Evaluating Vision Language Models Through Concept Hacking

Yijiang Li
Electrical and computer engineering, La jolla, California, United States

Bingyang Wang
Emory University, Atlanta, Georgia, United States

Tianwei Zhao
Johns Hopkins University, Baltimore, Maryland, United States

Qingying Gao
Computer Science, Baltimore, Maryland, United States

Hokin Deng
Harvard University, Boston, Massachusetts, United States

Dezhi Luo
University of Michigan, Ann Arbor, Michigan, United States

Abstract

Evaluating the cognitive abilities of Vision-Language Models (VLMs) is challenging due to their reliance on spurious correlations. To distinguish shortcut-taking from genuine reasoning, we introduce Concept Hacking, a paradigm manipulating concept-relevant information to flip the ground-truth but preserving concept-irrelevant confounds. For instance, in a perceptual constancy test, models must recognize that a uniformly wide bridge does not narrow in the distance; the manipulated condition using concept hacking altered the bridge to actually taper. We assessed 209 models across 45 experiment pairs spanning nine low-level cognitive abilities, encompassing all five core knowledge domains. Comparing performance on manipulated versus standard conditions revealed that models fell into shortcut-reliant or illusory understanding types, with none approaching human-level performance. Models of varying sizes appear in each category, indicating that scaling neither imparts core knowledge nor reduces shortcut reliance. These findings highlight fundamental limitations in current VLMs, reinforcing concerns about their ability to achieve genuine understanding.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*