FarExStance: Explainable Stance Detection for Farsi

Introducing the first Farsi dataset designed for explainable stance detection, bridging the gap in resources for Farsi NLP. FarExStance provides a robust framework for automated claim verification and misinformation detection.

Introduction

- Challenge in Stance Detection: The rapid spread of misinformation has made automated stance detection an important task. Stance detection helps identify the position of a piece of text towards a claim, which is crucial for tasks like fact-checking.
- Gap in Resources for Farsi: While many stance detection datasets exist for English, there is a significant lack of resources for Farsi, especially for explainable stance detection, which provides reasoning along with the stance label.

Introducing the FarExStance Dataset

Scale

26,307 instances across news agencies, Twitter (X), and Instagram.

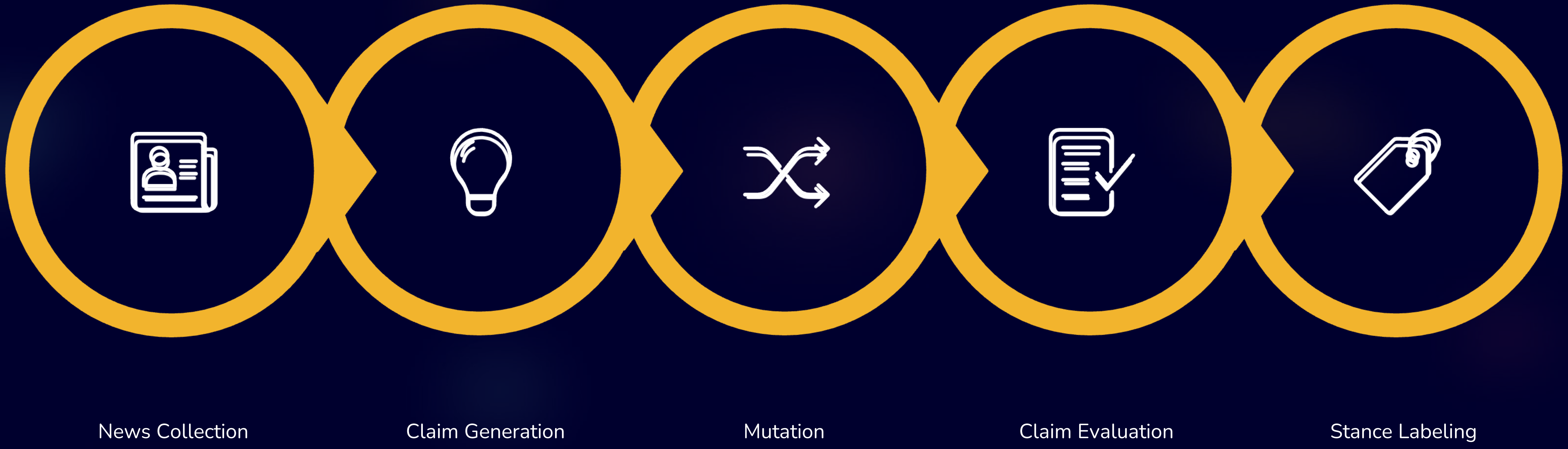
Unique Claims

5,874 unique claims manually curated and mutated for diversity.

Explanations

Includes extractive evidence for every stance label, providing a gold standard for explainability.

Data Collection & Annotation Process



News Collection

Claim Generation

Mutation

Claim Evaluation

Stance Labeling

Our manual process ensures high-quality data through expert pilot studies and a team of 16 native Farsi speakers.

Experimental Methodology

They benchmarked several state-of-the-art models using various learning paradigms to test the dataset's difficulty.

1

Fine-tuning

XLM-RoBERTa-Large and Aya-23-8B (using PEFT/QLoRA).

