LITERATURE SURVEY ON AI SAFETY EVALUATION METHODS

Vilash Naveen
(210413G)

University of Moratuwa

1 INTRODUCTION

The transformative capabilities of large language models (LLMs) have made them central to research, industry, and society. However, their increasing scale and deployment raise urgent questions about safety and alignment. Beyond measuring accuracy or task competence, it is critical to ensure that LLMs do not generate harmful outputs, mislead users, or increase bias, and that they remain aligned with human values and objectives. Failures in these areas can lead to harms, including misinformation, discrimination, privacy breaches, and malicious misuse.

Recent years have seen an significant increase in evaluation efforts designed to address these concerns. While some focus narrowly on specific risks, others attempt holistic coverage across multiple dimensions. These evaluations vary widely in scope, methodology, and openness, leaving users with little consensus on how best to measure alignment and safety.

This survey provides a structured overview of existing evaluation methods for AI safety and alignment. Our contributions are as follows:

1. We present a taxonomy of evaluation approaches, organizing them along dimensions such as risk domains, adversary models, and metric types.
2. We compare prominent frameworks—including **HELM, BIG Bench, TruthfulQA, RealToxicityPrompts, Anthropic red teaming**, and **OpenAI Evals**, highlighting their strengths and limitations.

2 BACKGROUND

In the context of LLMs, **Safety** refers to the prevention of unintended harmful outputs or behaviors, often framed in terms of minimizing harmful outcomes and adhering to normal or regular standards. **Alignment**, by contrast, addresses whether a model's goals and behaviors match human intentions and values, frequently ensured by the principles of being "helpful, honest, and harmless." Truthfulness is a core component of alignment, ensuring that models avoid propagating falsehoods or hallucinations gained from training data. The threat model for LLMs encompasses a wide spectrum of risks. These include:

– **Content safety**, such as preventing offensive, violent, or extremist text.
– **Deception**, including misinformation, disinformation, and manipulative narratives.
– **Data exfiltration**, involving the leakage of personally identifiable information (PII) or sensitive training data.
– **Malicious misuse**, where LLMs are used to facilitate fraud, disinformation campaigns, or other harmful activities.

Evaluation methods have accordingly diversified. Capability benchmarks like BIG Bench, BIG Bench Hard (BBH), and HELM test general reasoning abilities. Safety focused benchmarks (e.g., RealToxicityPrompts, CivilComments, BBQ, TruthfulQA) quantify specific harms such as toxicity, bias, or untruthfulness. Red teaming and adversarial testing probe models for vulnerabilities, as demonstrated by Anthropic's large scale red teaming efforts and automated adversarial techniques. Process and behavioral compliance evaluations assess whether models adhere to refusal guidelines or resist jailbreaks, often reinforced through RLHF. Reliability and robustness evaluations examine calibration, consistency, and performance under perturbations, while system level and agentic evaluations test safety in multistep or tool assisted environments such as Safety Gymnasium. Across all domains, human in the loop and governance linked audits remain indispensable for interpreting outputs, setting norms, and establishing safety standards.

3 TAXONOMY OF EVALUATION METHODS

LLM evaluation can be understood as spanning multiple fronts, which together provide a taxonomy for comparing frameworks.

Risk domains address specific harms, including harmful content, bias and fairness, misinformation and hallucination, privacy and IP leakage, jailbreak vulnerability, and autonomy risks in agentic settings. Frameworks vary in their emphasis: for instance, TruthfulQA targets misinformation, BBQ focuses on demographic bias, Anthropic's red teaming probes offensive and adversarial outputs, while HELM covers toxicity, fairness, copyright, and more.

Modalities differ by data type. While most evaluations are text based, specialized frameworks address code reasoning (e.g., HumanEval, APPS) and multimodal tasks, including vision based SafeRL in Safety Gymnasium and red teaming of image generation models like DALLE 2.

Adversary models range from benign prompts to adversarial queries crafted manually, scripted perturbations, or LLM generated attacks. Some frameworks emphasize multiturn and multiagent scenarios, reflecting real world adversarial elements.

Metrics vary with evaluation goals. Common measures include accuracy, F1 score, and pass@k for task performance, toxicity and refusal rates for safety, jailbreak success rates for robustness, calibration metrics (ECE, Brier scores), human ratings for truthfulness or harmfulness and efficiency measures such as runtime or computational cost.

Granularity spans different levels of assessment: response level (individual completions), conversation level (multiturn interactions), task level (collections of prompts), and system level (agents operating in environments).

