Emotions Where Art Thou: Understanding and Characterizing the Emotional Latent Space of Large Language Models

Benjamin Reichman
Georgia Institute of Technology
bzr@gatech.edu

Adar Avsian
Georgia Institute of Technology
aavsian3@gatech.edu

Larry Heck
Georgia Institute of Technology
larryheck@gatech.edu

Abstract

This work investigates how large language models (LLMs) internally represent emotion by analyzing the geometry of their hidden-state space. Using a synthetic dataset of emotionally rewritten sentences, we identify a low-dimensional emotional manifold via singular value decomposition and show that emotional representations are directionally encoded, distributed across layers, and aligned with interpretable dimensions. These structures are stable across depth and generalize to eight real-world emotion datasets spanning five languages. Cross-domain alignment yields low error and strong linear probe performance, indicating a universal emotional subspace. Within this space, internal emotion perception can be steered while preserving semantics using a learned intervention module, with especially strong control for basic emotions across languages. These findings reveal a consistent and manipulable affective geometry in LLMs and offer insight into how they internalize and process emotion.