2

In-Context Learning

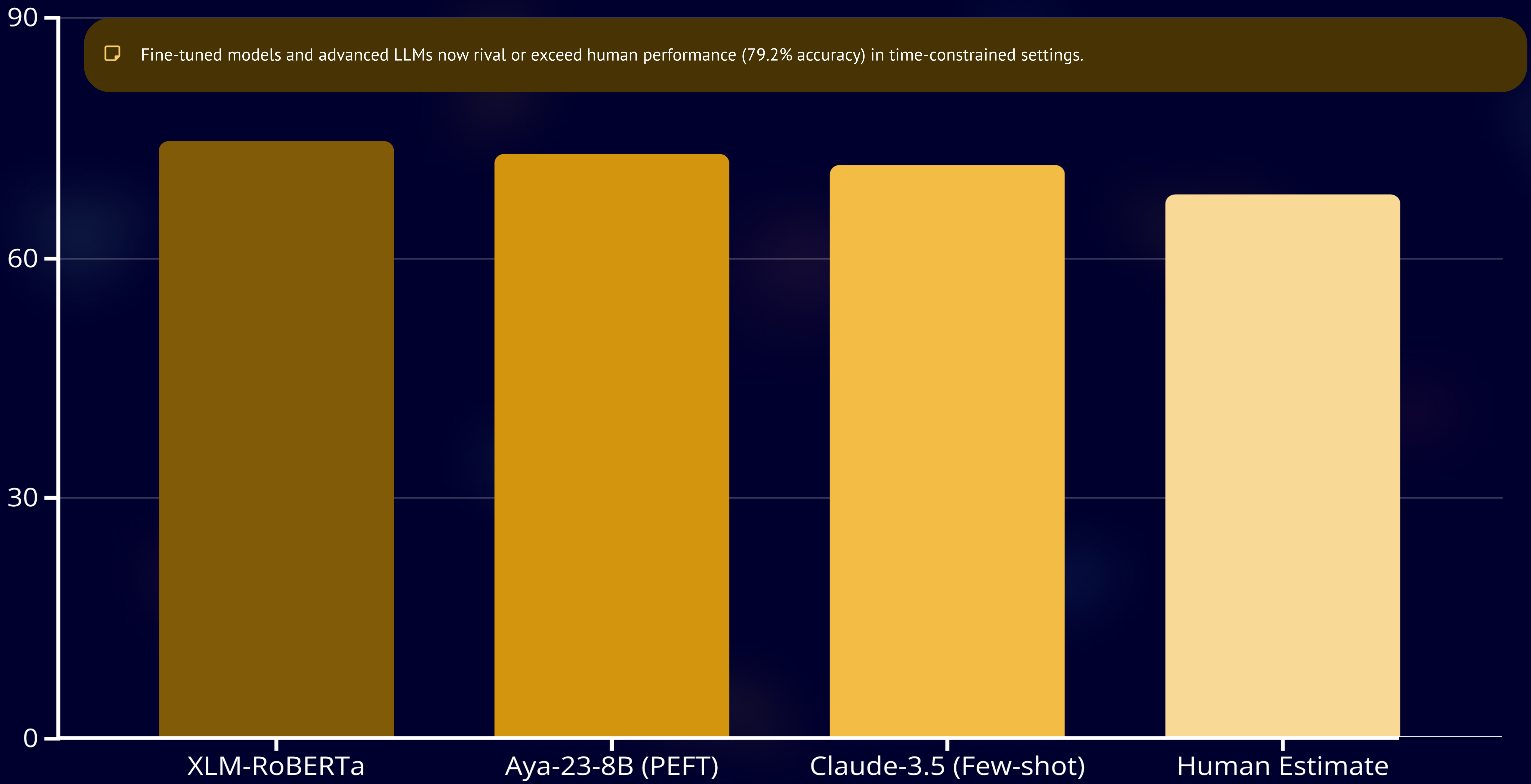
Zero-shot and few-shot prompting with GPT-4o and Claude-3.5-Sonnet.

3

RAG

Retrieval Augmented Generation to address hallucinations and improve evidence selection.

Stance Classification Performance



Fine-tuned models and advanced LLMs now rival or exceed human performance (79.2% accuracy) in time-constrained settings.

