Self-Rationalization in the Wild: A Large-scale Out-of-Distribution Evaluation on NLI-related tasks

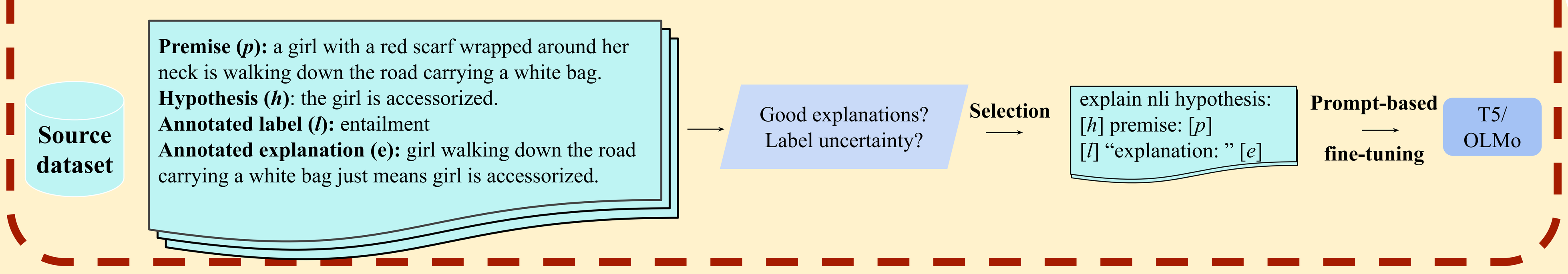
Jing Yang [1], Max Glockner[2], Anderson Rocha[1] and Iryna Gurevych[2]
[1]Artificial Intelligence Lab. (Recod.ai), Institute of Computing, University of Campinas, Brazil
[2] UKP Lab, Department of Computer Science, Technical University of Darmstadt, Germany

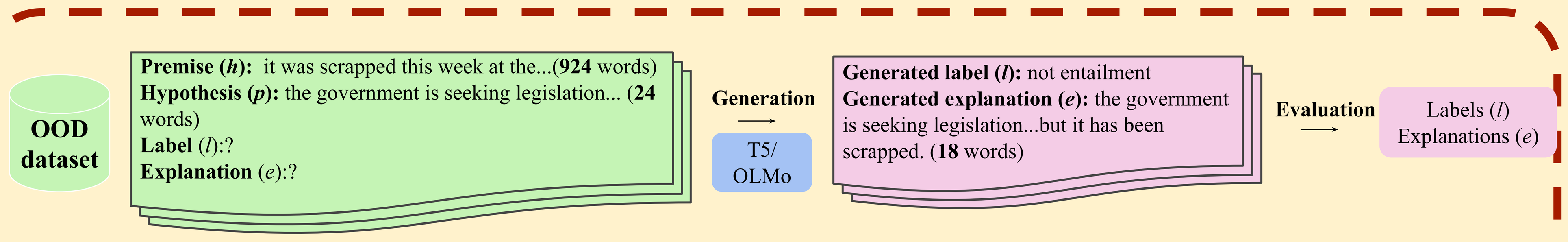
Motivation

- Datasets with explanations are scarce, raising challenges for learning and evaluation.
- Large-scale evaluation on large datasets with LLMs is computationally expensive, especially with long context input.

