

The Urgency Axis: Mapping Crisis Severity Representations in Large Language Models

Wang, Qi (School: Ghana International School)

Omane Boamah, Yaw (School: Ghana International School)

Language models are increasingly explored for classifying online messages during disasters to route resources, however interpretability is demanded for opaque models. No prior work has characterized how crisis urgency is geometrically encoded within LLM internal representations. We investigate the internal geometry of crisis representations in LLMs. We trained logistic regression probes on hidden-state activations from ~8,000 HumAID disaster tweets spanning 10 categories across 19 real disasters, evaluated layer-wise encoding patterns, and tested cross-disaster transfer. The PCA reveals a dominant first component (23.4% of variance) that is strongly correlated with human urgency ratings (Spearman $\rho=0.713$, $p=0.01$). This Urgency Axis distinguishes high-urgency and low-urgency tweets with a large effect size (rank-biserial $r=0.675$) but cannot fully resolve fine-grained categories (10-class $F1=0.841$). Layer-wise probing shows crisis information rapidly emerges in earlier layers, peaks in middle transformer layers following a slight decline. Cross-disaster transfer degrades only ~6%, indicating slight disaster-invariant representations. Role cosine-similarities reveal interpretable clustering among professional content. We conclude that crisis urgency is encoded as a coarse linear direction - reliable for binary severity discrimination but insufficient for operational triage. This representational factorization between role-urgency and content-urgency has direct implications for auditing humanitarian AI systems, establishing that persona-based interpretability methods transfer meaningfully but not completely to content classification.

Awards Won: