By: Nicole Poliak & Naveh Nissan

ADR Detect

Comparing LLM Generalization with Embedding-based Models for Adverse Drug Reaction Detection

Project Proposal

Motivation

Adverse Drug Reactions (ADRs) are a leading cause of patient harm and hospitalization, yet they are often underreported and buried within unstructured clinical text. Automatically detecting ADRs from these texts is essential for improving pharmacovigilance, enhancing patient safety, and reducing healthcare costs.

This project aims to explore and compare two NLP approaches for ADR detection:

General-purpose Large Language Models (LLMs) used in zero- or few-shot settings & Sentence embedding methods paired with a traditional classifier. By benchmarking both methods, we assess their effectiveness in enabling safer, AI-assisted medical decision-making.

Project Task

Goal: Compare general-purpose LLMs and embedding-based classifiers for ADR detection

- Input: A sentence from a medical case report.
- Output: Binary label (0/1) indicating whether the sentence describes an Adverse Drug Reaction (ADR).
- NLP Task: Binary sentence classification.

Two Paths:

Zero-/Few-shot LLM prompting:

GPT-4o-mini, GPT-4o, GPT-4.1, Phi-4-mini-instruct, LLaMA-3.2-3B-Instruct

Sentence embedding + classifier:

BioBERT, SBERT, InstructorXL → Logistic Regression

Key Challenges

01

Noisy and Ambiguous Text:
Patient-written reviews often include unrelated symptoms, making it hard to distinguish actual ADRs from coincidental mentions.

02

Domain-Specific Complexity:
Clinical and pharmaceutical texts contain medical jargon, abbreviations, and phrasing that may confuse general models.

03

Label Imbalance:
Only ~24% of the dataset contains ADR-positive sentences, which poses a challenge for both LLMs and classifiers.

Dataset Overview & Usage Strategy

Public Datasets

- ADE Corpus V2 : Expert-annotated PubMed case reports (23,516 sentences).
- PsyTAR: Patient-reported reviews from askapatient.com (6,009 sentences)

Data Type

Combined, structured and labeled dataset with three columns:

- Text: Sentence from medical abstract or patient review.
- Label: Binary (1 = ADR, 0 = non-ADR).
- Dataset: Dataset origin (ADE / PsyTAR).

Usage Strategy

- Embedding-based Models:

→ Train/dev/test split (60/20/20) using BioBERT, SBERT, InstructorXL.

- LLMs (GPT-4 models, Phi-4-mini-instruct, LLaMA-3.2-3B-Instruct):

→ No fine-tuning involved.

Input/Output Example

Negative Example (ADR Not Present):

Input: "The total amount of vitamin K received from the enteral feedings ranged from 50 to 115 micrograms/day, which is less than the normal daily intake of 300 to 500 micrograms."

Output: Label: 0

- This sentence only reports a vitamin K dosage range without mentioning any harm or negative reaction.

Positive Example (ADR Present):

Input: "Lupus-like syndrome caused by 5-aminosalicylic acid in patients with inflammatory bowel disease."

Output: Label: 1

- The sentence describes a drug (5-aminosalicylic acid) causing an adverse effect (Lupus-like syndrome).

Evaluation

Evaluation Metrics

- **Accuracy:** Overall percentage of correctly predicted labels.
- **Precision:** Of all predicted ADRs, how many were correct?
- **Recall:** Of all actual ADRs, how many were detected?
- **F1-Score:** Harmonic mean of precision and recall – balances both.

Evaluation Strategy

Embedding-Based Models:

- Train/dev/test split: 60/20/20, stratified by label.
- Input: Sentence → Embedding (InstructorXL, BioBERT, SBERT) → Logistic Regression
- Baseline: Naïve Bayes + Bag-of-Words.

LLM Models:

- Zero-shot & few-shot prompting only.
- Evaluated without fine-tuning.
- Prompts include task description + labeled examples (few-shot).
- Baseline: GPT-4o-mini zero-shot prompt.

Interim report

Description

Task

- **Input:** Sentence from biomedical literature (PubMed) / sentence from patient reviews (askapatient.com).
- **Output:** Binary label (0 = non-ADR, 1 = contains ADR).
- **Goal: Compare two paradigms for biomedical sentence classification:**

Zero-/Few-shot LLM prompting:
GPT-4o-mini, GPT-4o, GPT-4.1, Phi-4-mini-instruct, LLaMA-3.2-3B-Instruct

Sentence embedding + classifier:
BioBERT, SBERT, InstructorXL → Logistic Regression

Datasets:

- ADE Corpus V2: 23,516 expert-annotated sentences from biomedical literature.
- PsyTAR: 6,009 patient drug reviews.

