Final Report

"Bait Buster"

Presented by Lihi Nofar, Tomer Portal, Aviv Elbaz

Problem Description - Clickbait in Digital Media

Deceptive headlines using psychological manipulation to drive engagement

Why essential:

- 72% of users feel tricked by misleading headlines (Pew Research)
- Erodes trust: 68% say clickbait reduces overall media credibility
- Financial impact: Wastes $7B/year in misplaced ad revenue

Why challenging:

- Stylistic ambiguity: Legitimate teasers vs. manipulative hooks
- Cultural context: Humor/sarcasm varies across regions
- Adaptive tactics: Continuous evolution of manipulation methods

Project Objectives

1. Develop dual-capability system:

    1.1. Detection (binary classification)

    1.2. Method attribution (multi-label classification)

2. Compare architectural approaches:

    2.1. Single-step (joint detection + attribution) using LLMs

    2.2. Two-step (cascaded models) using fine-tuned BERT

3. Establish evaluation framework:

    3.1. Quantitative metrics (F1, accuracy, recall, precision)

    3.2. Human evaluation protocol

Formal Task Specification

Input

News headline text (original or clickbait-modified).
between 15-20 words avg.

Output

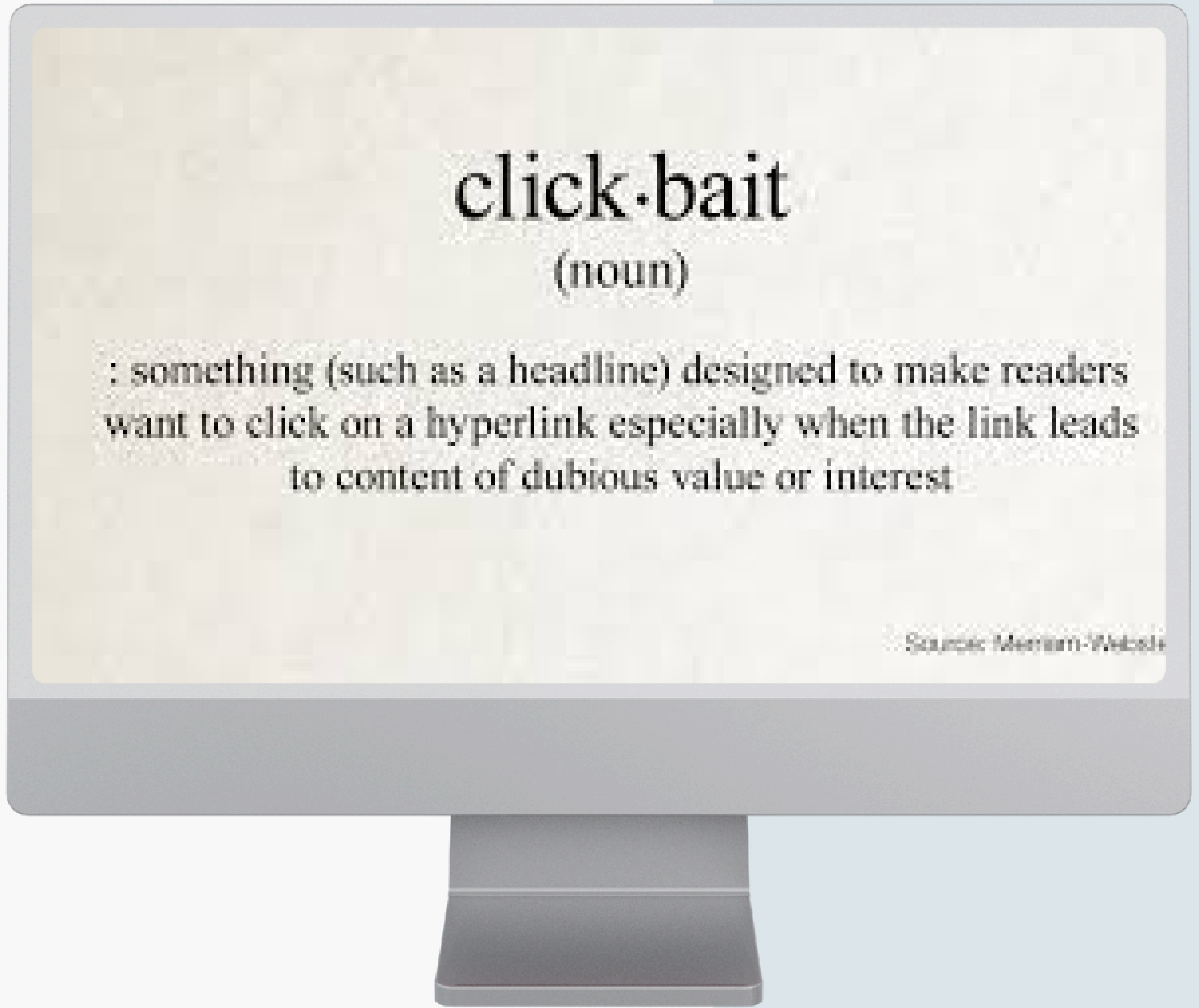
Binary classification: Clickbait (1) or Not (0). If clickbait, multi-label classification: Which tactics/styles were used?

Metrics

Accuracy, Precision, Recall, F1-score for detection. Multi-label metrics for tactics attribution (macro/micro F1).

Implementation Plan

1. **Generate dataset** with real headlines and clickbait variants using LLMs and style methods.
2. **Preprocess** and label the dataset.
3. **Train and evaluate**:
   - Single-step pipeline: simultaneous detection and tactics attribution. For example, using GPT-4o zero/few-shot.
   - Two-step pipeline: separate detection and attribution models.
4. **Analyze** and compare model performances.



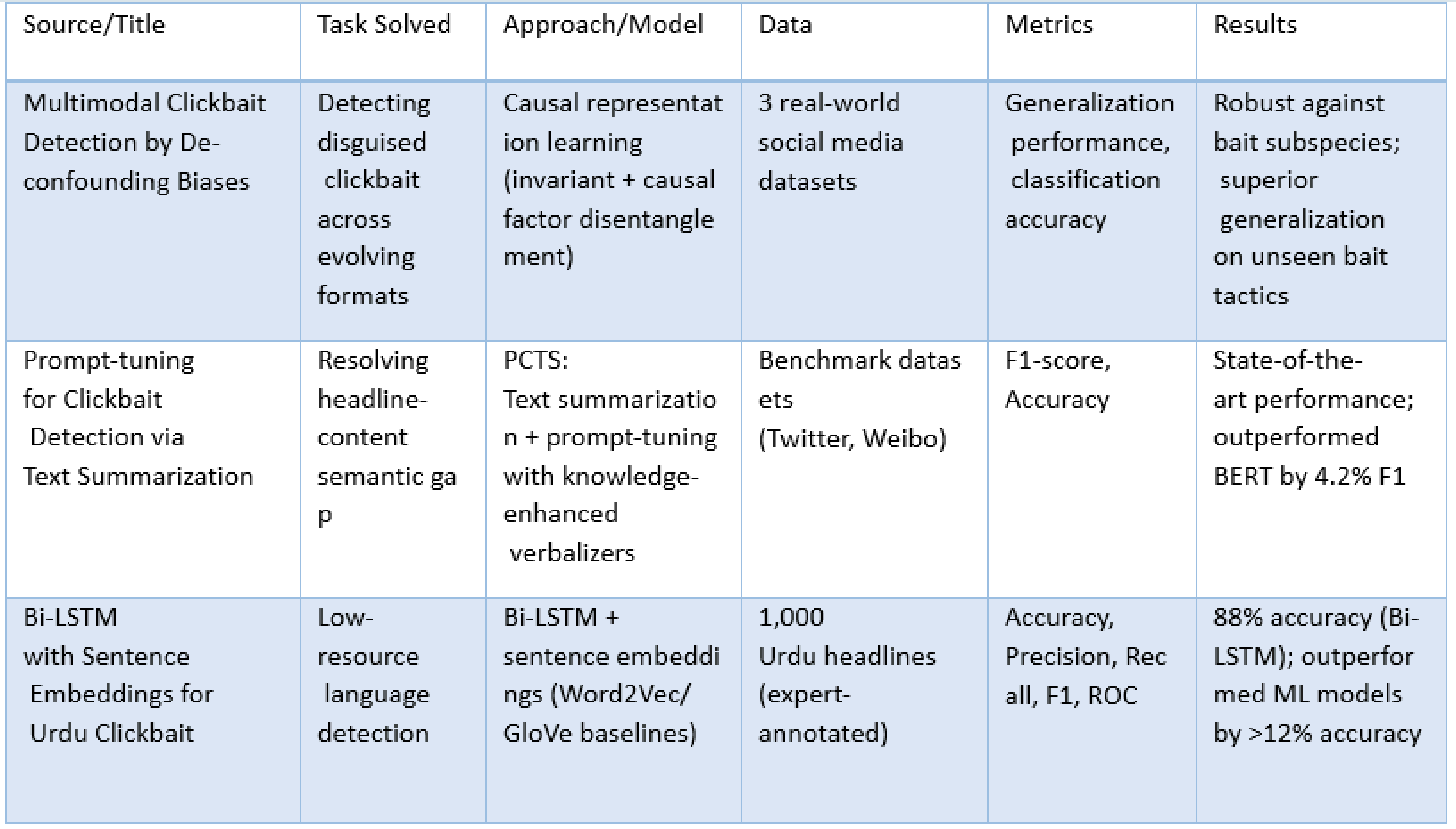
click·bait

(noun)