Self-Rationalization OOD Pipeline

Step **1: Learn to self-rationalize**



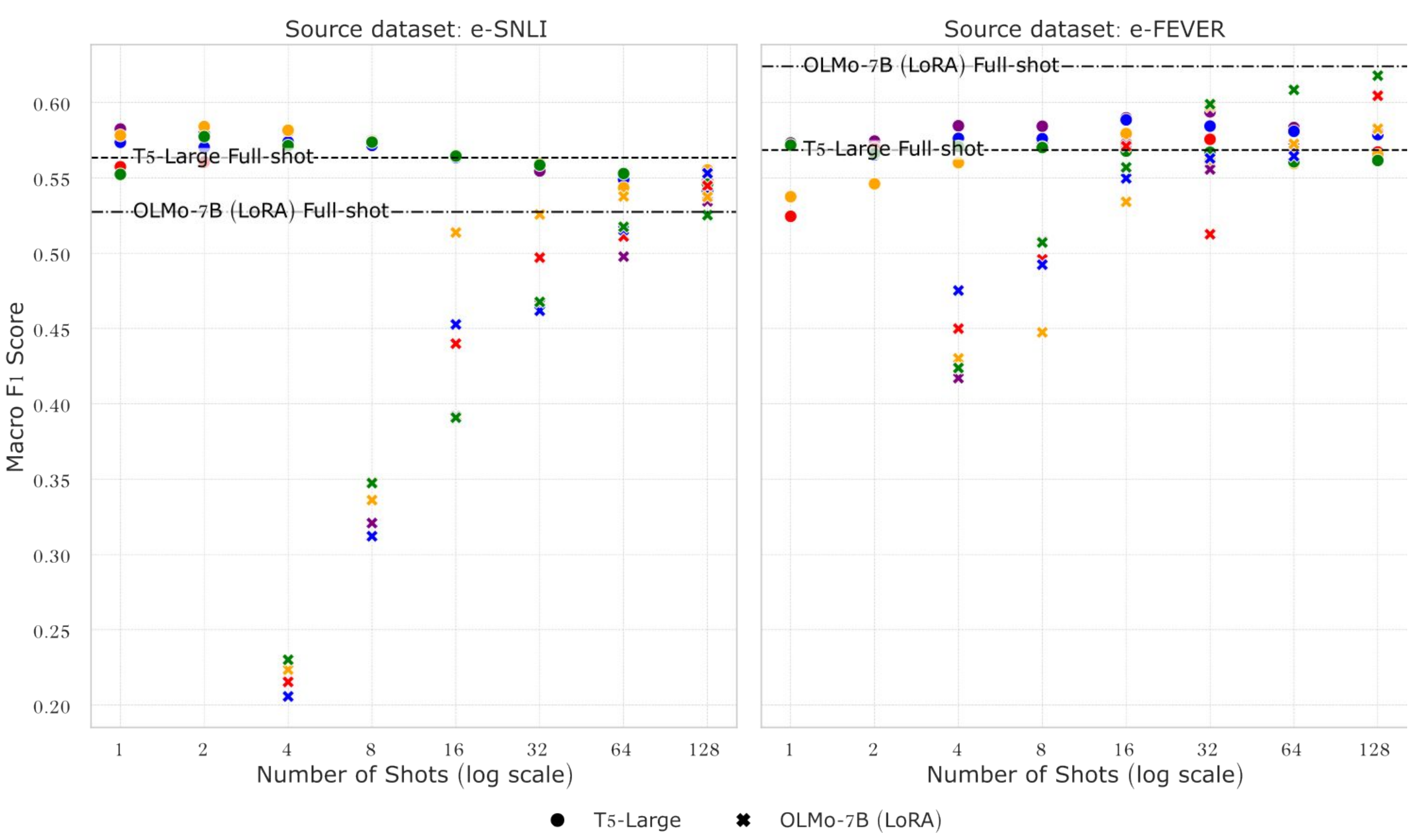


Step **2: Evaluate on OOD datasets**

OOD Datasets

Natural Language Inference	Fact-checking	Hallucination Detection of Abstractive Summarization
SICK, AddOneRTE, JOCI, DNC, MPE, HANS, WNLI, Glue Diagnostics, ConjNLI	Snopes, SciFACT, Climate-FEVER, VitaminC, COVID-Fact, FM2	FactCC, QAGS, XSUM Hallucination