Structure:

- Text: Sentence describing a clinical case.
- Label: ADR present (1) / not present (0).

Description

Usage:

- Embedding models: 60/20/20 stratified train/dev/test split.
- LLMs: Zero-shot and few-shot prompting (no fine-tuning).

Evaluation:

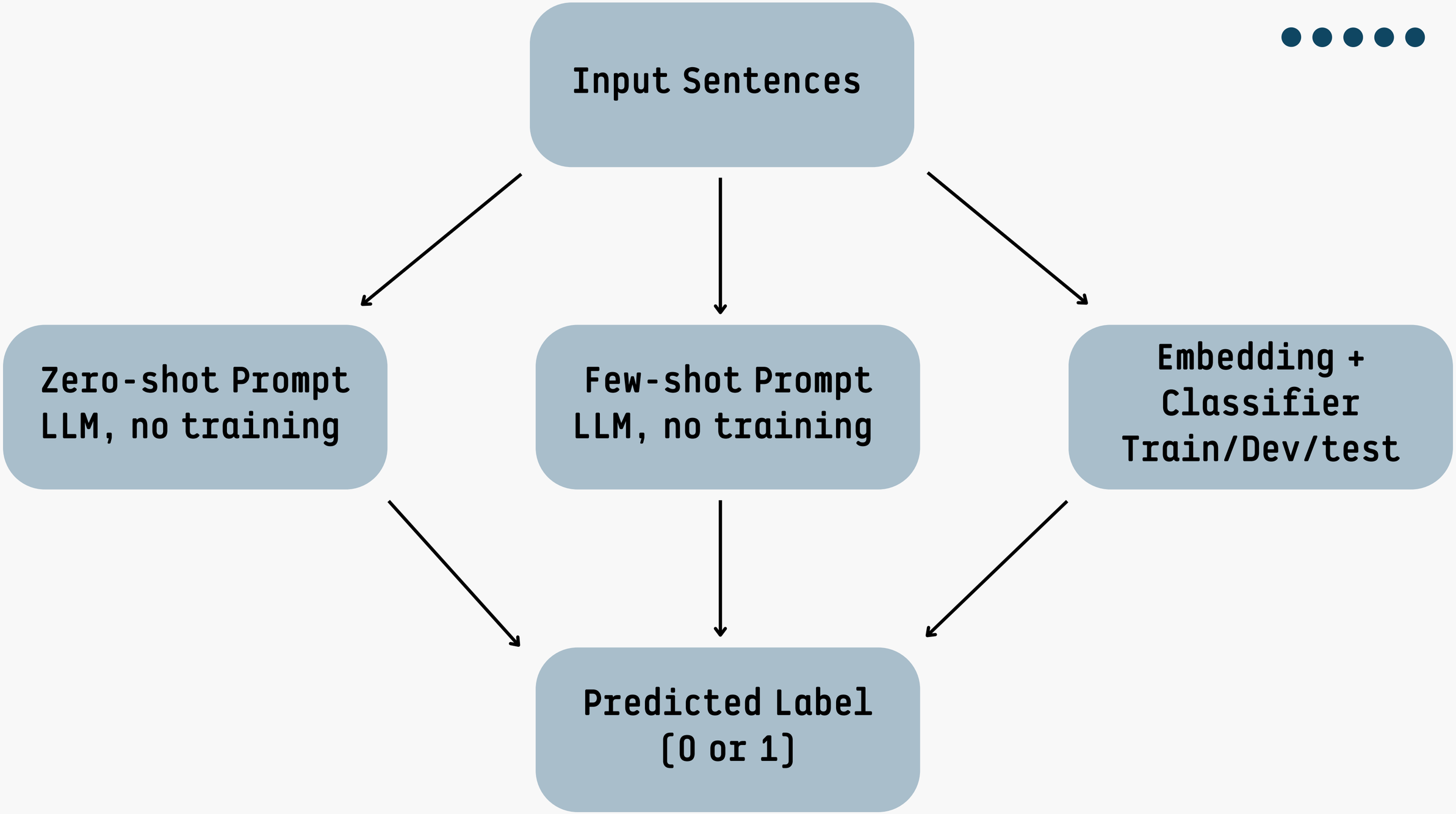
- Metrics: Accuracy, Precision, Recall, F1-Score.
- Embedding Models: Classifier trained on extracted embeddings.
- Baseline = Naïve Bayes + Bag-of-Words.
- LLMs: Zero/few-shot without training.
- Baseline = GPT-4o-mini zero-shot performance.