: something (such as a headline) designed to make readers want to click on a hyperlink especially when the link leads to content of dubious value or interest

Source: Merriam-Webst

Prior Art



Source/Title	Task Solved	Approach/Model	Data	Metrics	Results
Multimodal Clickbait Detection by De-confounding Biases	Detecting disguised clickbait across evolving formats	Causal representation learning (invariant + causal factor disentanglement)	3 real-world social media datasets	Generalization performance, classification accuracy	Robust against bait subspecies; superior generalization on unseen bait tactics
Prompt-tuning for Clickbait Detection via Text Summarization	Resolving headline-content semantic gap	PCTS: Text summarization + prompt-tuning with knowledge-enhanced verbalizers	Benchmark datasets (Twitter, Weibo)	F1-score, Accuracy	State-of-the-art performance; outperformed BERT by 4.2% F1
Bi-LSTM with Sentence Embeddings for Urdu Clickbait	Low-resource language detection	Bi-LSTM + sentence embeddings (Word2Vec/ GloVe baselines)	1,000 Urdu headlines (expert-annotated)	Accuracy, Precision, Recall, F1, ROC	88% accuracy (Bi-LSTM); outperformed ML models by >12% accuracy

Data Preparation

Source dataset description

link to News headlines 2024 in kaggle:
https://www.kaggle.com/datasets/dylanjcastillo/news-headlines-2024
scraped between April 25th and April 26th, 2024

Description of relevant fields:

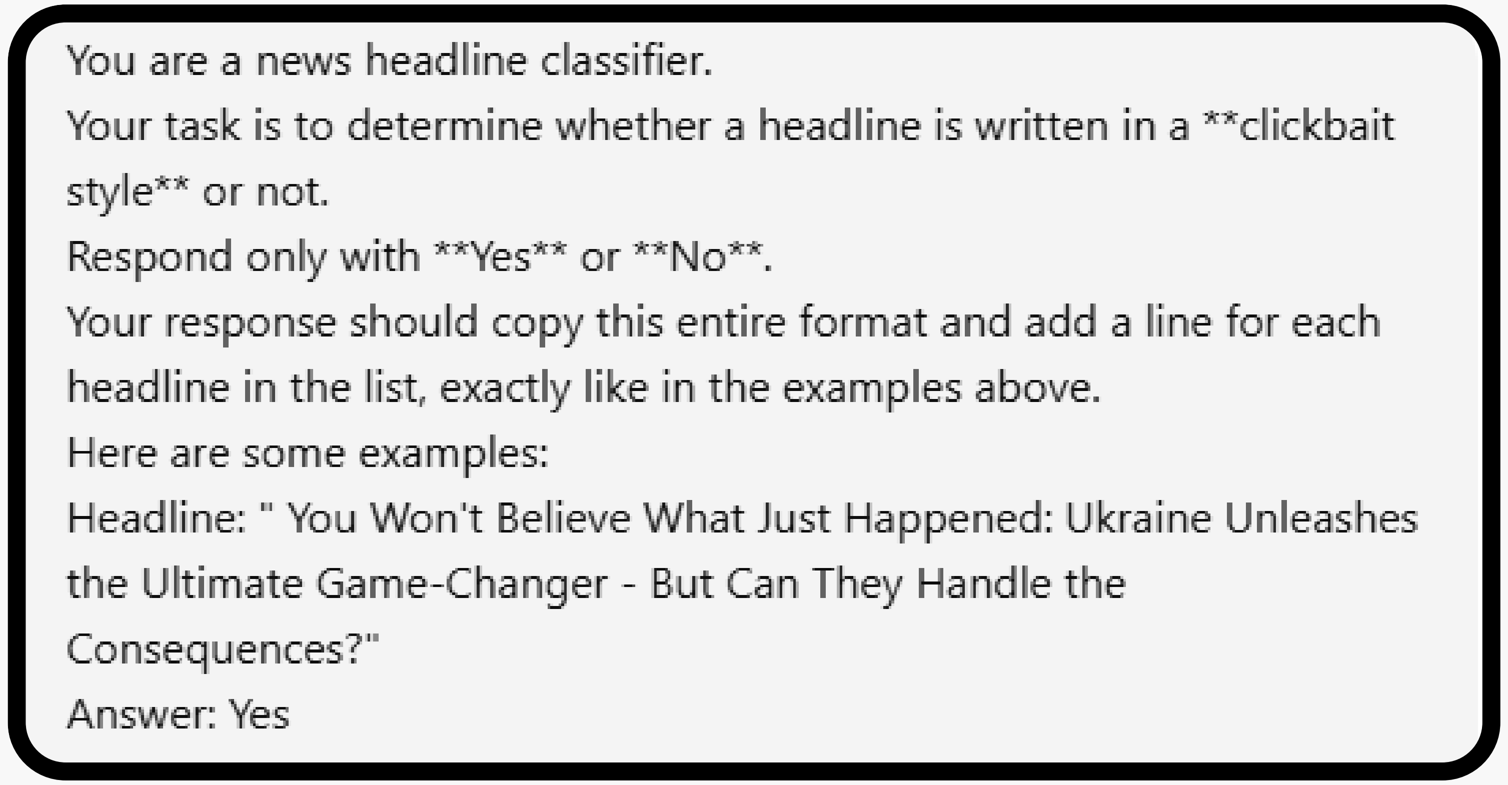
clickbait_dataset.csv — 3 columns:

- original: Original (non-clickbait) headline.
- clickbait: Clickbait version of the headline (used as model input).
- methods: List of clickbait tactics (binary vector of 10 values).

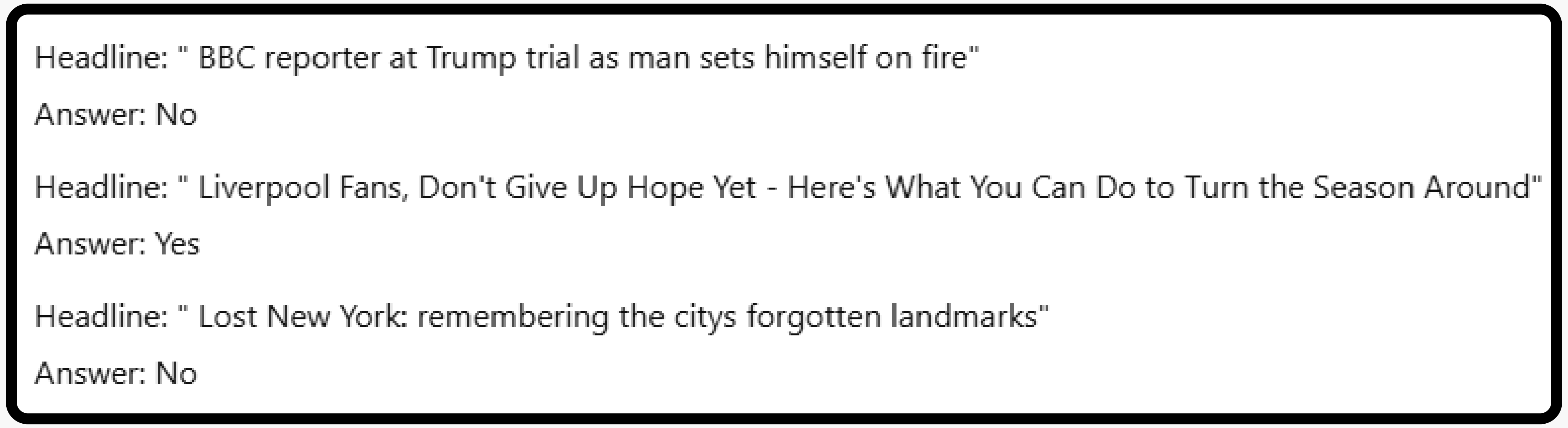
Data generation:

- Clickbait headline versions were created based on original headlines. The tactics were select randomly: for each headline, relevant tactics were selected from the predefined list of 10 clickbait tactics.
- Additional data augmentation was applied during training (using the augment_text() function), creating variations of clickbait headlines.

Prompts and examples



You are a news headline classifier.
Your task is to determine whether a headline is written in a **clickbait style** or not.
Respond only with **Yes** or **No**.
Your response should copy this entire format and add a line for each headline in the list, exactly like in the examples above.
Here are some examples:
Headline: " You Won't Believe What Just Happened: Ukraine Unleashes the Ultimate Game-Changer - But Can They Handle the Consequences?"
Answer: Yes



Headline: " BBC reporter at Trump trial as man sets himself on fire"
Answer: No

Headline: " Liverpool Fans, Don't Give Up Hope Yet - Here's What You Can Do to Turn the Season Around"
Answer: Yes

Headline: " Lost New York: remembering the citys forgotten landmarks"
Answer: No

- For GPT-based one-step pipeline, few-shot prompting was used to generate clickbait/tactics predictions.
- The same clickbait headlines and tactics were used as inputs to the BERT/RoBERTa-based models.

Model Used

Clickbait Detection (Step 1):

Model: TFBertForSequenceClassification (binary classification). Pretrained model: bert-base-uncased.

Tactic Attribution (Step 2):

Model: TFRobertaForSequenceClassification (multilabel classification, 10 labels). Pretrained model: roberta-base. With custom threshold tuning for better multilabel performance.

GPT-based one-step pipeline:

Models used: GPT-4o (OpenAI), Gemini 2.5 Flash (Google DeepMind). Using prompting only (no fine-tuning).

model/pipelines are used

Two pipelines were evaluated:

One-step pipeline - a single large language model (LLM) prompt performs both clickbait detection (binary: clickbait vs. non-clickbait) and tactic attribution (identifying which of the 10 clickbait tactics are used, or "non-clickbait" if none).

