A Survey on Evaluation of Large Language Models

Patrycja Wysocka, Mieszko Mirgos, Łukasz Jaremek, Tomasz Krupiński

https://arxiv.org/pdf/2307.03109

Agenda

1. Introduction,
2. What to evaluate,
   a. Ethics, bias, trustworthiness,
   b. Applications,
   c. Tasks
3. Where to evaluate,
4. How to evaluate,
5. Success and failure cases,
6. Grand challenges.

Introduction

Large Language Models (LLMs): Background

Difference Between LMs and LLMs

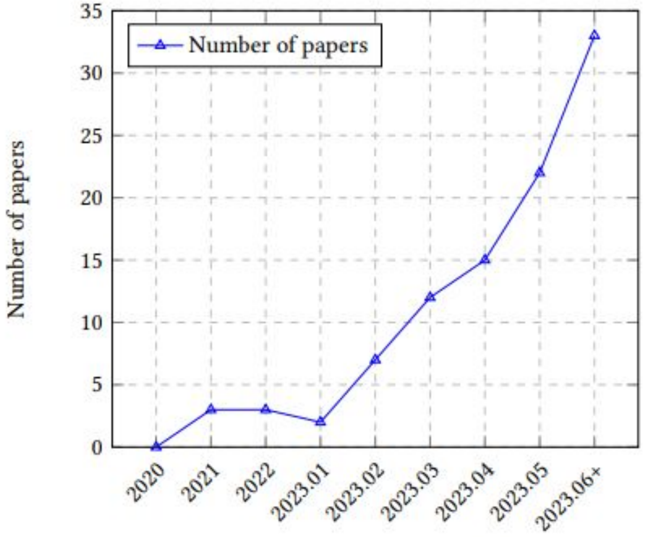
- **LMs**: Predict text sequences; task-specific; smaller datasets.
- **LLMs**: Handle diverse tasks; trained on vast datasets; examples include GPT-3 and GPT-4.

Core Technology of LLMs

- **Transformers**: Efficiently process sequential data; capture long-range dependencies.
- **In-Context Learning**: Adapt to tasks using prompts without retraining.
- **Reinforcement Learning from Human Feedback (RLHF)**: Fine-tune models with human feedback to improve responses.

Interaction with LLMs

- **Prompts**: Guide responses by providing specific instructions.
- **Q&A**: Enable natural, conversational interactions.



Why LLMs?

- ★ **Revolutionized AI with capabilities beyond task-specific models** - Traditional AI models were typically designed to perform well on narrowly defined tasks, while (LLMs) like GPT-4 have shifted this paradigm by excelling across multiple domains without needing task-specific fine-tuning.
- ★ **Address diverse needs like:**
  - ○ education
  - ○ healthcare
  - ○ customer service automation
  - ○ documents translation
- ★ Spark debates about their potential as **Artificial General Intelligence (AGI).**

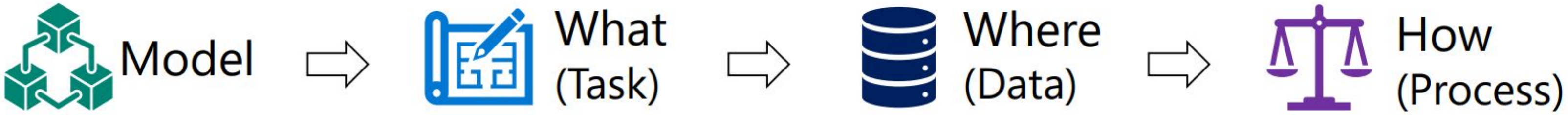
Evaluation: Background

Standard Evaluation Methods:

- **k-Fold Cross-Validation**: Splits data into *k* parts; reduces training data loss.
- **Holdout Validation**: Divides data into training and testing sets; simpler but prone to bias.
- **Leave One Out Cross-Validation (LOOCV)**: Uses one data point as a test set.
- **Reduced Set**: Trains on one dataset and tests on another; limited applicability.

Challenges for LLM Evaluation

- LLMs' growing popularity and poor interpretability challenge current protocols.
- Existing methods may not fully capture LLMs' capabilities and potential risks.



Model → What (Task) → Where (Data) → How (Process)

Importance of evaluation

Key Significance of Evaluation:

★ Identifies **strengths and weaknesses** of LLMs (e.g., sensitivity to adversarial prompts, requiring robust prompt engineering).

★ **Guides better human-LLM** interaction design and implementation.

