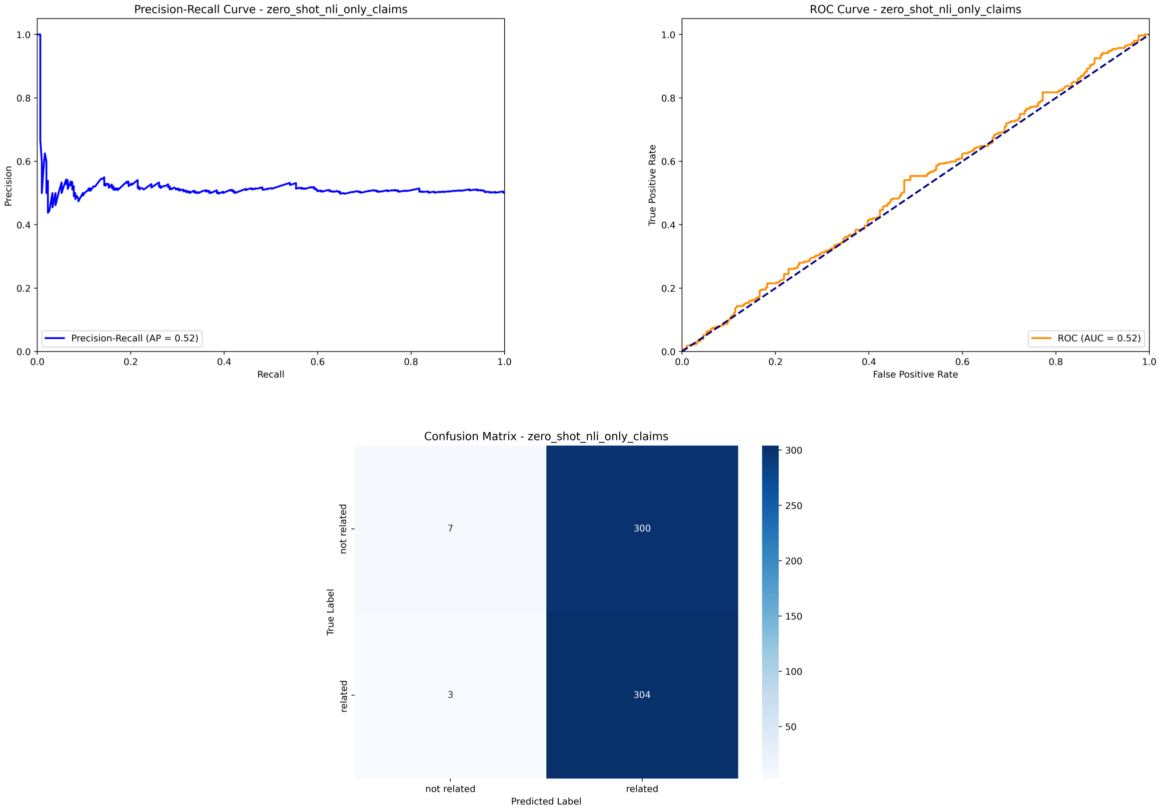

**Figure 3:** Metrics for zero-shot learning with the climate-related claim as the only input.

In our application, we do 10 iterations of this process in which a new instance of the transformer is trained each time. The goal of doing this is to make sure that the outcoming performance metrics are not due to a particularly good/bad sample. At the end of the process, we compare the metrics from all 10 train models.

We can see in our 16-examples set that performance is just above that of random classification. All 10 models exhibit accuracy of around 0.52-0.53, indicating that we *can* consider these results better than random. Below are the evaluation metrics.