Label Prediction Results



Fine-tuning on few-shot samples have comparable performance with Full-shot

Base model and source dataset has a large impact on label prediction performance

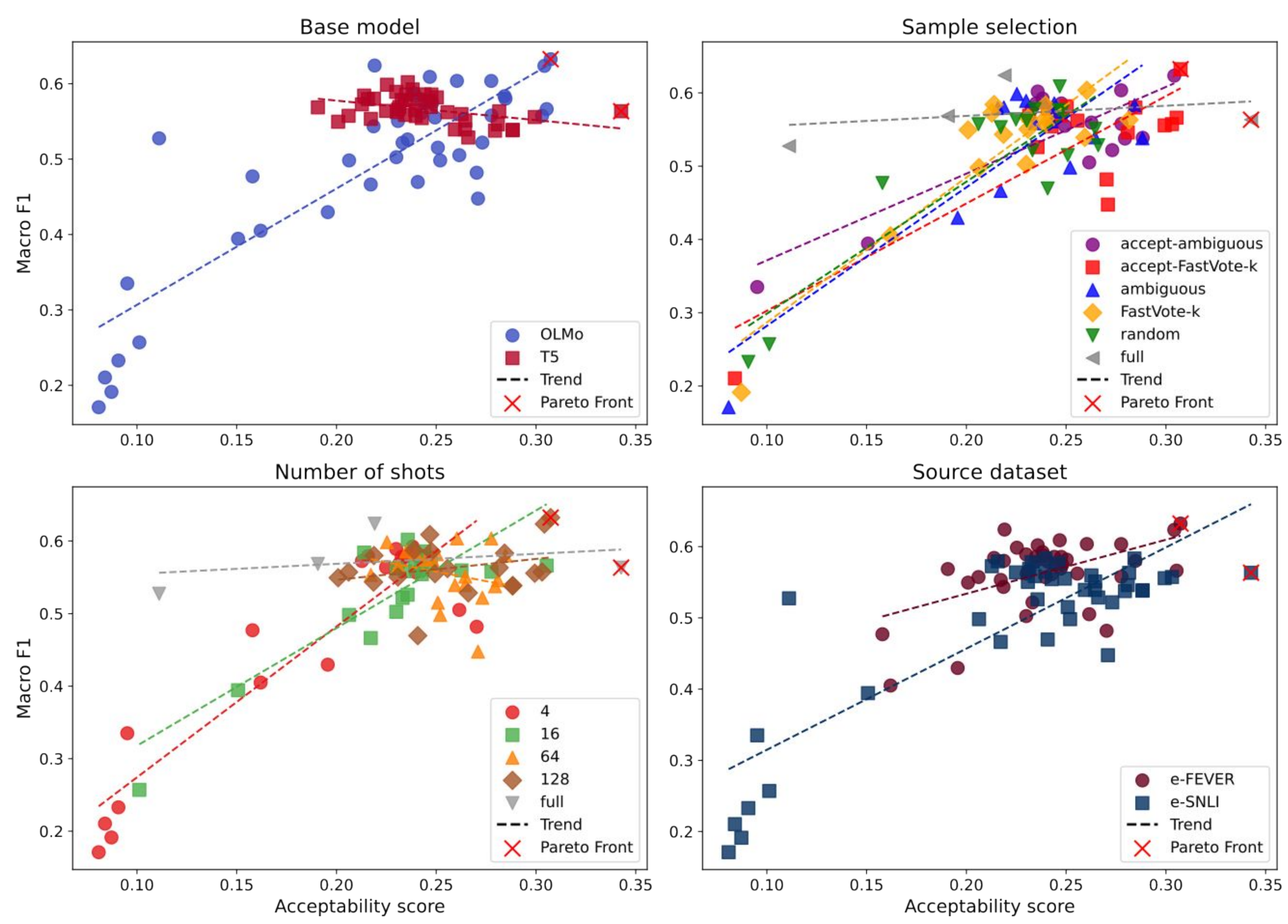
Automatic and Human evaluation

Automatic evaluation: Acceptability score, Themis, Auto-J, and TigerScore
Human evaluation: 468 crowd-workers from Prolific, judging the quality of explanations from 1,560 data instances
Table shows the correlation between the four metrics and human evaluation

Dataset	Auto-J	TigerScore	Themis	Accept.
SICK	-0,011	-0,220	0,400	**0,466**
VitaminC	0,163	-0,263	0,394	**0,469**
XSUM H.	0,223	-0,216	0,326	**0,475**
All	0,123	-0,219	0,387	**0,484**

Acceptability has the highest correlation with humans, while other three general LLM-as-a-judge evaluation metrics were not as reliable

Label Prediction vs Explanation Quality



Acceptability score (explanation quality) is positively related to Label Prediction Performance

Acknowledgement

FAPESP
CNPq
HORUS
Bundesministerium für Bildung und Forschung
ATHENE National Research Center for Applied Cybersecurity

Paper & Code