Evaluating Large Language Models : A Survey

The survey aims to categorize and analyze the evaluation of LLMs across three primary dimensions: knowledge and capability evaluation, alignment evaluation, and safety evaluation, while also exploring their performance in specialized domains and the organization of comprehensive evaluation platforms. The paper emphasizes the importance of rigorous evaluation to harness LLMs' potential responsibility, mitigate risks such as bias and misinformation, and guide their development toward societal benefit.

Introduction

The survey begins contextualizing LLM evaluation within the broader history of intelligence assessment, drawing parallels with human intelligence benchmarks like IQ tests. It traces the evolution of NLP evaluation from simple benchmarks e.g. MUC in the 1990s to sophisticated frameworks like GLUE and SuperGLUE, reflecting the growing complexity of language models. The advent of LLMs, exemplified by models like BERT and ChatGPT, has shifted evaluation from task-specific metrics to capability-centered assessments, especially under zero- and few-shot settings. The rapid public adoption of LLMs, such as ChatGPT's 100 million users in two months, underscores their transformative potential and the urgent need to address risks like bias, misinformation, and privacy breaches.

Taxonomy and Roadmap

The authors propose a taxonomy dividing LLM evaluation into five key areas: knowledge and capability, alignment, safety, specialized domains, and evaluation organization. This frameworks aims to answer fundamental questions about LLMs' capabilities, deployment considerations, applicable domains, and performance metrics, providing a structured approach to the survey's analysis

Knowledge and Capability Evaluation

This section assesses LLMs' foundational abilities across four subcategories:

- **Question Answering**: Evaluates practical knowledge application using datasets like SQuAD and HotpotQA, noting that general, broad-source datasets are ideal for pure A assessment.
- **Knowledge Completion**: Tests knowledge retention using benchmarks like LAMA and KoLA, which leverage knowledge bases like Wikidata to probe factual and commonsense understanding.
- **Reasoning**: Divided into common sense e.g. CommonsenseQA, logical e.g. SNLI, multi-hop e.g. HotpotQA, and mathematical reasoning e.g. GSM8K. Studies show LLMs like ChatGPT excel in deductive reasoning but struggle with inductive and complex multi-hop tasks.
- **Tool Learning**: Explores LLMs' ability to manipulate e.g. WebGPT for search engines and create tools, with benchmarks assessing execution success and output quality in domains like robotics and code generation.

Alignment and Evaluation

Alignment evaluation ensure LLMs' outputs align with human values, covering:

- **Ethics and Morality**: Examines ethical decision-making
- **Bias**: Assesses societal biases in downstream tasks (e.g. hiring) and inherent model biases using datasets like StereoSet.
- **Toxicity**: Identifies harmful content via datasets like RealToxicityPrompts, with metrics for classification and generation quality.
- **Truthfulness**: Takes hallucinations using datasets like TruthfulQA, emphasizing the need for factual accuracy.

Safety Evaluation

Safety evaluation focuses on robustness and risk:

- **Robustness**: Tests resilience to prompts and tasks e.g. adversarial attacks, using metrics like accuracy under perturbation.
- **Risk**: Evaluates catastrophic potential (e.g., power-seeking behaviors) as LLMs approach AGI, with emerging benchmarks like ARC Evals assessing autonomous replication

Specialized LLMs Evaluation

This section reviews LLMs' performance in specific fields:

- **Biology and Medicine**: Benchmarks like USMLE and PubMedQA test medical knowledge and application scenarios, with human evaluation highlighting gaps and clinicians.
- **Education**: Assesses teaching (e.g. pedagogical competence) and learning support (e.g., math hints), with LLMs lagging behind human experts.
- **Legislation**: Evaluates legal exams e.g. UBE and reasoning e.g., COLIEE, showing promise but limitations in factual consistency.
- **Computer Science**: Focuses on code generation e.g. HUMANEVAL+ and programming assistance, with LLMs improving functional correctness.
- **Finance**: Tests financial literacy e.g. BloombergGPT and advisory roles, revealing coherent reasoning at scale but ethical concerns

Evaluation Organization

This section outlines benchmarks for comprehensive LLM assessment:

- **NLU and NLG**: GLUE, SuperGLUE, and LongBench evaluate understanding and generation, adapting to long-context challenges.
- **Knowledge and Reasoning**: MMLU, C-Eval, and AGIEval test subject-specific proficiency, revealing uneven performance across domains.
- **Holistic Evaluation**: Frameworks like HELM and OpenCompass integrate multiple dimensions, with leaderboards (e.g., Chatbot Arena) facilitating model comparison.

Future Directions

The survey proposes advancements in:

- **Risk Evaluation**: Beyond QA to situational risk analysis.
- **Agent Evaluation**: Diverse environments to assess autonomy.

- **Dynamic Evaluation**: Adaptive benchmarks to prevent data leakage and test evolving knowledge.
- **Enhancement-Oriented Evaluation**: Focus on actionable insights for model improvement.

Conclusion

The authors conclude that while LLMs exhibit remarkable progress, comprehensive evaluation is essential to define their limits, ensure safety, and maximize benefits. The survey aims to guide future research and development toward responsible LLM advancement.

Insights Gained

- **Holistic Evaluation is Critical**: The survey underscores the need for multi-dimensional evaluation frameworks that go beyond accuracy to include alignment, safety, and domain-specific performance. This holistic approach is vital as LLMs integrate into real-world applications.
- **Capability Gaps Persist**: Despite advancements, LLMs struggle with complex reasoning (e.g., multi-hop, mathematical), computational proficiency, and consistent truthfulness. These gaps suggest a divergence from human cognition, where foundational knowledge drives complex problem-solving.
- **Alignment and Safety Challenges**: Issues like bias, toxicity, and potential catastrophic risks (e.g., power-seeking) highlight the dual-edged nature of LLMs. Current evaluation methods are insufficient for fully mitigating these risks, necessitating innovative approaches like dynamic and agent-based assessments.
- **Domain-Specific Potential and Limits**: LLMs show promise in specialized fields but often fall short of human expertise (e.g., medicine, education). Domain-specific fine-tuning and human evaluation are key to bridging these gaps.
- **Evaluation Evolution**: Static benchmarks risk obsolescence due to data leakage and LLMs' rapid progress. Dynamic, enhancement-oriented

evaluations that adapt to model capabilities and provide improvement insights are the future.

- **Model Size vs. Tuning**: Larger models generally perform better, but fine-tuning with high-quality data can yield competitive results in smaller models, emphasizing the role of training strategies over sheer scale.
- **Multilingual Disparities**: LLMs excel in high-resource languages like English but falter in non-Latin or low-resource languages, indicating a need for diverse training data and language-specific evaluations.