Beyond Testers' Biases: Guiding Model Testing with Knowledge Bases using LLMs

Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace A. Lewis, Christian Kästner, Tongshuang Wu

Motivation
Coarse-grained metrics over benchmarks cannot reveal nuanced model behaviors.
Model testing aims to explore nuanced behaviors, but is often conducted
ad-hoc and biased by testers.

Users tend to explore locally overfit to their intuition, domain knowledge, confirmation bias.
Expect: Comprehensive testing More systematically cover the space

Introduce requirements elicitation for systematic model testing.

TL;DR: Weaver systematizes ad-hoc model testing with requirements elicitation, by extracting knowledge bases from LLMs for any tasks.

Workflow

Interactive exploration of knowledge base generated by LLM for any tasks.

Weaver Workflow

Weaver helps users explore model behaviors systematically and find model bugs beyond their biases.

Intrinsic Evaluation
Weaver covers 90.6% existing concepts.

Weaver helps users test more diverse (+47.7%) concepts and find bugs beyond their biases.

Experiment setup: We conducted a within-subject controlled experiment (N=20), where users test ChatGPT on two stance detection tasks.

Findings:

Weaver helps users explore model behaviors systematically and find model bugs beyond their biases.

User Evaluation

Case Studies
Weaver helps LLM practitioners test (and iterate) their prompts.

Checkout our paper! Try our tool!