Prior Art

Source / Title	Approach / Model	Data	Metrics	Results
ModernBERT vs LLMs for Detecting Adverse Drug Reactions **Simmering.dev, 2025**	Comparison between: ModernBERT-base and ModernBERT-large as the structured language models with Llama 3.2-3B-instruct as the LLM.	ADE-Benchmark Corpus (23.5k labeled sentences)	Recall, Precision, F1 Score, Speed, Cost	**1a-modernbert-base:** f1-score - 86.0, recall - 90.3, precision - 82.2 **1b-modernbert-large:** f1-score - 89.2, recall - 91.8, precision - 86.8 **2-DSPy-25-threads-Llama-3.2-3B-Instruct:** f1-score - 80.7, recall - 87.9, precision - 74.6
LLMs are not Zero-Shot Reasoners for Biomedical Information Extraction **ACL Anthology, 2025**	Various open LLMs— BioMistral and Llama-2 models —using standard prompting, Chainof-Thought (CoT), Self-Consistency based reasoning and Retrieval-Augmented Generation (RAG).	14 different classification datasets from the BigBio collection like: PsyTAR dataset	F1 Score	Standard prompting (VANILLA); highest average F1 scores: **BioMistral-7B** (36.48%), **Llama-2-70B-Chat-AWQ** (40.34%), **Llama-2- 7b-chat-hf** (34.92%).
Using LLMs to Extract ADR's from Short Text **SCITEPRESS, 2025**	Evaluate various LLMs and ML approaches for ADR extraction and detection, Using multiple ADR datasets and a range of prompt formulations.	ASU-CHOP dataset, SMM4H dataset, WEB-RADRSMM4H dataset and ADE Corpus v2	Recall, Precision, F1 Score	**GPT-4o-mini** (Precision - 1.0, Recall - 0.91, F1 - 0.95) **GPT-4** (Precision - 1.0, Recall - 0.97, F1 - 0.98) **Llama** (Precision - 1.0, Recall - 1.0, F1 - 1.0)

Steps

Step	Description	Input → Output	Method/Tool	Metrics
1a. Preprocessing	Prepare text (for BoW & TF-IDF)	Raw sentence → Cleaned text	Lowercasing & punctuation removal	None (text preparation only)
1b. Explorative Data Analysis	Analyze dataset structure & class distribution	Data → Summary stats, visualizations	Pandas, Matplotlib, Seaborn	• Label distribution • Sentence/word length stats • Top frequent terms
2a. Baseline model	Simple lexical baseline using BoW features	Cleaned text → Binary label	CountVectorizer + Naïve Bayes	• Accuracy • Precision • Recall • F1-Score
2b. InstructorXL Embedding	Extract task-aware semantic features	Instruction + sentence → Dense vector	Instructor-XL	Evaluated in step 3a
2c. SBERT / BioBERT	Extract semantic features	Raw sentence → Dense vector (768D)	sentence-transformers / BioBERT / SBERT	Evaluated in step 3a
3a. Classification (Embeddings)	Train classifier on embeddings	Embedding → Binary label	Logistic Regression	• Accuracy • Precision • Recall • F1-Score • ROC-AUC
3b. LLM Prompting	Prompt LLM to return label	Sentence + prompt → Binary label	GPT-4o-mini, GPT-4o, GPT-4.1, Phi-4-mini-instruct, LLaMA-3.2-3B-Instruct (zero/few-shot, no train)	• Accuracy • Precision • Recall • F1-Score
4. Evaluation	Compare model outputs to true labels	Predictions + Labels → Scores	sklearn.metrics, seaborn, visualizations	• Confusion Matrix & ROC-AUC Curve (embedding models) • Metric comparison across models • Zero-/Few-shot performance gap (LLMs)


Input Sentences
Zero-shot Prompt
LLM, no training
Few-shot Prompt
LLM, no training
Embedding +
Classifier
Train/Dev/test
Predicted Label
(0 or 1)