Automation and openness further distinguish frameworks. Some rely heavily on manual human red teaming, while others (e.g., HELM, OpenAI Evals, Safety Gymnasium) provide modular toolkits for reproducible, large scale testing. Open source datasets and harnesses enhance transparency and community use, though access to proprietary models and data often constrains reproducibility.

Together, these dimensions provide a structured lens for understanding existing evaluation frameworks. They also highlight that no single method suffices: robust AI safety assessment requires combining multiple approaches across risk domains, modalities, and adversarial settings.

4 BENCHMARKS AND FRAMEWORKS

4.1 CAPABILITY BENCHMARKS

Capability evaluations remain essential for measuring general reasoning, comprehension, and problem solving skills. BIG Bench (Srivastava et al., 2022) and its successors (BBH, BBEH) test models on diverse, challenging tasks spanning reasoning, language understanding, and world knowledge. HELM (Liang et al., 2022) integrates capability tasks alongside safety metrics, enabling comparable, multi metric analysis across models. HumanEval (Chen et al., 2021) and APPS (Hendrycks et al., 2021) specifically assess code generation, which has direct implications for software reliability and security. These benchmarks form a foundation for evaluating LLM competence, but they do not directly capture safety or alignment.

4.2 SAFETY AND HARM BENCHMARKS

Specialized datasets probe harmful behaviors. RealToxicityPrompts (Gehman et al., 2020) evaluates toxicity generation. CivilComments supports toxicity detection, while BBQ (Parrish et al., 2022) exposes social biases in question answering. TruthfulQA (Lin et al., 2021) evaluates truthfulness by testing models against adversarially constructed questions prone to get imitative falsehoods. Together, these benchmarks quantify specific risks, though they often represent narrow slices of safety concerns.

4.3 RED TEAMING AND ADVERSARIAL PROBING

Red teaming plays a critical role in uncovering failure modes not captured by static benchmarks. Anthropic pioneered large scale manual red teaming of LLMs, identifying categories of harmful outputs and developing automated adversarial prompting strategies. OpenAI Evals provides an robust framework for red teaming, enabling users to script adversarial tests and evaluate models against safety criteria. Automated red teaming approaches, including the use of LLMs to generate adversarial prompts, show promise for scalable, dynamic evaluation.

4.4 PROCESS COMPLIANCE

Models are increasingly assessed for adherence to behavioral norms, such as refusing harmful requests, avoiding jailbreaks, and maintaining honesty. Reinforcement Learning from Human Feedback (RLHF) systems operationalize principles like "helpful, honest, harmless." Evaluations in this category often measure refusal rates, jailbreak success rates, or behavioral compliance under adversarial conditions. OpenAI's safety testing pipelines and Anthropic's "Constitutional AI" highlights frameworks designed to align behavior with acceptable constraints.

4.5 RELIABILITY AND ROBUSTNESS EVALUATION

Robustness evaluations measure consistency under perturbations. Calibration studies (e.g., Hendrycks & Gimpel, 2017) assess probability accuracy, while adversarial robustness tests measure resilience to input variations like typos, dialect changes, or adversarial paraphrases. HELM includes robustness metrics across language shifts and adversarial scenarios. These evaluations highlight vulnerabilities in models and reliability.

4.6 SYSTEM LEVEL AND AGENTIC EVALUATION

As LLMs evolve into autonomous agents capable of tool use and multistep planning, system level evaluations gain importance. Safety Gymnasium and related environments test SafeRL in robotics and multiagent domains under constrained action spaces. HELM also incorporates reasoning focused evaluations such as code execution and mathematical problem solving. Such system level tests are critical for capturing emergent risks not visible in isolated text completions.

4.7 HUMAN IN THE LOOP AUDITS

Human judgment remains indispensable. TruthfulQA relies on human annotation for correctness, while Anthropic's red teaming emphasizes structured human probing. Governance linked efforts include proposals for community based red teaming norms, oversight bodies, and shared safety standards. The combination of automated and human evaluation remains central to building trust in AI deployment.

5 METRICS AND MEASUREMENT PARADIGMS

Evaluation metrics differ widely across benchmarks:

1. **Task accuracy metrics**: accuracy, F1, exact match, pass@k (for coding).
2. Safety specific metrics: toxicity scores, bias differentials, refusal rates, jailbreak success rates.

3. **Calibration metrics**: Expected Calibration Error (ECE), Brier score.
4. Human judgment metrics: Likert ratings, pairwise preference comparisons, expert annotations.
5. **System level metrics**: cumulative reward under constraints (SafeRL), efficiency measures (runtime, compute cost).
6. **Metrics remain a point of contention**: many benchmarks rely on proxy measures (e.g., lexical overlap for truthfulness), which may fail to capture nuanced safety concerns. Multi metric evaluation, as exemplified by HELM, provides a more balanced perspective.