★ Ensures **safety and reliability** in critical sectors like healthcare and finance.

★ Adapts to emerging abilities of larger models with new evaluation protocols.

★ Ethical issues can hinder societal trust and deployment of LLMs.

Current Gaps in Evaluation:

★ Existing research lacks a comprehensive framework.

★ Evolving capabilities of LLMs challenge current protocols, emphasizing the need for multifaceted evaluation techniques.

Structured of evaluation framework

Three core dimensions:

1. **What to Evaluate**: Tasks and capabilities assessed in various domains.
   a. Tasks (NLU, Reasoning, Text generation)
   b. Robustness, Ethics, Bias, and Trustworthiness
   c. Applications
2. **Where to Evaluate**: Datasets and benchmarks used for evaluation.
3. **How to Evaluate**: Methods and processes to assess LLMs' performance (human, automatic).

What to evaluate

Ethics, bias, trustworthiness

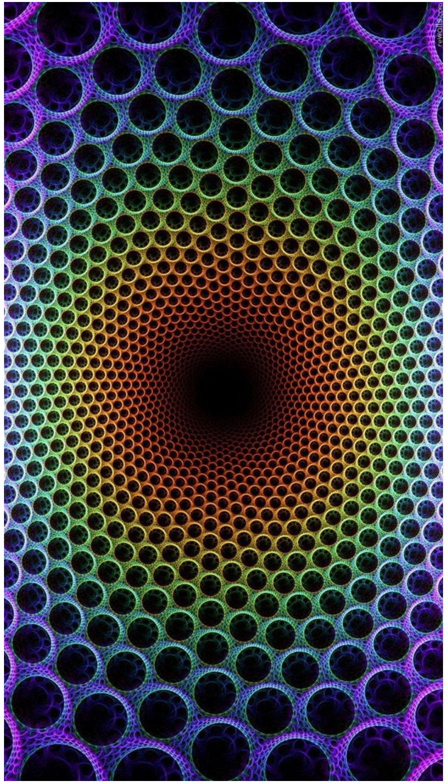
Robustness and Ethics in LLM Evaluation

	Robustness	Ethics and Bias
Key issues	**Out-of-Distribution (OOD)**: Studies how well models perform on unseen data. **Adversarial Robustness**: Evaluates resistance to malicious inputs.	**Toxicity**: Offensive or harmful content (e.g., hate speech, insults). **Social Biases**: Stereotypes in gender, race, religion, etc.
Tools	PromptBench for adversarial text attack evaluation. **PromptBench**	CHBias dataset for Chinese conversational bias evaluation.
Findings	Contemporary LLMs are vulnerable to adversarial inputs. Visual input manipulation reveals risks in vision-language models.	Role-playing amplifies toxicity in responses. Biases extend to political tendencies and cultural values. GPT-4 alignment studies reveal systematic biases.

Hallucinations

Hallucinations: Factual inaccuracies or ungrounded statements in generated content.

- **Key Insights**:
  - Models like GPT-4 generate seemingly factual but sometimes incorrect outputs.
  - Illusions in visual language models are significant.
- **Mitigation Tools**:
  - Datasets: LRV-Instruction for robust evaluation.
  - Methods: POPE (Polling-based query method) for assessing visual instructions.
- **Objective**:
  - Improve training techniques to reduce hallucinations and enhance trustworthiness.



Tasks

Natural Language Understanding

1. **Sentiment Analysis: Strong in English, weak in low-resource languages**
2. **Text Classification: High accuracy (~85%) for general tasks**
3. **Natural Language Inference: Good factual comprehension but struggles with representing human disagreement**
4. **Semantic Understanding: Can grasp individual events but limited in semantic connections**

Reasoning

1. **Strong: Arithmetic reasoning, causal reasoning, temporal reasoning**
2. **Weak: Abstract reasoning, symbolic reasoning, multi-hop reasoning**
3. **ChatGPT vs GPT-3.5: Better at arithmetic, worse at symbolic tasks**
4. **Complex tasks remain challenging for all LLMs**

Natural Language Generation

1. **Summarization: Comparable to traditional models**
2. **Dialogue: Latest models (Claude/ChatGPT) outperform earlier versions**
3. **Translation: Superior to commercial MT systems but struggles with non-English**
4. **QA: Strong performance with 2%+ improvement over previous models**

Multilingual Tasks

1. **Major challenge: Poor performance on non-Latin scripts**
2. **Limited effectiveness in low-resource languages**

Factuality

Best models achieve 80%+ accuracy on factual questions

Novel measures proposed:

- Information theory metrics
- Atomic fact breakdown
- Natural language inference & question generation

Applications

Applications

- Social Science,
- Natural Science,
- Engineering,
- Medical.

