

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

From Spiky to Spectral: Aligning LLM Confidence with Human Uncertainty for Psychological Defense Mechanism Classification

Permalink

<https://escholarship.org/uc/item/6j94x12t>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Chen, Chuzi

Li, Hao

Xiao, Yankang

et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

From Spiky to Spectral: Aligning LLM Confidence with Human Uncertainty for Psychological Defense Mechanism Classification

Chuzi Chen

Wuhan University, Wuhan, Hubei, China

Hao Li

Wuhan University, Wuhan, Hubei, China

Yankang Xiao

Wuhan University, Wuhan, Hubei, China

Hao Jiang

Wuhan University, Wuhan, Hubei, China

Abstract

Psychological defense mechanisms serve as unconscious cognitive strategies aimed at regulating anxiety and protecting the self. These processes are not limited to clinical settings but are also prevalent in social media interactions. While Large Language Models (LLMs) can capture semantic ambiguity, a structural misalignment arises as their decoding patterns yield spiky probability distributions with high confidence, failing to reflect the uncertainty in the spectral nature of human cognition. To bridge this gap, we propose DefenseAlign for psychological defense mechanism classification, aligning LLM confidence from pre-softmax logits with human uncertainty represented by a human belief distribution. DefenseAlign pools logits across label verbalizations, applies learnable temperature scaling to fit the human belief distribution, and distills the aligned distributions into a student model. We evaluate the framework on DefenseSpectrum, a dataset that estimates human uncertainty directly from annotation frequencies. Experiments show reduced divergence and improved preservation of secondary mechanisms.