Do Machines Think Emotionally? A Cognitive Appraisal Analysis of Large Language Models

Anonymous Submission

1 Introduction and Motivation

Emotional intelligence is central to human life, and equipping LLMs with coherent emotional reasoning is critical for their use as conversational agents, simulators in human-subject research, and models of cognition. Yet, most existing approaches remain limited to surface-level, categorical prediction of emotions from stimuli, which often fail to generalize to novel, ambiguous, or culturally nuanced scenarios. To build systems that engage in true emotional reasoning we focus on using the cognitive appraisal theory [1].

Prior work leveraging appraisal theory has mostly studied other-appraisal [3], often restricted to a few coarse cognitive. Other-appraisal has the caveat of confounding assumptions about external agents' demographics, not reflecting true emotional reasoning. Thus, a lack of systematic analysis remains regarding how LLMs internally represent emotions through their own cognitive structures.

We introduce CoRE (Cognitive Reasoning for Emotions), the first large-scale benchmark of self-appraisals for emotions. CoRE spans 15 emotion categories, 16 appraisal dimensions, and an analysis across 7 models.

2 Dataset and Experimental Setup

We construct a benchmark dataset of emotion-related scenarios inspired by [2], covering 15 emotions (e.g., Happiness, Pride, Fear, Sadness) and 16 appraisal questions targeting 8 core cognitive dimensions. A total of 308 high-quality scenarios and 4,928 prompts are created, requiring LLMs to provide appraisal ratings for each dimension. We evaluate seven models: DeepSeek R1 (671B), GPT-o4-mini, Gemini 2.5 Flash, LLaMA 3 (8B), Phi 4 (14B), Qwen 3 (32B), and Qwen QwQ (32B), on two tasks: (i) open-ended emotion identification and (ii) appraisal responses (open-text plus numerical ratings).

3 Main Results

Through three specific themes of experiments, we find that **LLMs are broadly similar to humans in making plausible connections between different cognitive dimensions and emotions, while also showing emotion and cognitive dimension-specific nuances and idiosyncrasies.** We uncover the latent cognitive dimensions, using PCA, that are used implicitly by each model and find that pleasantness, effort, and agency (responsibility) are the three most important factors across most models (and humans). We then study which cognitive dimensions are the best predictors of different emotions. LLMs again show coherent associations, with some nuances – for example, anger is predicted the best by dimensions of legitimacy, as opposed to valence. Finally, we study the appraisal rating distributions, comparing across emotions for a single model, and across different models. We find that **intra-model representations** for emotions are **coherent and similar** across most models, with Gemini 2.5 Flash showing inconsistencies in representing abstract emotions like Hope and Interest. **Inter-model comparisons**, however, show that models represent emotions with **significantly different distributions**, highlighting that there exists **no concept of a universal emotion representation** across popular models.

References

[1] Klaus R Scherer. On the nature and function of emotion: A component process approach. In *Approaches to Emotion*, pages 293–317. Psychology Press, 2014.

[2] Craig A Smith and Phoebe C Ellsworth. Patterns of Cognitive Appraisal in Emotion. *Journal of Personality and Social Psychology*, 48(4):813, 1985.

[3] Ala N. Tak, Jonathan Gratch, and Klaus R. Scherer. Aware yet biased: Investigating emotional reasoning and appraisal bias in large language models. *IEEE Transactions on Affective Computing*, pages 1–11, 2025.