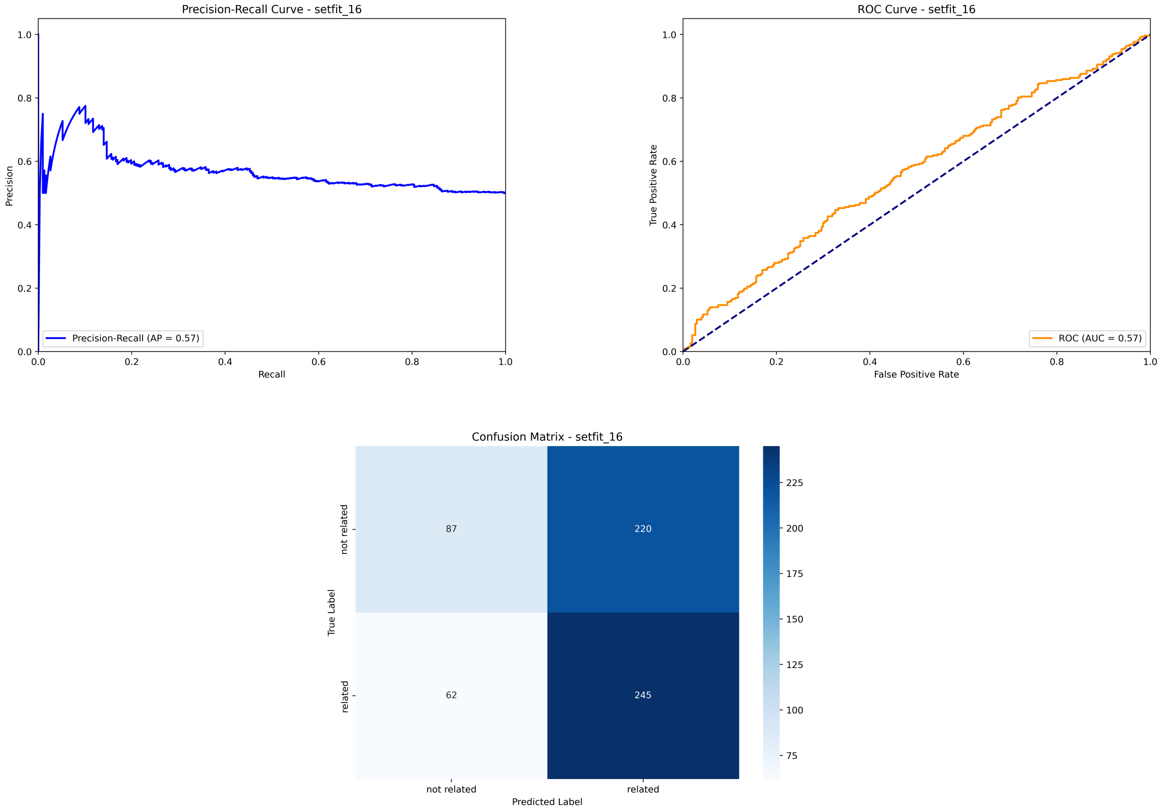

**Figure 4:** Metrics for few-shot learning with contrastive learning (SetFit) and 16 labelled instances.

### 2.2.2 Contrastive learning with 32 examples

This model works exactly like the one above, but with 32 examples.

The performance increases slightly, bringing accuracy to around 0.56. This reinforces our hypothesis that the performance of this model is not due to randomness.

### 2.2.3 Fine-tuning a fact-checking model with 16 examples

In this setup, we fine-tune a pre-trained transformer model for binary classification using only 16 labeled examples. The base model is `tals/albert-xlarge-vitaminc-mnli`, which is a variant of ALBERT that has been trained specifically for tasks related to fact-checking and natural language inference (NLI). Its pre-training includes exposure to datasets like VitaminC and MNLI, which are designed to help models reason about relationships between claims and evidence — exactly the structure of our task.