6 COMPARATIVE ANALYSIS

Existing frameworks can be compared along three dimensions:

- **Coverage**: HELM provides coverage across multiple risk domains, while TruthfulQA, BBQ, and RealToxicityPrompts are narrower but more targeted.
- **Depth**: Red teaming efforts (Anthropic, OpenAI Evals) probe deeper into adversarial vulnerabilities compared to static benchmarks.
- **Practicality and openness**: OpenAI Evals and HELM emphasize reproducibility and openness, while many proprietary evaluations remain opaque, limiting independent verification.
- **Fragmentation**: no single framework provides comprehensive coverage of all risks. Instead, researchers must assemble diverse evaluations, trading off between depth and breadth. Moreover, evaluations often lag behind evolving model capabilities, underscoring the need for adaptive and forward looking approaches.

7 FUTURE DIRECTION

Several open problems remain. First, evaluations must adapt to frontier risks, including autonomy, situational awareness, and persuasion. Second, the shift toward multimodal models demands new safety benchmarks that integrate vision, audio, and embodied interaction. Third, scalable oversight mechanisms, such as AI assisted red teaming and human AI hybrid audits, are needed to keep pace with rapidly scaling models. Finally, stronger community norms and international coordination are necessary to align evaluation efforts with governance frameworks and societal values.

8 CONCLUSION

The evaluation of AI safety and alignment has advanced significantly, with benchmarks like BIG Bench, HELM, TruthfulQA, and RealToxicityPrompts, alongside adversarial frameworks such as Anthropic's red teaming and OpenAI Evals. Yet, no single framework offers complete coverage. Robust safety

assessment requires a portfolio approach, combining static benchmarks, adversarial testing and human oversight. Future efforts must prioritize integration, adaptability, and transparency to ensure that evaluation keeps pace with the accelerating capabilities of large language models.

9 TIMELINE

Phase 1: Topic Refinement & Baseline Selection (Week 5–6)

- Select a baseline safety evaluation suite - TruthfulQA and SafetyGym.

Deliverable: Baseline model/suite + justification for selection.

Phase 2: Gap Analysis & Problem Definition (Week 8–9

- Identify limitations of the baseline (e.g., coverage of harmful behaviors, robustness to adversarial prompts, or lack of domain-specific safety tests).
- Define incremental enhancement scope, (e.g. Adding adversarial prompt benchmarks)

Phase 3: Methodology Development (Week 8–9)

- Data Processing Improvements.
- Loss/Metric Enhancements: Add or refine safety-specific metrics (toxicity scores, calibration metrics, robustness under adversarial attack).
- Training Strategy Enhancements

Deliverable: Experimental design + implementation plan.

Phase 4: Implementation & Validation (Week 9–12)

- Implement proposed improvements in code.
- Run evaluations on specified datasets.
- Compare results baseline vs improved.
- Paper completion.

Deliverable: Experimental results (tables, graphs, statistical tests).

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
2. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
3. Christiano, P., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*.

4. Ganguli, D., Askell, A., Bai, Y., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Jones, A., Kadavath, S., Kundu, S., et al. (2022). Red teaming language models to reduce harms. *arXiv preprint arXiv:2209.07858*.
5. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.
6. Ji, J., Yang, K., Xu, Y., Wang, Z., Lu, Y., Zhang, H., Chen, X. (2023). Safety Gymnasium: Safe reinforcement learning benchmarks. *arXiv preprint arXiv:2306.12626*.
7. Liang, P., Bommasani, R., Zong, A., Yu, A. W., Zhang, T., Narayanan, D., et al. (2022). Holistic evaluation of language models (HELM). *arXiv preprint arXiv:2211.09110*.
8. Lin, S., Hilton, J., Evans, O. (2021). TruthfulQA: Measuring how models imitate human falsehoods. *arXiv preprint arXiv:2109.07958*.
9. OpenAI. (2023). OpenAI Evals. https://github.com/openai/evals.
10. Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Bowman, S. (2022). BBQ: A hand built bias benchmark for question answering. In *Proceedings of the Association for Computational Linguistics (ACL)*.
11. Perez, E., Kundu, S., Chen, A., Baidoo, N., Ganguli, D., Askell, A., Goldie, A., Mirhoseini, A., McKinnon, C., Bai, Y., et al. (2022). Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.
12. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A., Abid, A., Fisch, A., Karia, A., et al. (2022). Beyond the Imitation Game Benchmark (BIG Bench). *arXiv preprint arXiv:2206.04615*.