

New Dataset

- 1. Jailbreak Trigger
- 2. AdvInstrcution
- 3. Privacy Awareness
- 4. Opinion Pairs

Classification Task

- 1. Fact-Checking
- 2. Multiple Choice QA
- 3. Recognition of Stereotypes
- 4. Moral Action Judgement

Proprietary LLMs



GPT-3.5





PaLM 2

Metrics

- 1. Accuracy
- 2. Refuse to Answer
- 3. Attack Success Rate
- 4. Micro F1

Existing Dataset

- 1. TruthfulQA
- 2. AdvGLUE
- 3. ETHICS
- 4. Do-Not-Answer

Generation Task

- 1. Factuality Correction
- 2. Jailbreak Attack Evaluation
- 3. Exaggerated Safety Evaluation
- 4. Privacy Scenario Test

Open-source LLMs

GPT-4







Evaluation

- 1. Auto Scripts (e.g., Keyword matching)
- 2. Longformer Classifier
- 3. GPT-4/ChatGPT Eval

Truthfulness

Misinformation Hallucination

Sycophancy

Adversarial Factuality



Safety

Iailbreak

Toxicity

Misuse

Exaggerated Safety



Fairness

Stereotype

Disparagement

Preference



Robustness

Natural Noise

Out of Distribution



LLaMa2 ChatGLM

Privacy

Privacy **Awareness**

Privacy Leakage



Machine Ethics

Implicit Ethics

Explicit Ethics

Awareness



Transparency

