We explain adversarial robustness from the holistic lens of training data

A Curious Case of Searching for the Correlation between Training Data and Adversarial Robustness of Transformer Textual Models

Cuong Dang,
Dung D. Le,
Thai Le

FPT Software AICenter

VINUNIVERSITY
College of Engineering and Computer Science

Indiana University Bloomington

ACL 2024

Motivation

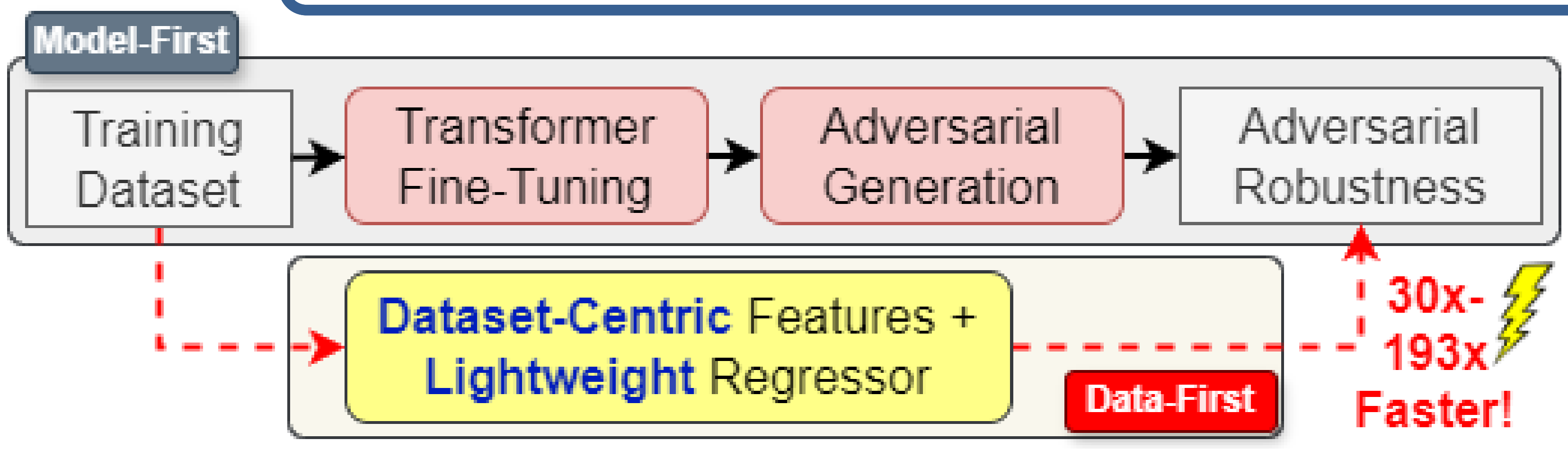
I am actively looking for a PhD position

How do **training data** correlate to **adversarial robustness**?

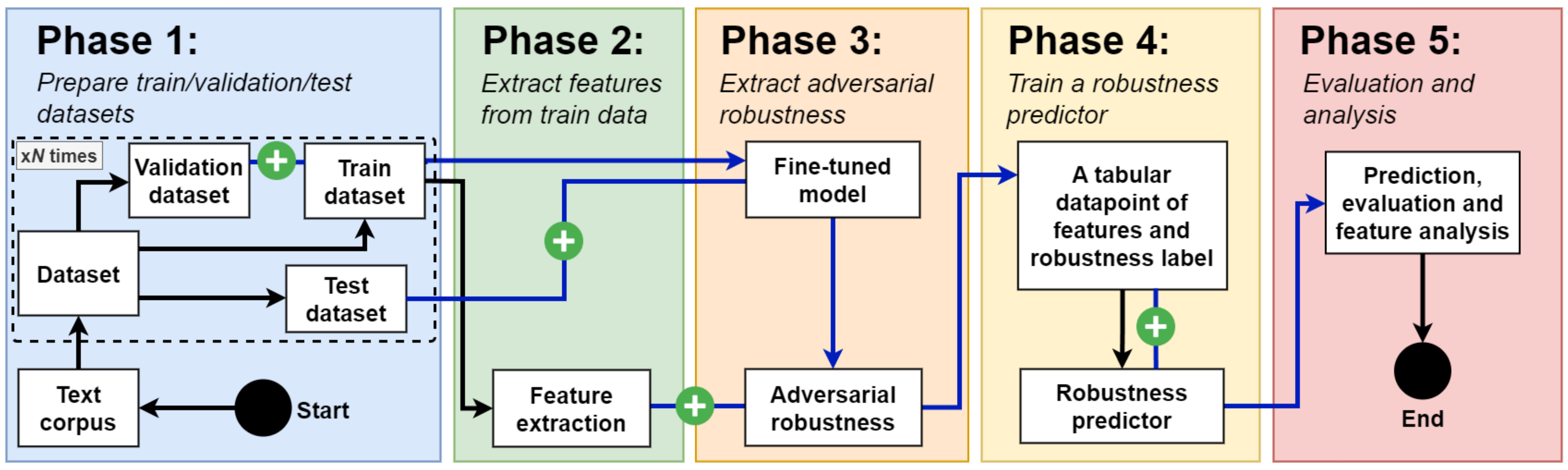
Can we **estimate the adversarial robustness** before models are fine-tuned without generating adversarial examples?