Two-step pipeline -the process is split into two independent parts: Clickbait detection and tactic attribution.

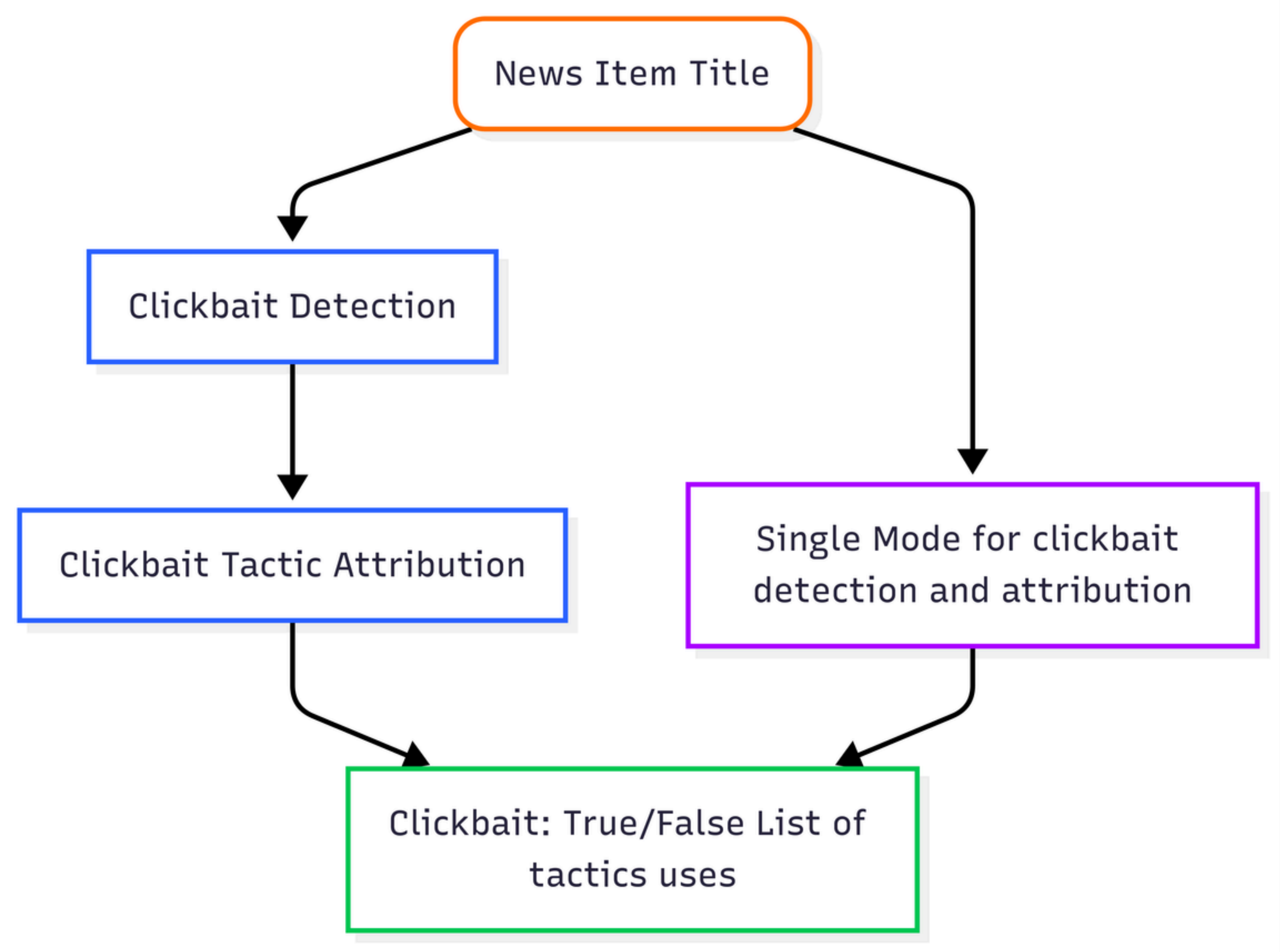
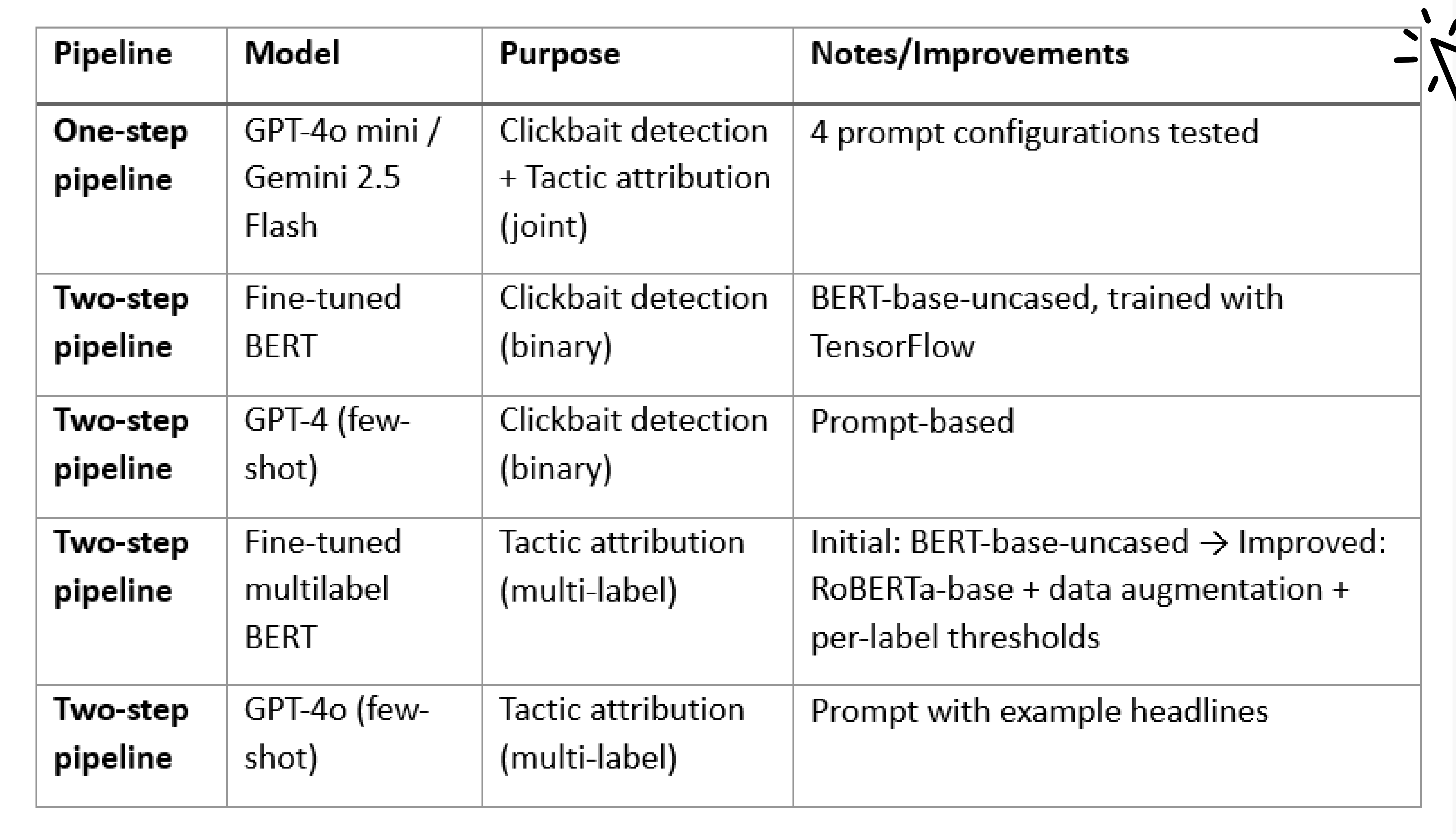

News Item Title
Clickbait Detection
Clickbait Tactic Attribution
Single Mode for clickbait detection and attribution
Clickbait: True/False List of tactics uses

Table of models/configurations



Pipeline	Model	Purpose	Notes/Improvements
One-step pipeline	GPT-4o mini / Gemini 2.5 Flash	Clickbait detection + Tactic attribution (joint)	4 prompt configurations tested
Two-step pipeline	Fine-tuned BERT	Clickbait detection (binary)	BERT-base-uncased, trained with TensorFlow
Two-step pipeline	GPT-4 (few-shot)	Clickbait detection (binary)	Prompt-based
Two-step pipeline	Fine-tuned multilabel BERT	Tactic attribution (multi-label)	Initial: BERT-base-uncased → Improved: RoBERTa-base + data augmentation + per-label thresholds
Two-step pipeline	GPT-4o (few-shot)	Tactic attribution (multi-label)	Prompt with example headlines

How the models were trained

The Clickbait detection model

used a TFBertForSequenceClassification (binary classification) trained on merged source and clickbait headlines, tokenized with BertTokenizer. It was optimized with Adam (learning rate 0.00002), trained for 3 epochs with batch size 32. Evaluation metrics: accuracy, precision, recall, and F1 on the test set.

The Tactic attribution model

used TFBertForSequenceClassification (multi-label, 10 labels) trained on clickbait_dataset.csv, with headlines and tactic vectors. Tokenization was done with BertTokenizer. The model used Adam (learning rate 0.00002), 3 epochs, batch size 16. Performance was evaluated with precision, recall, and F1 per tactic. Improvements included switching to RoBERTa-base, adding data augmentation, and applying per-label thresholds.

Data split

Clickbait detection:

- Dataset: 1,740 headlines (870 clickbait + 870 regular).
- Split: 80% training (1,392 examples), 20% testing (348 examples).

Tactic attribution:

- pre-improvement model: 870 clickbait titles. 20% test and 80% training.
- Improved model (used with augmented): 1740 clickbait titles. (also 80% and 20%)

Platform: Google Colab (CPU runtime).

Metrics

Metric was used at each step:

- Clickbait detection: Accuracy, Precision, Recall, F1-score
- Tactic attribution: Precision, Recall, F1-score per tactic plus micro, macro, and weighted averages.

Classification:

Both steps are classification tasks:

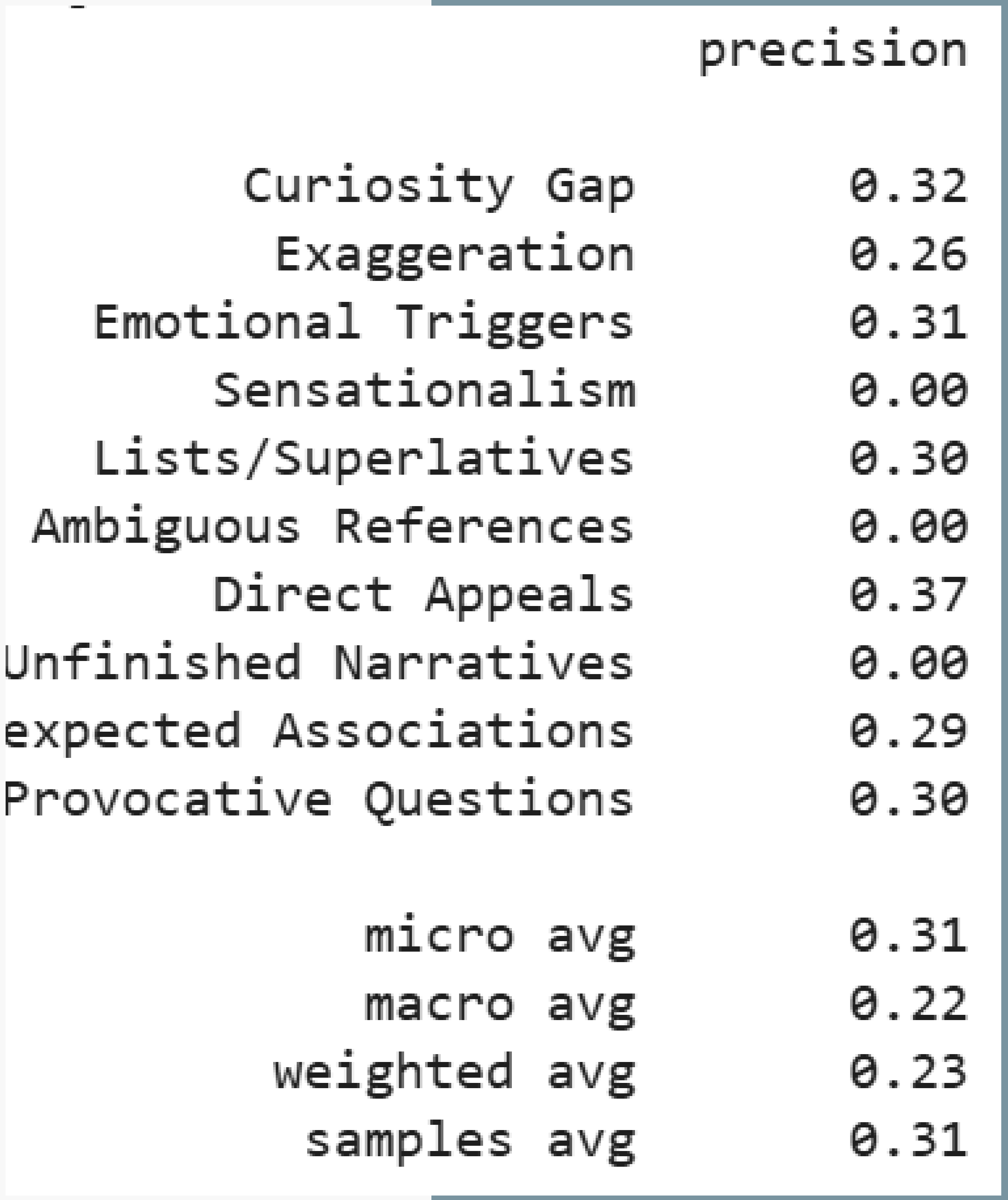
Clickbait detection - binary classification

Tactic attribution- multilabel classification (10 independent binary labels).

validations set and test set:

During training: validation set = the test set (20% of the data). No additional validation is used (like validation_split inside the train).

During evaluation: the same test set, used for the final performance evaluation (classification_report, precision/recall/F1).



```
                       precision

        Curiosity Gap       0.32
         Exaggeration       0.26
   Emotional Triggers       0.31
       Sensationalism       0.00
   Lists/Superlatives       0.30
 Ambiguous References       0.00
       Direct Appeals       0.37
Unfinished Narratives       0.00
expected Associations       0.29
Provocative Questions       0.30

            micro avg       0.31
            macro avg       0.22
         weighted avg       0.23
          samples avg       0.31
```

Code Organization

Results files:

The evaluation results were generated in the notebook using sklearn.metrics. The evaluation report includes Precision, Recall, and F1-score, as well as support and overall micro, macro, and weighted averages.

Link to GitHub:

nlp-hit-2025/clickbait

Contribute to nlp-hit-2025/clickbait development by creating an account on GitHub.

GitHub

https://github.com/nlp-hit-2025/clickbait

Data files:

contains original headline, clickbait headline (used as model input), and a binary vector of 10 clickbait tactics.

Major tasks and code files/functions

There are 4 code files in the project. One for generating clickbait headlines, the second and third for a two step pipeline, and the last for a one step pipeline.

Single-Step Detection Pipeline

Single-Step Pipeline

Simultaneous clickbait detection and method inference through one LLM prompt.

Unified Assessment

Each model assesses a headline (some models receive examples of clickbait and non-clickbait), having been given a dictionary of clickbait creation tactics.

Efficiency, At a Cost

Streamlined, Single-prompt-fits-all approach, with no training or fine-tuning.
Tendency towards False-Positives and over attribution of methods.

LLM Configurations for Clickbait

Evaluating Performance in Detection

1

Evaluated 4 LLM Configurations

We assessed GPT-4o mini and Gemini-2.0 Flash, each on both Zero-Shot and Two-Shot prompts.

2

Prompted with Headlines

Each model was prompted with various headlines to analyze their classification capabilities.

3

Performance Metrics

Calculated Micro and Macro F1-Scores for multi-label method inference, Accuracy and F1-Score for binary clickbait detection.

4

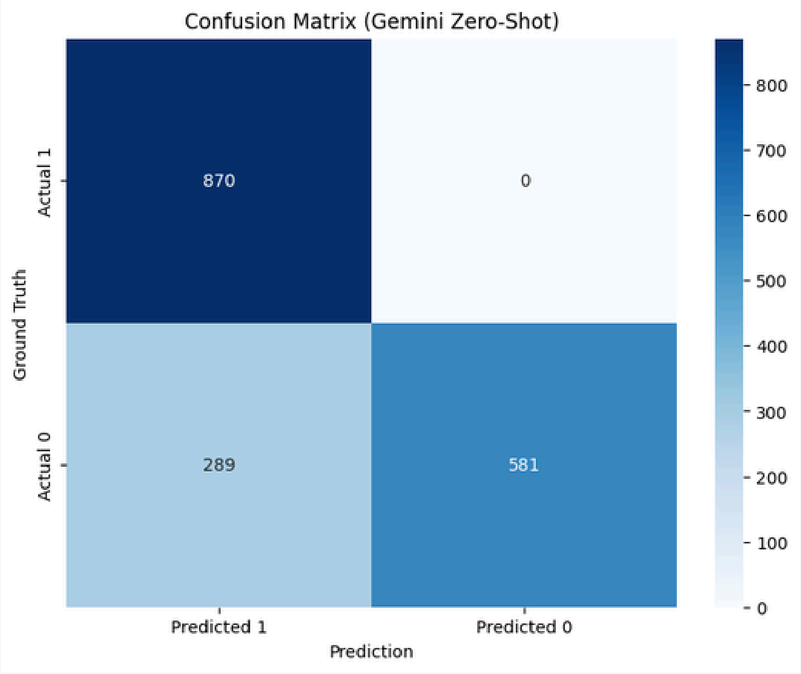
Capabilities and Limitations

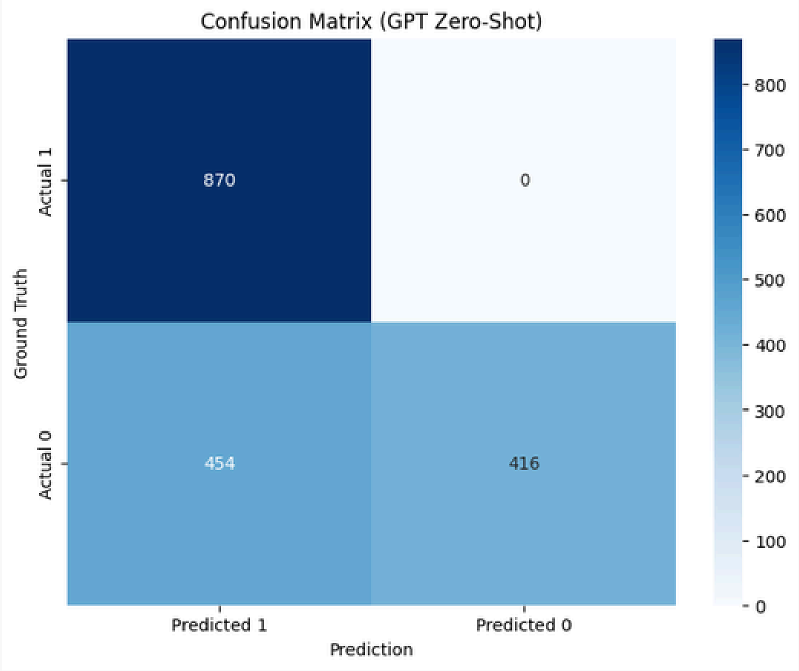
Results led to a clearer understanding of the models' detection capabilities and inherent limitations.