Rather than updating the entire model, we freeze the large pre-trained transformer and only fine-tune the classification head. This approach, common in few-shot learning, helps preserve

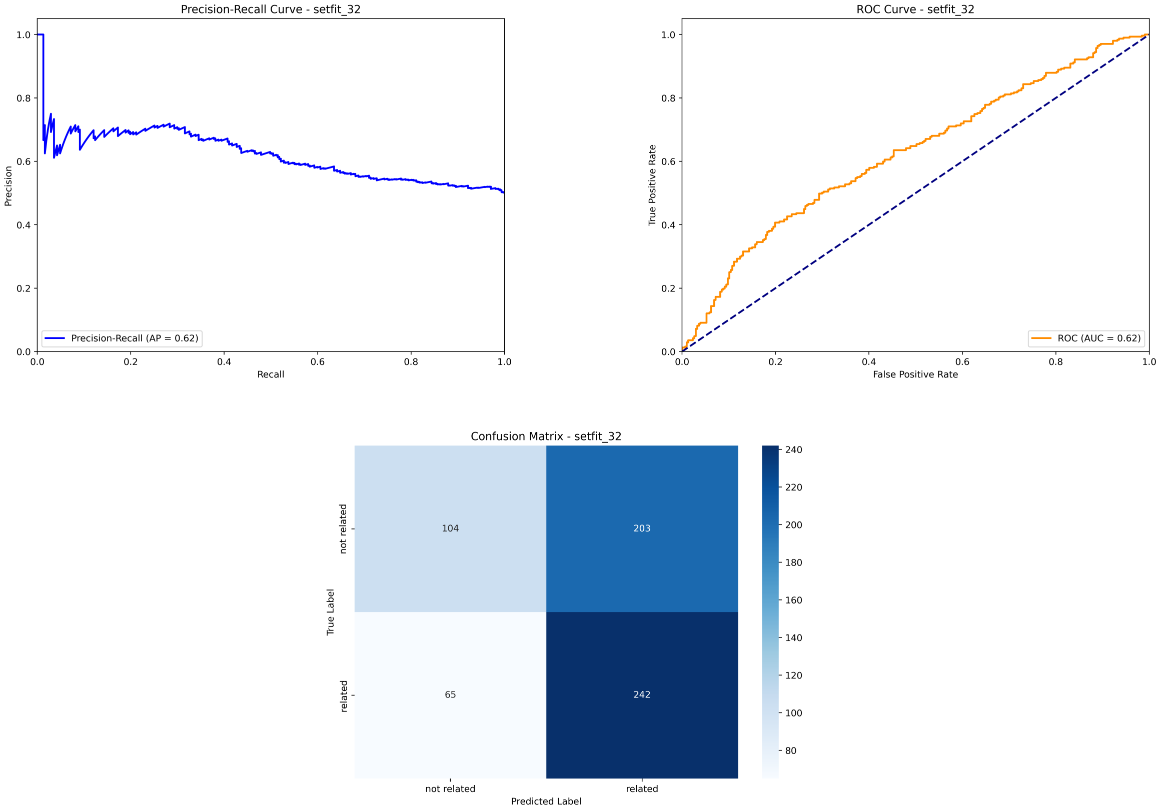

**Figure 5:** Metrics for few-shot learning with contrastive learning (SetFit) and 32 labelled instances.

the model's general understanding of language and reasoning, while avoiding overfitting to our tiny training set. The classification head is therefore trained to make predictions on our two labels.

Again, we train it over 10 epochs and save the checkpoint with the lowest validation loss.

Compared to the contrastive learning approach, the results here are slightly stronger. We observe an accuracy of 0.53, with an area under the ROC curve of 0.61 and an average precision of 0.61. Although performance is still quite low, these metrics suggest an improvement indicating that the fact-checking pre-training truly sets up the model to better understand relationships between our claim-evidence pairs.

## 3 Discussion

In our experiments, the zero-shot NLI models—both with default and custom hypothesis templates demonstrated being extremely limited for our task-specific reasoning. While the requirement of no additional labeled data make these models widely accessible and therefore a great option, their performance showed now improvement over random classification (0.50 accuracy, AUC 0.50, AP 0.52) and they overwhelmingly predicted "related." This reflects two key weaknesses: first, generic NLI pretraining (even on broad-domain datasets like MNLI) does not equip the model with nuanced understanding of climate-claim evidence relations, and second, our hypotheses ("This text is about X" or "This claim is X to the evidence") were too abstract to guide the model's reasoning. We believe that ZSL for this task could be improved by leveraging stronger, domain-tuned NLI backbones, or by providing additional context (e.g., longer claims) to reduce ambiguity.