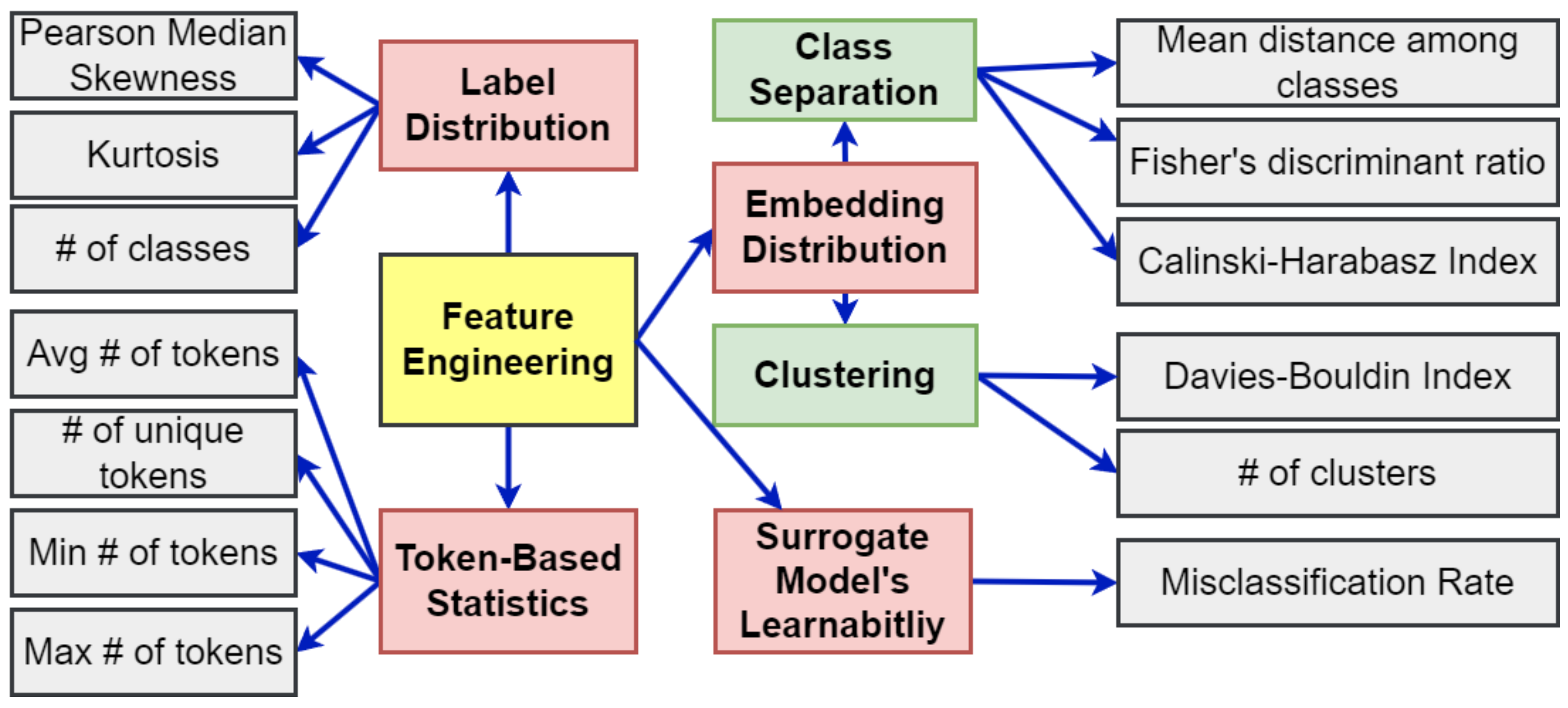
Propose an **interpretation framework**.