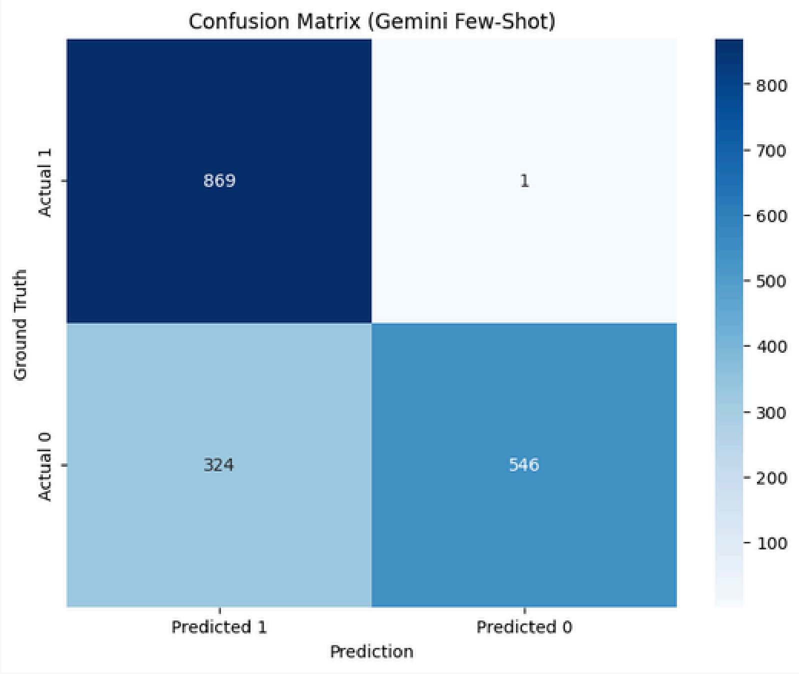
Evaluation Metrics for Clickbait Detection

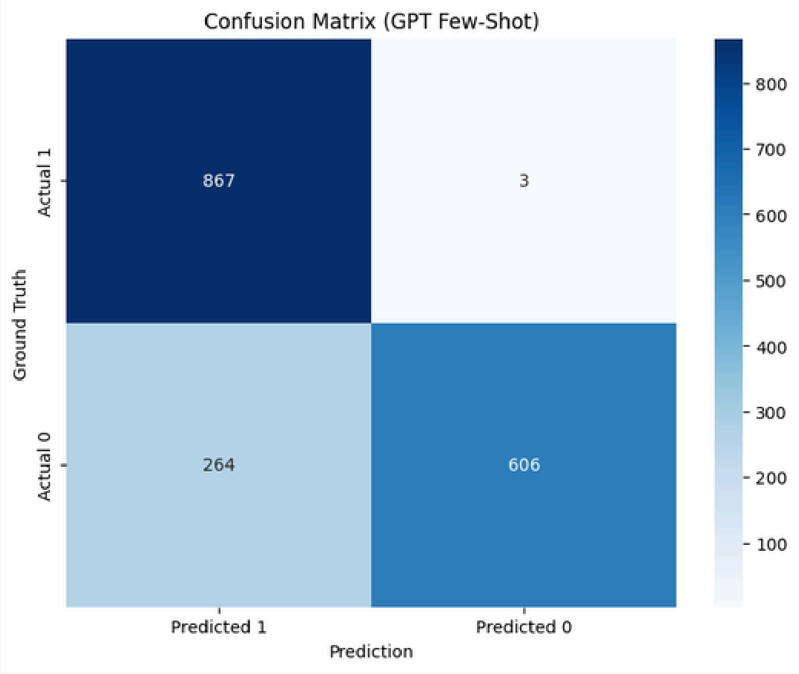
Measuring model performance with F1-scores and precision

			Worsened	Improved
Task	**Gemini Zero-Shot**	**GPT Zero-Shot**	**Gemini Few-Shot**	**GPT Few-Shot**
Method Inference	**Highest Micro-F1 0.474**	**Micro-F1 0.400**	**Lower Micro-F1 0.466**	**Improved Micro-F1 0.434**
Classification	**Accuracy 0.834**	**Lowest Accuracy 0.739**	**Lower Accuracy 0.813**	**Improved Accuracy 0.847**