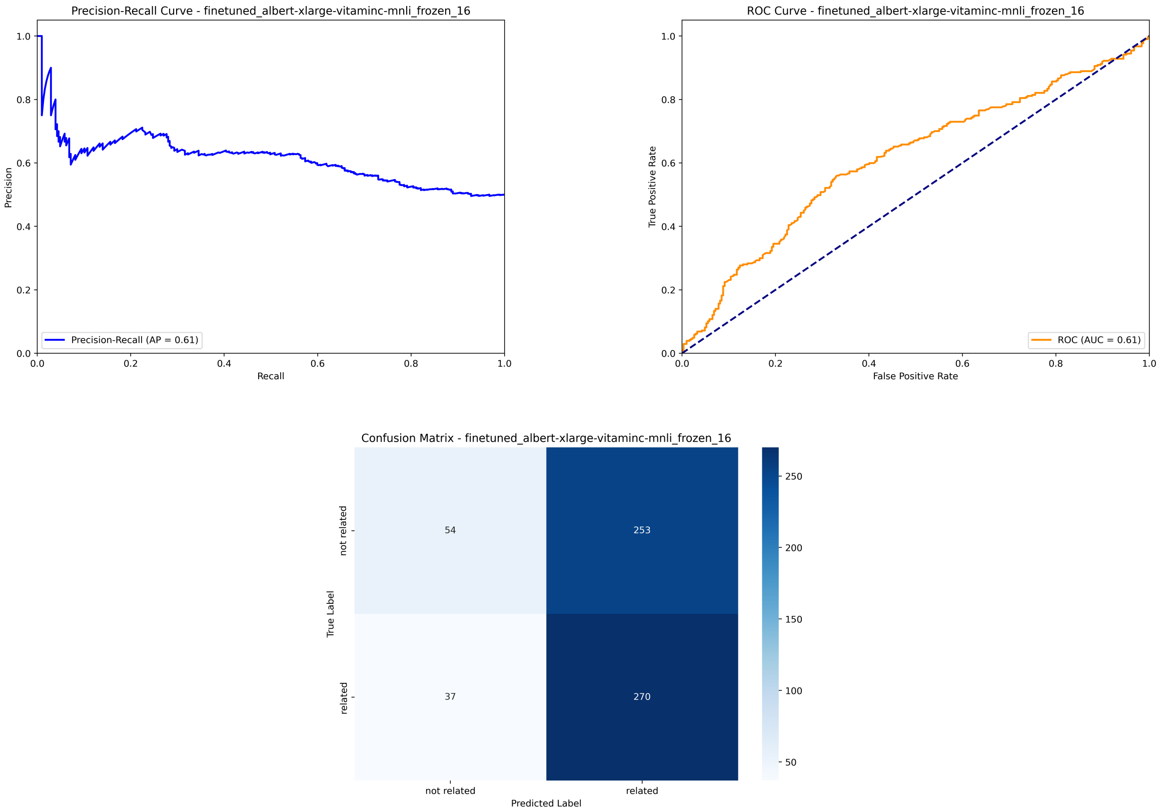

**Figure 6:** Metrics for few-shot learning after fine-tuning `albert-xlarge-vitaminc-mnli` with 16 labelled instances .

By contrast, our few-shot fine-tuning of the fact-checking model `albert-xlarge-vitaminc-mnli` achieved modest but consistent gains (0.53 accuracy, 0.61 AUC, 0.61 AP) with only 16 examples—slightly outperforming SetFit's contrastive approach ( 0.52–0.53 accuracy) on the 32-sample mdoel. Freezing the ALBERT transformer and training only the classification head preserved the model's pre-learned fact-checking reasoning while avoiding severe overfitting on our tiny training pool. However, overall performance remains limited by the very small sample size and by potential domain mismatch between VitaminC pretraining and our climate-focused claims. Future improvements could include a larger training pool, longer learning loops to select the most informative examples, or a domain-specific (as opposed te only task-specific) trained model.