ABDN-NLP at CoMeDi Shared Task: Predicting the Aggregated Human Judgment via Weighted Few-Shot Prompting

Ying Xuan Loke[1], Dominik Schlechtweg[2], Wei Zhao[1]

University of Aberdeen[1], University of Stuttgart[2]

Paper in a Nutshell

Motivation: Human annotation in semantic proximity refers to how close or how far two usages of a word are in meaning is extremely subjective. It is also rather expensive and often results in a disagreement between the annotators.

Aim: This paper tackles the challenge by using large language models (LLMs) to automatically predict the aggregated human judgment of semantic proximity. It further proposes a weighted few-shot prompting strategy that factors in class importance and distribution.

Main results: The *weighted few-shot* method outperforms both zero-shot and standard few-shot approaches on average in the CoMeDi 2025 subtask 1, tested across 7 languages. It shows improved alignment with human annotations in predicting aggregated judgments of semantic proximity.

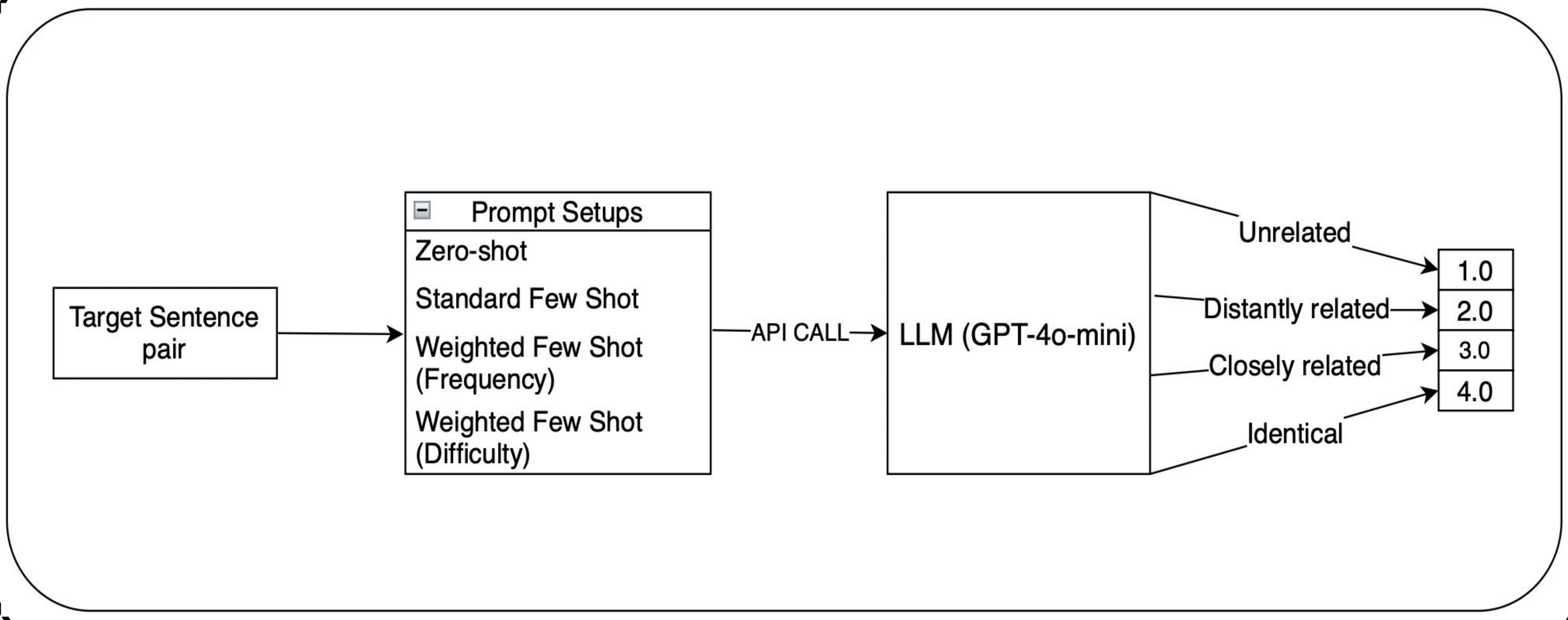
Research Questions

- How can we automatically predict the aggregated human judgment of semantic proximity between word usages?
- Does weighted few-shot prompting help with class imbalance and improve predictions?

Contributions

- Outline the characteristics of human judgment of semantic proximity (class imbalance in difficulty/frequency).
- Introduce a new few-shot method that gives higher weight in the prompt to more difficult or more frequent classes.
- Discuss the results (e.g., GPT-4o-mini struggles with Norwegian and Chinese) and limitations of our approach.

System Workflow



Results – Subtask 1

Setup	Russian	Swedish	Spanish	Norwegian	English	German	Chinese	Avg
zero-shot (n=0)	0.504	0.351	0.491	0.207	0.610	0.529	0.026	0.388
standard few-shot (n=20)	0.423	0.441	**0.587**	0.197	**0.626**	0.675	-0.127	0.403
weighted few-shot (frequency, n=20)	0.478	**0.509**	0.569	**0.431**	0.625	0.673	**0.209**	**0.499**
weighted few-shot (difficulty, n=20)	**0.512**	0.389	0.543	0.183	0.600	**0.690**	-0.056	0.408
deep-change (Kuklin and Arefyev, 2025)	**0.623**	**0.675**	**0.748**	**0.668**	**0.732**	**0.723**	**0.424**	**0.656**
comedi-baseline (Schlechtweg et al., 2025)	0.112	0.018	0.175	0.124	0.102	0.274	0.059	0.123

Limitations

- When the proportion of classes differs considerably between the test set on the one hand and training and development data on the other, the weighted strategy is likely to lose a part of its effectiveness.
- Our approach is based on a single LLM, which is not representative of the broader LLM community. Therefore, our findings may differ when other LLMs are applied.

Paper link

Github link

CoMeDi @ COLING 2025 Abu Dhabi UAE