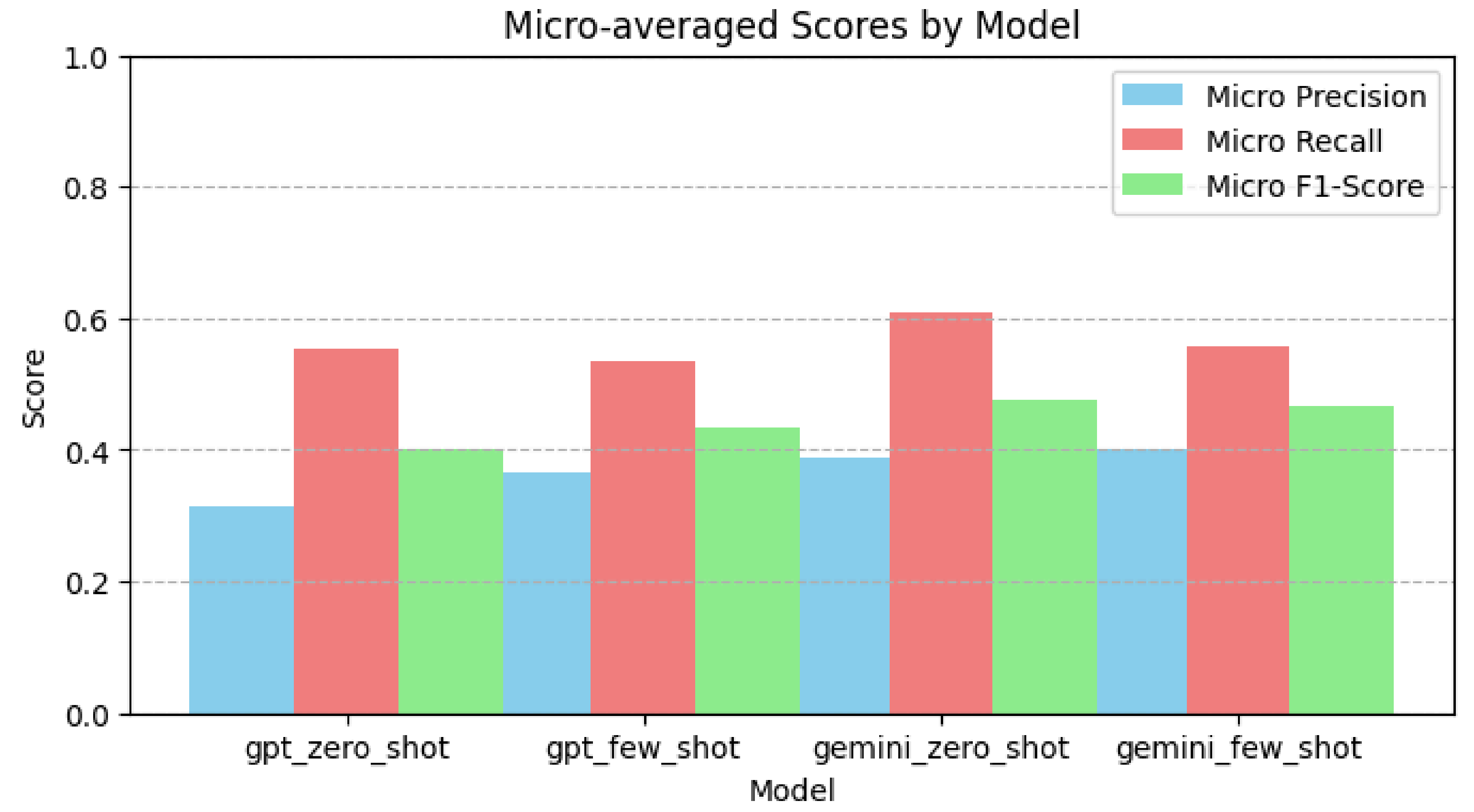


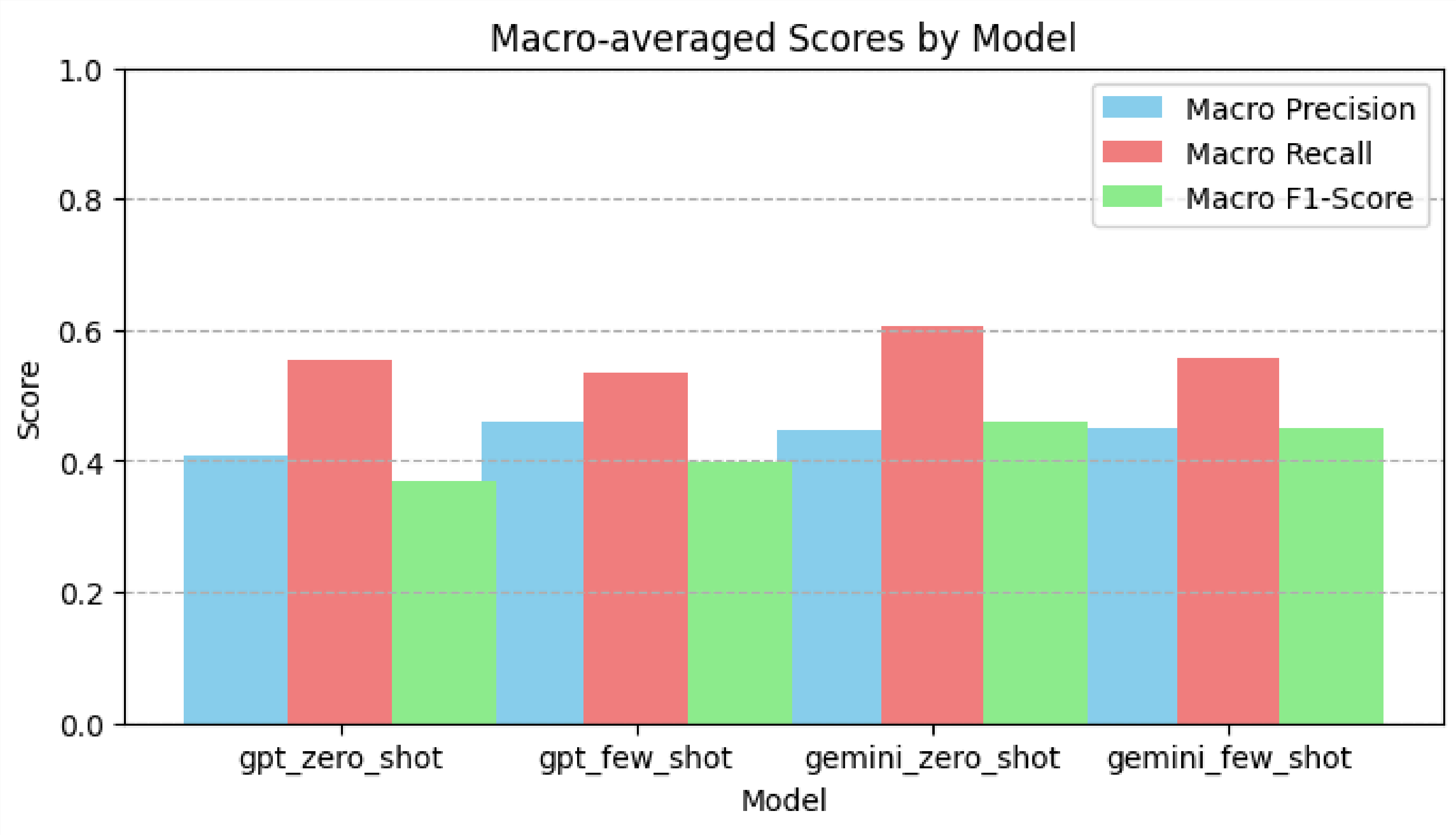


Evaluation Metrics for Clickbait Detection

Measuring model performance with F1-scores and precision

Task	Gemini Zero-Shot	GPT Zero-Shot	Worsened Gemini Few-Shot	Improved GPT Few-Shot
Method Inference	**Highest Micro-F1 0.474**	**Micro-F1 0.400**	**Lower Micro-F1 0.466**	**Improved Micro-F1 0.434**
Classification	**Accuracy 0.834**	**Lowest Accuracy 0.739**	**Lower Accuracy 0.813**	**Improved Accuracy 0.847**





Two-Step Clickbait Pipeline

Clickbait Detection and Method Inference on Different Models

Task Separation

Clickbait detection and tactic attribution are performed sequentially and independently.

Allows each model to specialize, with no mutual interference.

Accuracy and Robustness

Fine-tuned BERT exhibits near-perfect performance in clickbait detection. Tactics attribution benefits from enhanced architectures (e.g., RoBERTa), data augmentation, and custom thresholds.

Dedicated Models

Clickbait Detection: Fine-tuned BERT or GPT-4 with few-shot prompting is used for binary classification.

Tactics Attribution: Multi-label BERT or GPT-4 with few-shot prompting identifies clickbait tactics.

Efficiency Trade-Offs

Offers fine-grained insights and higher adaptability compared to the single-step pipeline, but demands more computational resources and time.

BERT and GPT-4 Overview

Understanding Their Roles in Detection Pipelines

BERT

Used primarily for classification tasks, providing enhanced context understanding which is crucial for accurate results.

GPT-4

Employed with few-shot prompting, significantly improving tactic attribution by leveraging its advanced language generation capabilities for diverse scenarios.

Detection Pipeline

The integration of BERT and GPT-4 forms a robust detection pipeline that enhances the identification and classification of clickbait content through sophisticated model interactions.

BERT for Clickbait Detection

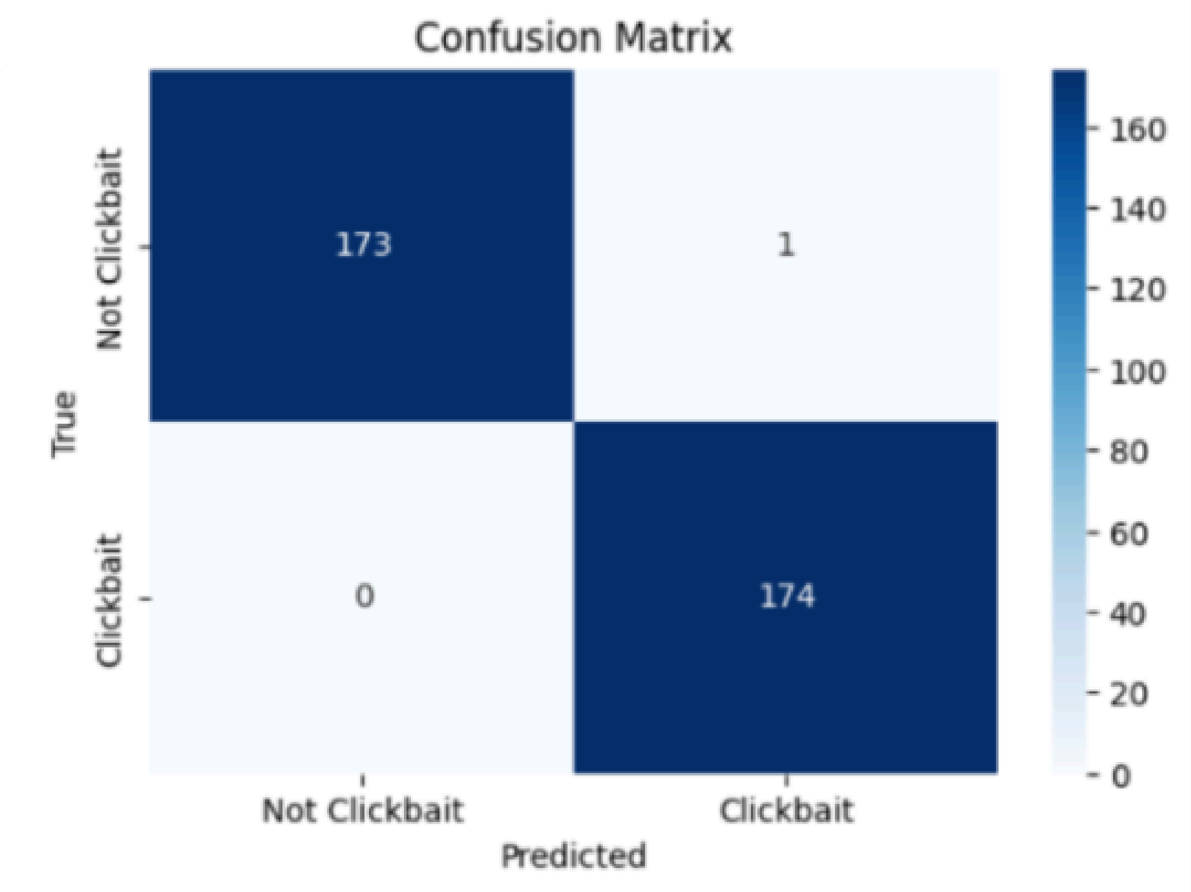
Leveraging BERT to identify clickbait headlines

Fine-tuning BERT model

We adapted the BERT model for clickbait detection, optimizing its parameters to improve performance and accuracy in distinguishing between types of headlines.

Achieved 100% accuracy

The fine-tuned model reached an impressive accuracy rate of 100%, indicating its effectiveness in clickbait detection scenarios.