Evaluating Explanation Quality

Automatic Metrics

- **Aya-23-8B:** Highest ROUGE-L (17.5), aligning closest to reference text.
- **GPT-4o:** Best Global Coherence (82.6), ensuring logical consistency.

Human Evaluation (OES)

- **Claude-3.5-Sonnet:** Top LLM with 87.8 Overall Explanation Score.
- **Gold Standard:** Human explanations remain the benchmark at 88.5.

Error Analysis: Where Models Fail

Despite high accuracy, models struggle with specific nuances in Farsi stance detection.

The "Discuss" Challenge

The most difficult category for all models; neutral reporting is often misclassified.

Missed Details

Open-source models struggle to capture fine-grained details in explanations (up to 50% failure for Command-R).

Stance-Only Errors

Models sometimes provide a correct explanation but predict the wrong stance label.

Critiques and Limitations

Dataset Limitations:

→ ### Computational Constraints

Fine-tuning large models like Command-R-32B and Llama-3.1-70B was not possible due to hardware limitations.

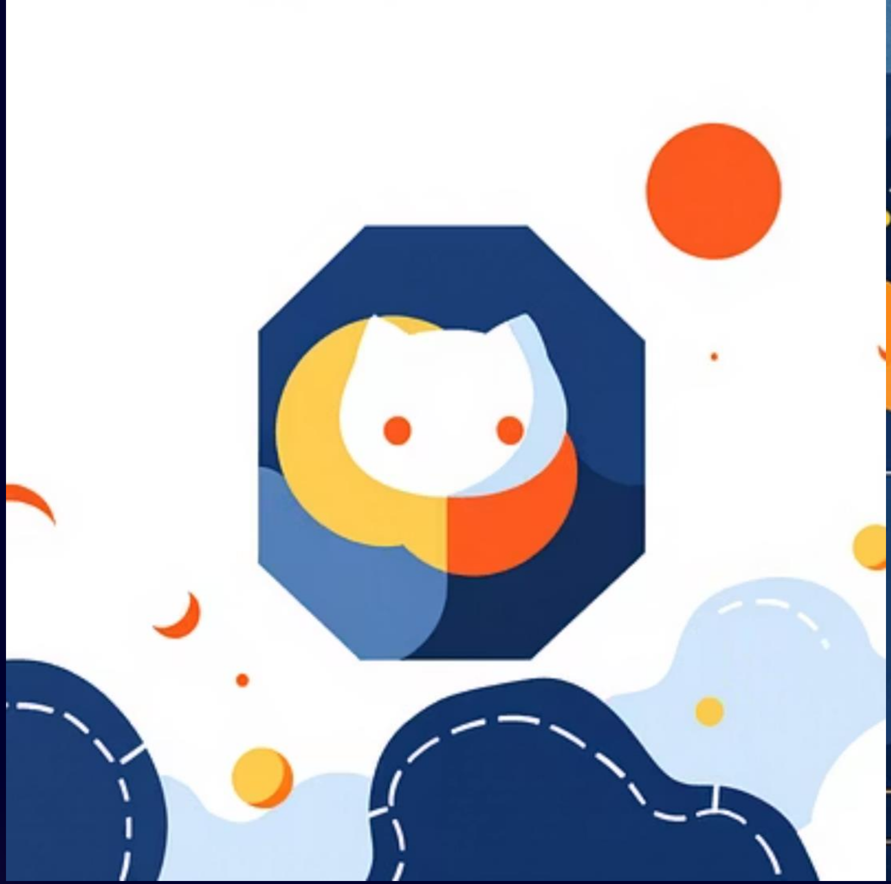
→ ### Limited LLMs Evaluated

Only a few closed-source models (Claude-3.5-Sonnet and GPT-4o) were tested. Expanding to more models could provide broader insights.

→ ### Social Media Domain:

The paper did not explore social media perspectives in-depth, leaving an opportunity for future research.

Key Contributions & Impact



- First Farsi dataset with extractive explanations.
- Publicly available resource to foster Farsi NLP research.
- Comprehensive benchmarks across SLMs and LLMs.

- Link to paper
- Github