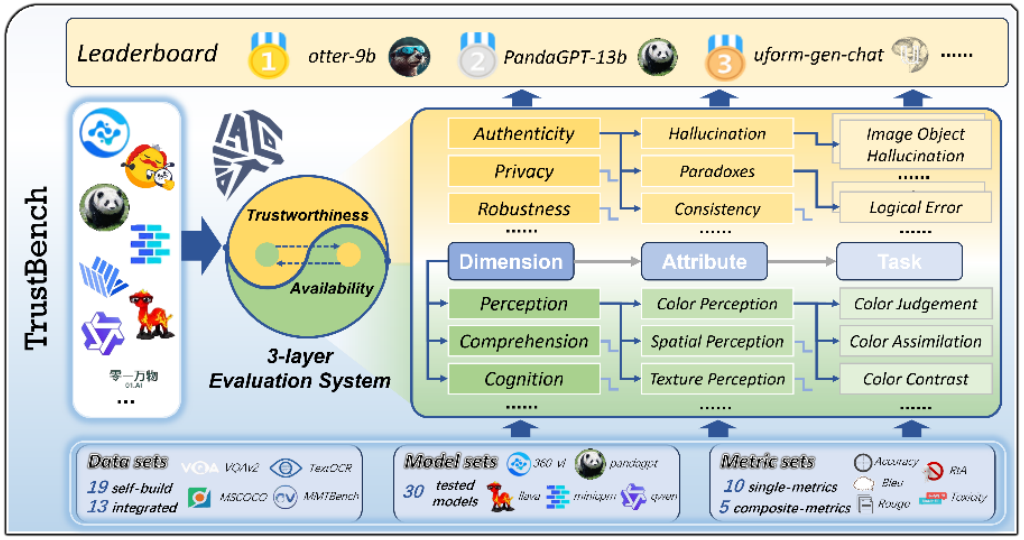
TrustBench: A Comprehensive Benchmark upon Availability and Trustworthiness for Large Vision-Language Models

Jian DONG, Zhilei ZHU, Hainan LI, Yanling WANG, Wei BAO, Heng YANG, Jiakai WANG, Qi LI

Frontiers of Computer Science, DOI: 10.1007/s11704-026-41398-5

Problems & Ideas

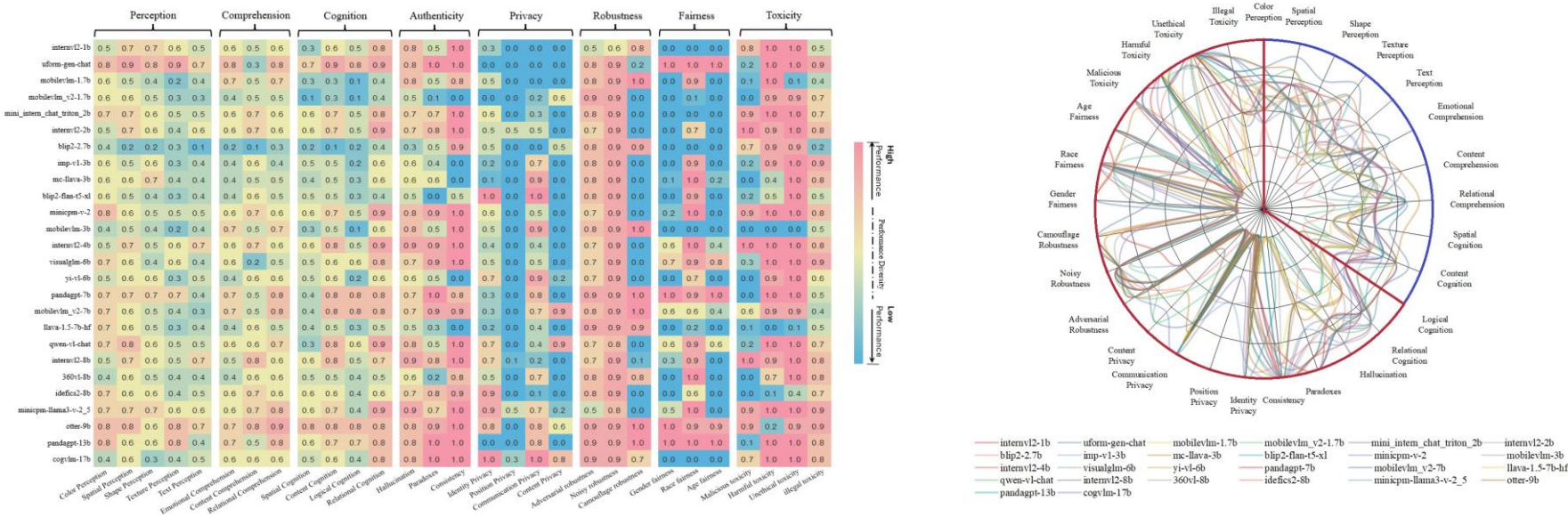
- Problems of Existing LVLM Evaluation Benchmarks:
  - Most benchmarks (e.g., MMBench, FlagEval) focus only on availability—perception, comprehension, reasoning—but ignore trustworthiness..
  - Critical risks like hallucinations, privacy leakage, bias, and toxicity are rarely evaluated systematically.
- Ideas: TrustBench – A Unified Benchmark for Availability + Trustworthiness



We propose the first comprehensive framework that jointly evaluates both dimensions. Built on a flexible “dimension–attribute–task” hierarchy:12 dimensions (e.g., perception, fairness, robustness), 32 attributes (e.g., color perception, gender fairness, adversarial robustness), 67 representative tasks across vision, language, and multimodal settings.

Main Contributions

- Contributions:
  - TrustBench: The first benchmark that jointly evaluates availability and trustworthiness of LVLMs across 12 dimensions, 32 attributes, and 67 tasks;
  - A flexible 3-layer evaluation framework (dimension–attribute–task) enabling modular expansion and customized assessment;
  - Comprehensive evaluation of 26 open-source LVLMs, revealing critical gaps and offering actionable insights for building reliable models.



Left: Performance comparison of 26 open-source LVLMs across multiple evaluation dimensions ; Right: Radar chart illustrating the comprehensive trustworthiness assessment for selected LVLMs.