Multi-Label BERT Insights

Enhancements in Clickbait Detection Accuracy

Multi-Label BERT Model

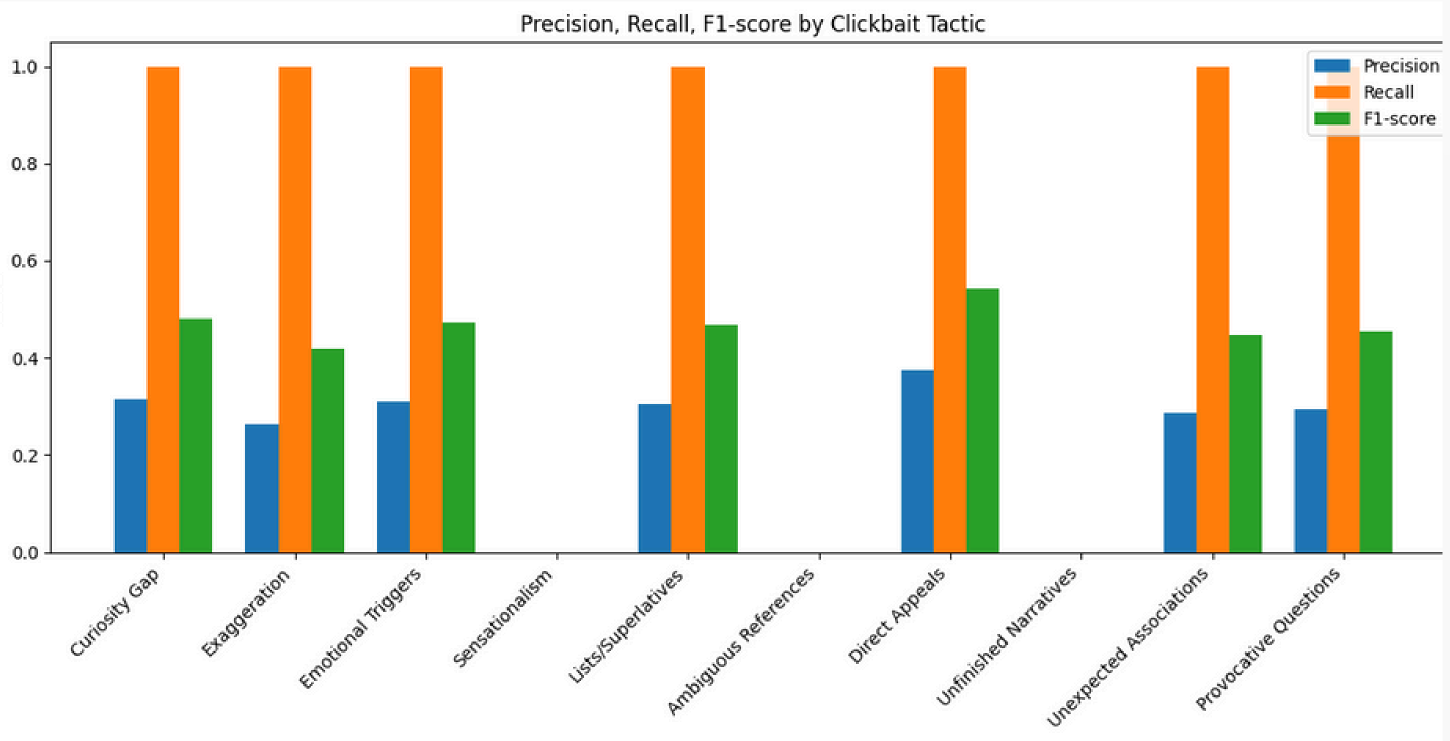
We trained a multi-label BERT model to classify various tactics in clickbait headlines.

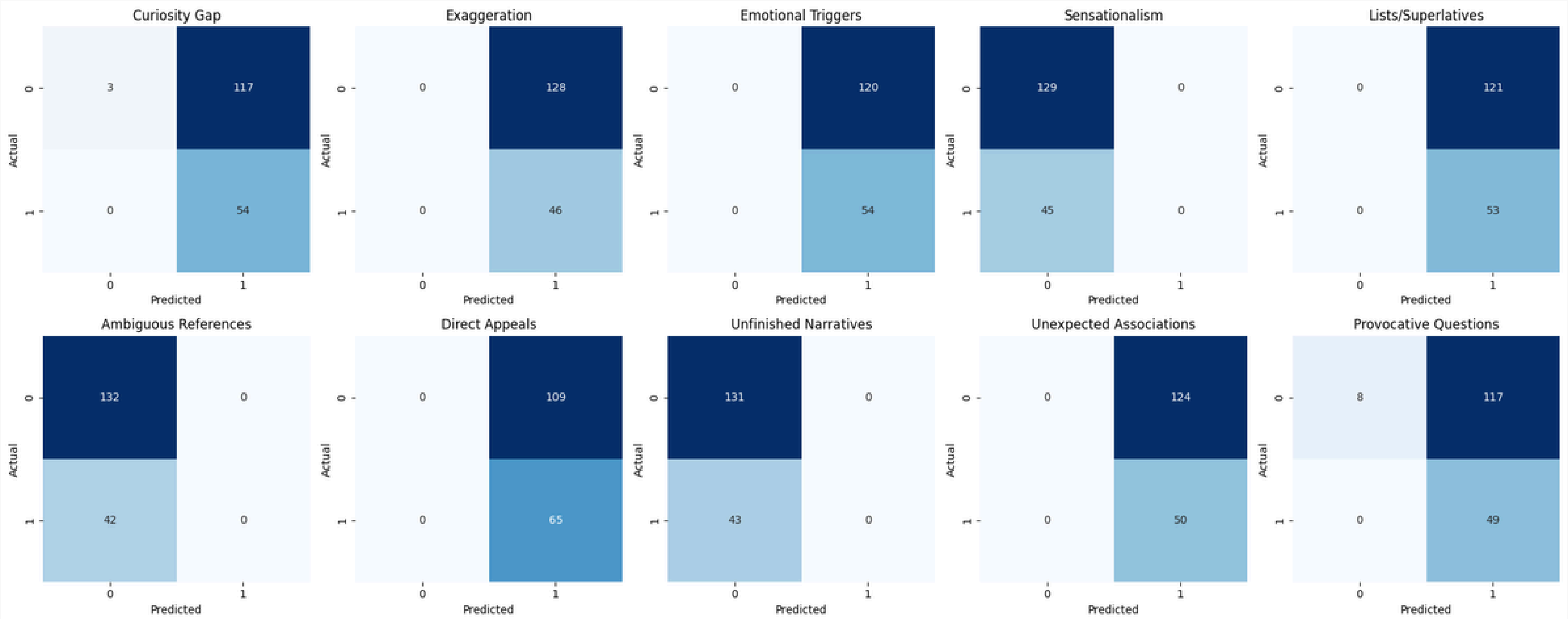
High Recall, Low Precision

Recall of 1.00 for most tactics means the model recognizes almost all cases where a tactic is present;
Precision around 0.30 means it also “guesses” many labels that are not actually present.
Conclusion: The model tends to "over-recognize" labels – too sensitive but not accurate.

Missed Tactics

Sensationalism Ambiguous References Unfinished Narratives – The model did not recognize these tactics at all, perhaps they are too rare or obfuscated in the data.





Multi-Label BERT Improvements

Enhancements in Clickbait Detection Accuracy

Utilizing RoBERTa

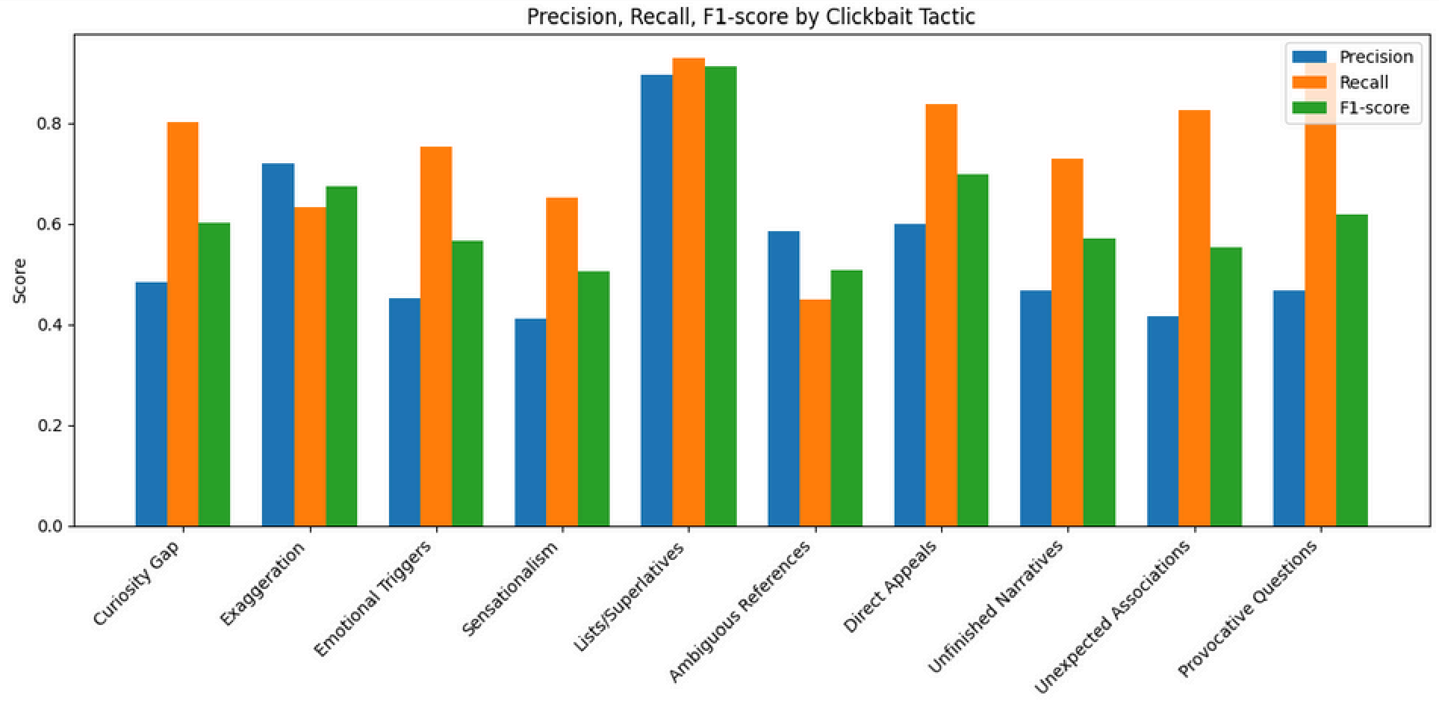
Switching to RoBERTa enhanced the model’s accuracy, providing better context understanding for clickbait classification.