Introduce a **robustness predictor**.



Method


Phase 1:
Prepare train/validation/test datasets
xN times
Validation dataset
Train dataset
Dataset
Test dataset
Text corpus
Start
Phase 2:
Extract features from train data
Feature extraction
Phase 3:
Extract adversarial robustness
Fine-tuned model
Adversarial robustness
Phase 4:
Train a robustness predictor
A tabular datapoint of features and robustness label
Robustness predictor
Phase 5:
Evaluation and analysis
Prediction, evaluation and feature analysis
End

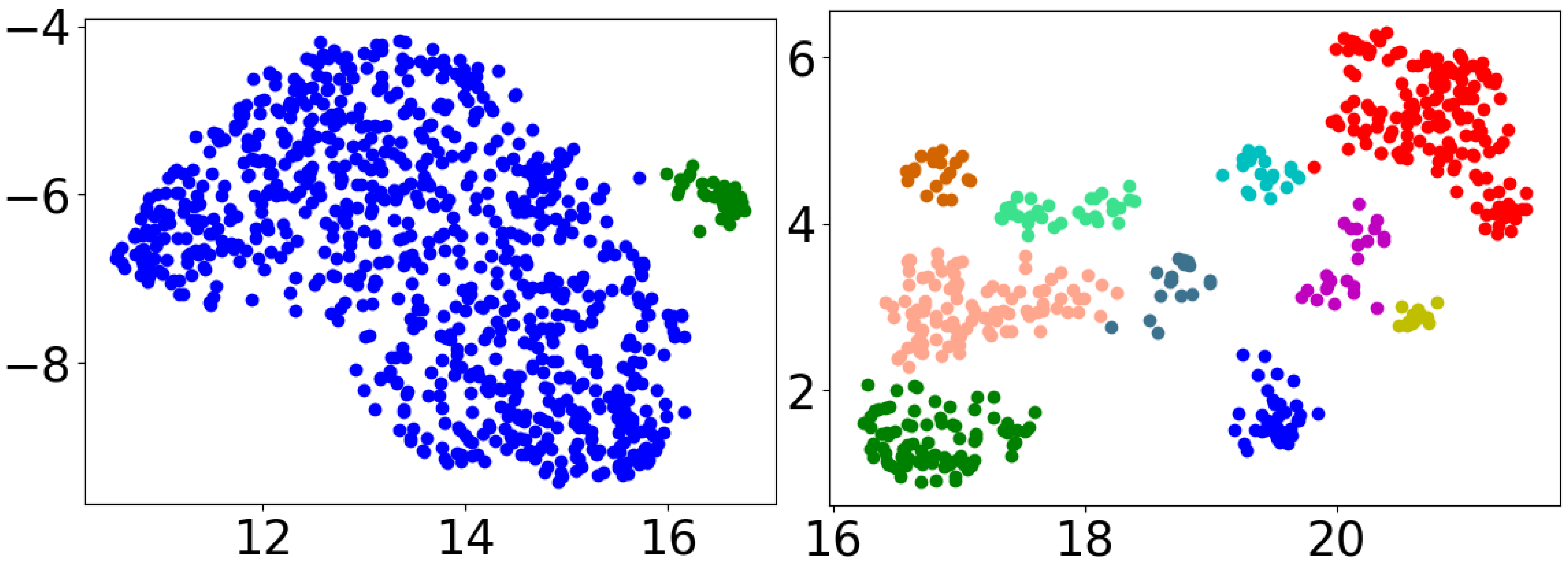


Main Results

Result 1: Fine-tuning data have a **strong correlation** with model robustness.

	METRIC	INTERPOLATION	EXTRAPOLATION
BERT	RMSE↓	0.055 ± 0.000	0.063 ± 0.001
	R^2↑	0.904 ± 0.005	0.885 ± 0.033
	MAE↓	0.037 ± 0.000	0.045 ± 0.000
	EVS↑	0.907 ± 0.005	0.908 ± 0.021
	MAPE↓	0.071 ± 0.000	0.102 ± 0.004
RoBERTa	RMSE↓	0.031 ± 0.000	0.061 ± 0.001
	R^2↑	0.972 ± 0.000	0.900 ± 0.019
	MAE↓	0.025 ± 0.000	0.044 ± 0.000
	EVS↑	0.972 ± 0.000	0.922 ± 0.010
	MAPE↓	0.048 ± 0.000	0.095 ± 0.004