Exploration & Baseline

Dataset

- ADR classification dataset combining ADE Corpus V2 and PsyTAR
- 26,867 sentences, 2 classes (ADR = 1, Non-ADR = 0) after cleaning & removing duplicates
- Minimal text cleaning: lowercasing, punctuation removal
- Mean sentence length: 17.2 ± 8.91 words (max 155 words)
- Notable class imbalance: ~76% Non-ADR, ~24% ADR → downsampled majority class to balance (6,431)

Baseline Results

1. Bag-of-Words (CountVectorizer) + Multinomial Naive Bayes

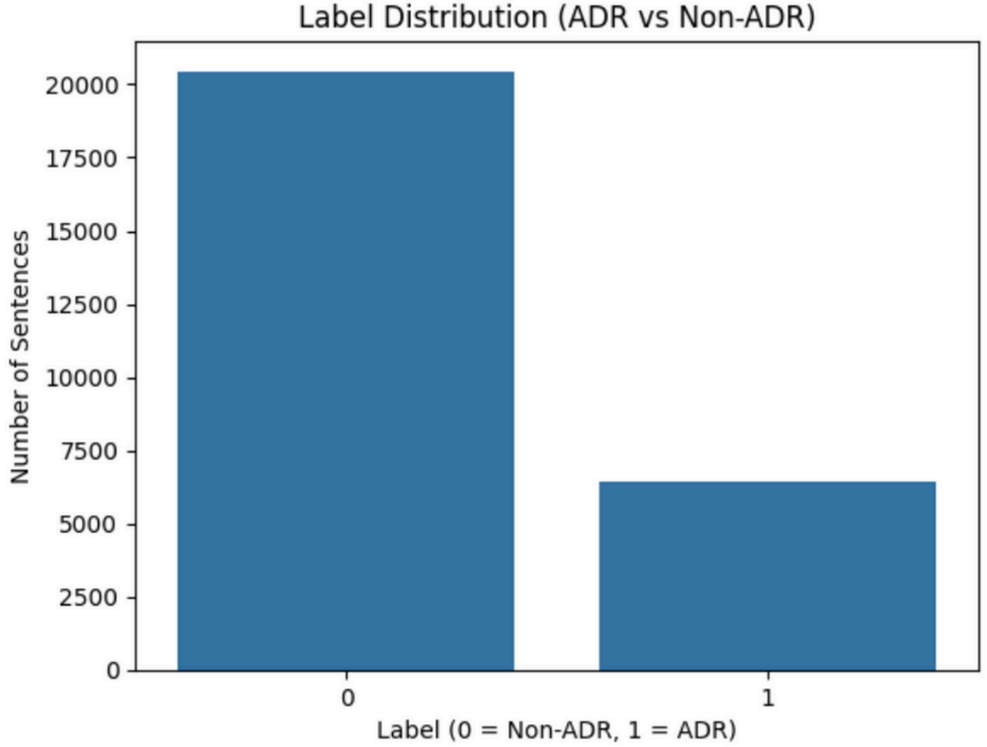
- 60/20/20 stratified train-dev-test split
- Performance → Test: Accuracy = 0.76, Precision = 0.73, Recall = 0.81, F1 = 0.77