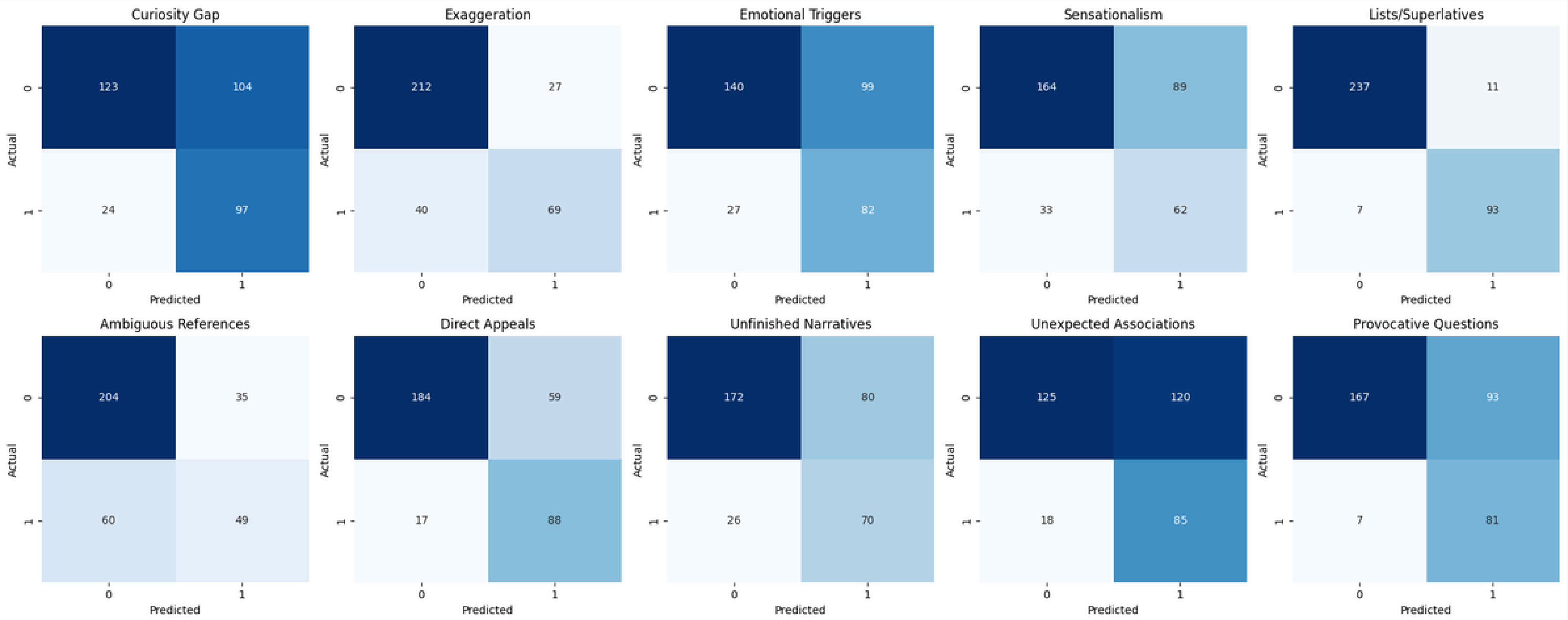
Data Augmentation

Duplicate versions of titles with different wordings (e.g. bolded words, exclamation marks, marketing wording) were added, to diversify the data and improve the overall ability of the model.

Adjusted Prediction Thresholds

instead setting p > 0.5 as indicating a positive label, a threshold was calculated separately for each tactic using ROC analysis on the validation set





Evaluating GPT-4's Few-Shot Prompting

Analyzing the effectiveness of few-shot prompting in classifying clickbait tactics with GPT-4

High accuracy in classification

GPT-4 demonstrated high accuracy when classifying clickbait tactics through few-shot prompting.

- **False positives observed**

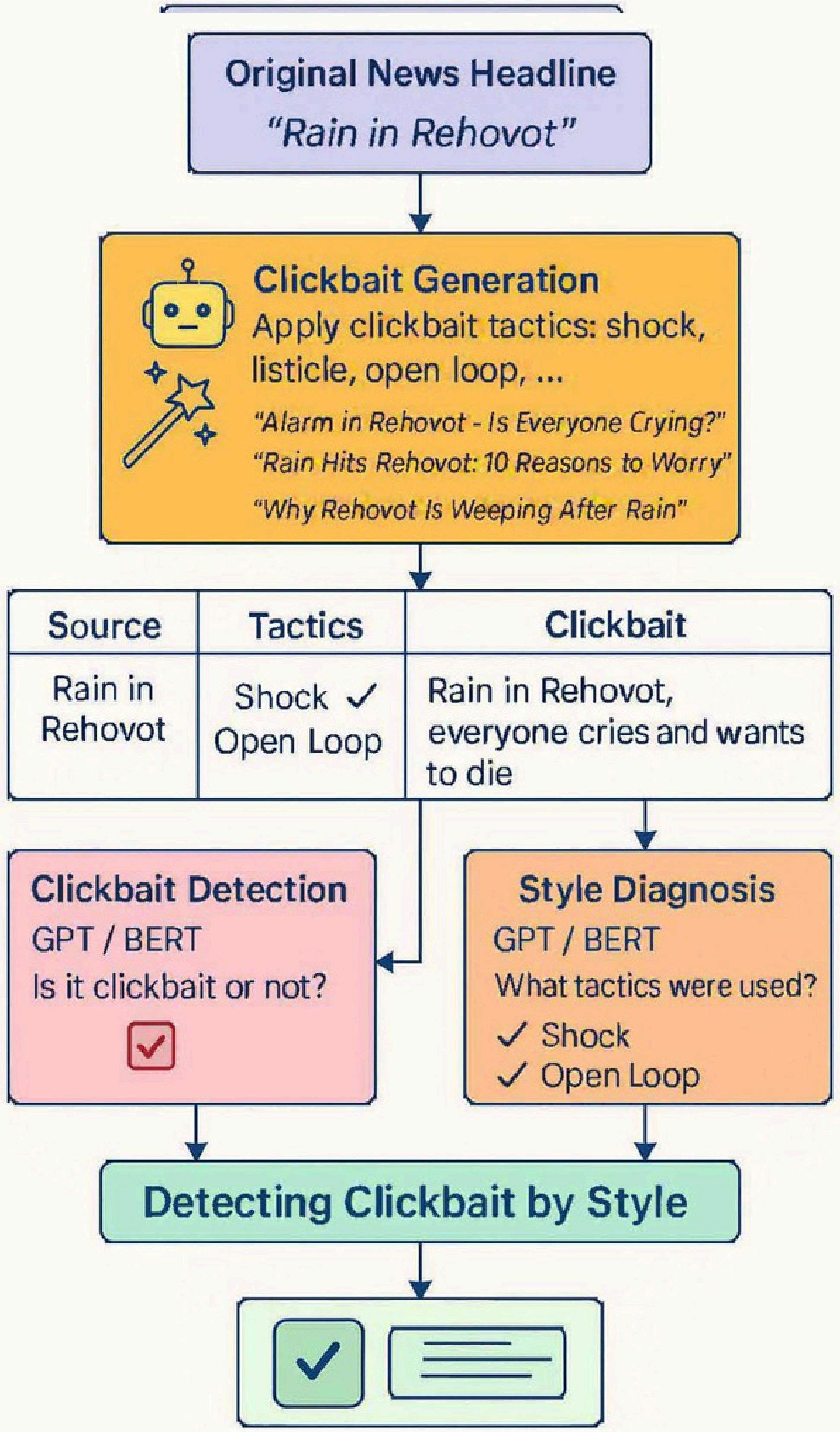
  Despite its strengths, the model produced false positives, indicating a cautious approach that may lead to misclassification.

- **Challenges for Clickbait Attribution**

Further False-Positives: Model over-attributes tactics, including correct ones but inferring extras, leading to high Recall but lower Precision.

Ambiguous Tactic Interpretation: Model tends to interpret tactics too broadly. For example, emotional expressions like “You’ll Be Furious” will also be labeled as Direct Appeals, even though their main meaning is to activate emotion.

Graphical Abstract



Original News Headline

"Rain in Rehovot"

Clickbait Generation

Apply clickbait tactics: shock, listicle, open loop, ...

"Alarm in Rehovot - Is Everyone Crying?"

"Rain Hits Rehovot: 10 Reasons to Worry"

"Why Rehovot Is Weeping After Rain"

Source	Tactics	Clickbait
Rain in Rehovot	Shock ✓ Open Loop	Rain in Rehovot, everyone cries and wants to die

Clickbait Detection

GPT / BERT

Is it clickbait or not?

☑

Style Diagnosis

GPT / BERT

What tactics were used?

✓ Shock

✓ Open Loop

Detecting Clickbait by Style