Good performance in robustness prediction

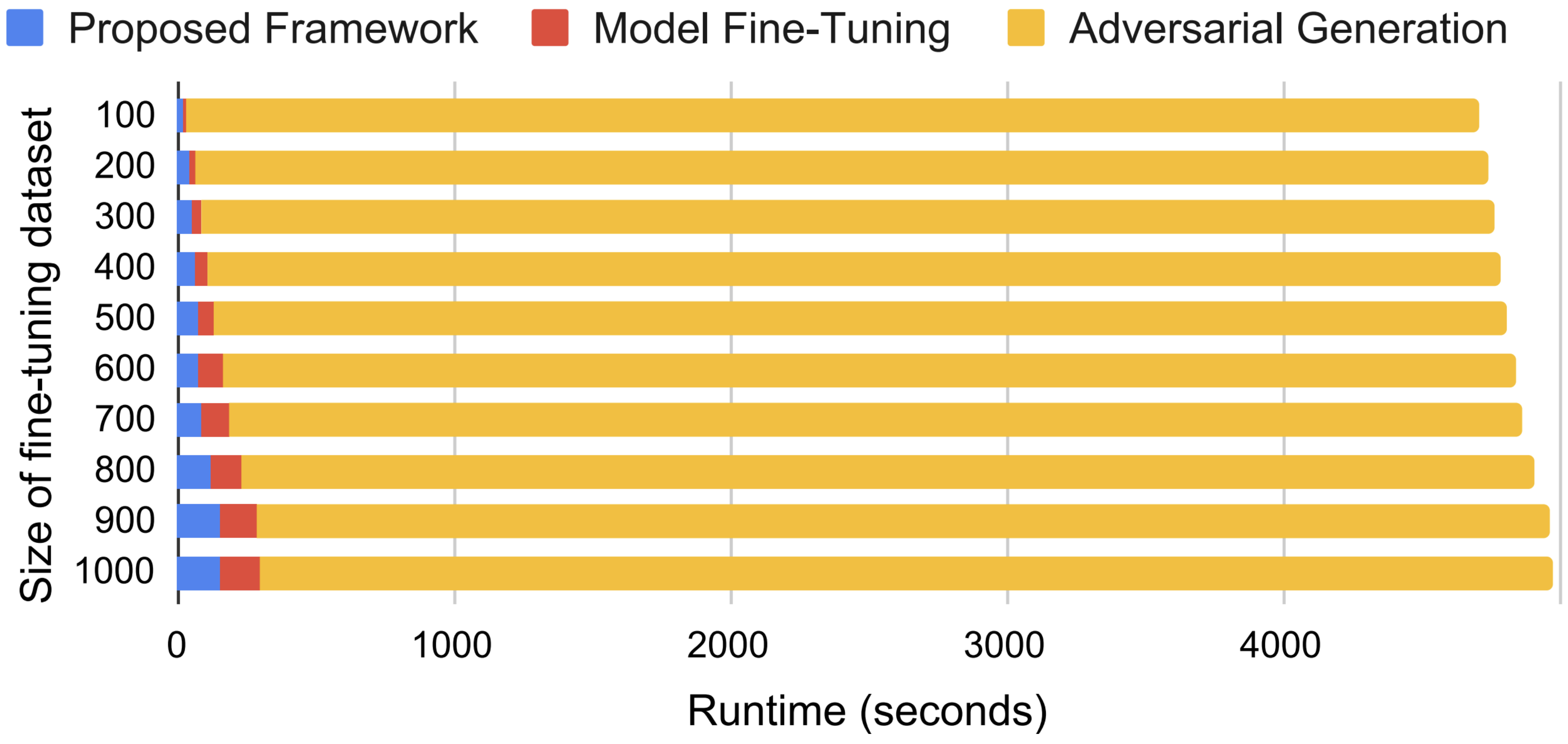


Models trained on data whose embedding is denser are more robust

Result 2: Interpretation framework can be used as a **robustness predictor**.

METRIC	BERT	Distil-BERT	RoBERTa	Distil-RoBERTa
RMSE↓	0.070	0.100	**0.061**	0.072
R^2↑	**0.806**	0.621	0.782	0.740
MAE↓	**0.045**	0.075	0.052	0.049
EVS↑	0.812	0.790	**0.918**	0.760
MAPE↓	0.145	0.173	0.139	**0.109**

Transferable between transformer models



Boost robustness evaluation from 30x to 193x