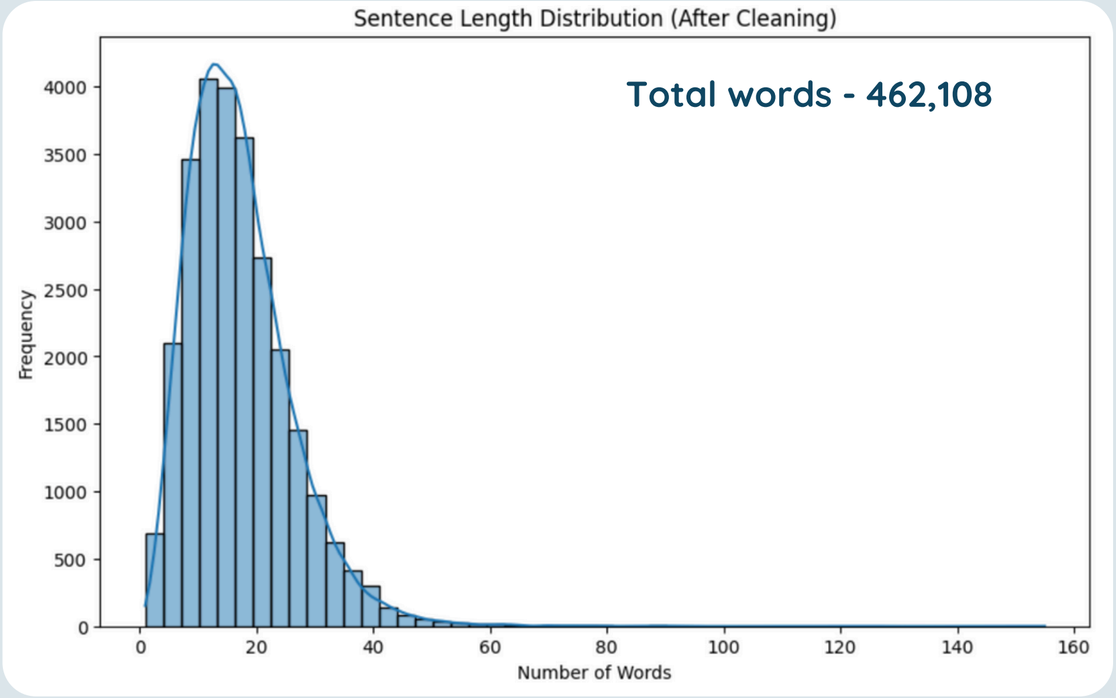
2. GPT-4o-mini Zero-Shot

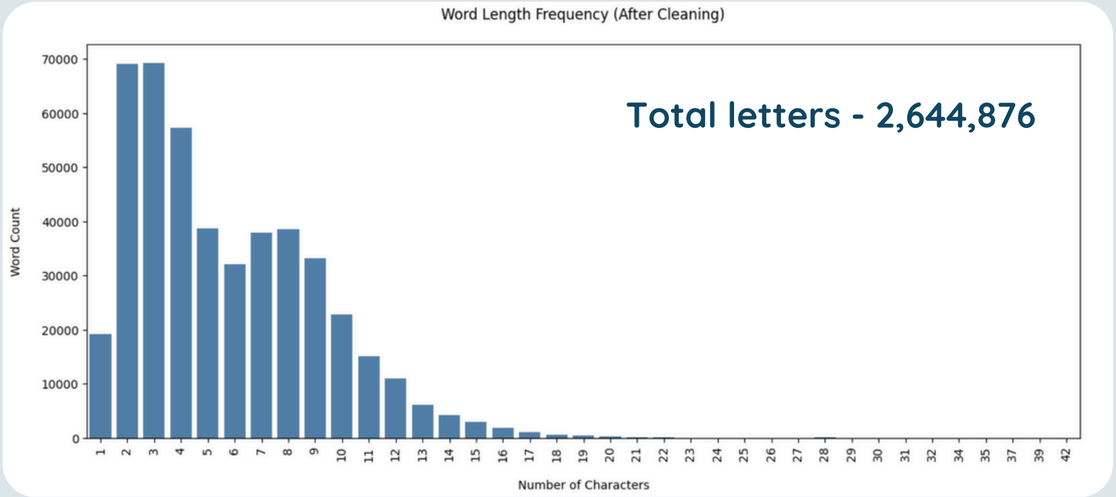
- Trained only on test data (20%)
- Performance → Accuracy = 0.83, Precision = 0.79, Recall = 0.90, F1 = 0.84

EDA Visualizations

	label	count	percentage
0	0	20436	76.06
1	1	6431	23.94







Insights & Recommendations

- From Simmering.dev (2025): ModernBERT models outperformed a few-shot LLaMA → **Our embedding models approach may outperform zero-/few-shot LLMs.**
- From ACL Anthology (2025): From ACL Anthology (2025): Advanced prompting (CoT, Self-Consistency) didn't always improve over basic prompting in open source LLMs with zero-shot → **Keep LLM prompting simple and controlled when relevant.**
- From SCITEPRESS (2025): GPT-4 and Llama showed better performance → **supports including them as a competitive LLM models, even without fine-tuning.**
- Mean sentence length = 17.2 words (Non-ADR has more outliers) → **Sentences are short enough to fit within BERT and LLM token limits.**
- Strong class imbalance (~76% Non-ADR) required downsampling → LLMs and BERT/InstructorXL models will be evaluated on the downsampled dataset.
- Duplicate rows were removed before & after cleaning.
- **Level of Sensitivity (Recall) is the primary metric** for our task - FN are more important than FP

Thank you