Social Science Applications

- **Key Contributions**:

  LLMs support text analysis, policy-making, and psychology by:

  - Scaling text-based research in political ideologies.
  - Understanding complex social phenomena like hate and empathy, despite accuracy limitations.
  - Assisting legal experts in summarizing case law, though precision remains a challenge.
  - Offering insights into human cognition through alternative evaluation methods.
- **Limitations**:

  LLMs cannot fully replace domain experts due to accuracy and contextual understanding challenges.

Natural Science Applications

- **Mathematics**:
  - GPT-4 excels in basic tasks (addition, subtraction) but struggles with complex problems like division and trigonometry.
  - Performance decreases with problem complexity, indicating a need for better reasoning capabilities.

- **General Science**:
  - Limited accuracy in chemistry and physics evaluations (25–100% in chemistry tasks).
  - Further improvements are essential for broader scientific applications.

Engineering Applications

- **Code Generation**:
  - GPT-4 shows advanced skills in dynamic programming and code understanding.
  - ChatGPT struggles with advanced concepts in data structures and graph theory.
- **Software Engineering**:
  - Performs well in general tasks but fails in complex issues like vulnerability detection.
  - Limited capability in commonsense and advanced planning tasks.

By 2025, Meta will be able to develop and deploy AI systems that program at the level of an intermediate engineer.

~Mark Zuckerberg during a podcast interview with Joe Rogan, January 2025

Medical Applications

- **Medical Queries**:

  LLMs provide accurate responses in genetics, oncology, and biomedicine but often fail to cite reliable sources.

- **Medical Examinations**:

  Performance in tests like USMLE remains below the passing threshold.

- **Medical Assistance**:

  Promising in diagnostics and therapy but requires significant optimization for clinical reliability.

Where to evaluate

Benchmarks

- A lot of different benchmarks, authors listed out 46 of them.
- Different types of focus, for example: chat-bots, software tools, dynamic QA
- Three types of domain: **General Language Task, Specific Downstream Task, Multi-modal Task**
- Different types of evaluation criteria, for example: social language understanding, winrate judged by GPT-4, Accuracy

Benchmarks For General Language Tasks

General task benchmarks assess LLMs' performance across a broad spectrum of tasks, including language understanding, reasoning, and generation.

They are designed to evaluate models' general capabilities and adaptability in diverse scenarios.

These benchmarks help identify both the strengths and limitations of models in handling varied linguistic and contextual challenges.

Benchmarks For General Language Tasks

- **Chatbot Arena** and **MT-Bench**: Focus on chatbot interaction, multi-turn dialogues, and user engagement.
- **HELM**: Assessment across tasks like language understanding and domain-specific reasoning.
- **PromptBench**: Assesses robustness to adversarial prompts.

Benchmarks For Specific Downstream Tasks

Benchmarks for specific downstream tasks are tailored to evaluate LLMs' performance on particular applications and domains.

These benchmarks focus on specialized tasks such as reasoning, ethical considerations, or domain-specific knowledge, providing detailed insights into the models' capabilities in those areas.

Benchmarks For Specific Downstream Tasks

- **MultiMedQA**: Medical benchmark that focuses on medical examinations, medical research, and consumer healthcare questions.
- **FRESHQA**: Assessment of the ability of LLMs in dynamic QA about current world knowledge.
- **GAOKAO-Bench**: Testing the proficiency of LLMs in intricate and context-specific tasks, utilizing questions sourced from the Chinese Gaokao examination.

Benchmarks For Multi-modal Tasks

Benchmarks for multi-modal tasks evaluate LLMs' ability to process and integrate information from multiple data modalities, such as text, images, and videos.

These benchmarks assess both the perceptual and cognitive capabilities of models, focusing on tasks that require understanding and generating across different formats.

Benchmarks For Multi-modal Tasks

- **SEED-Bench**: 19,000 multiple-choice questions. Covers 12 different aspects, including the models' proficiency in understanding patterns within images and videos
- **MME**: Specially crafted instructions with answers, that guarantee equitable evaluation conditions.

How to evaluate

Automatic Evaluation

- **LLM-EVAL**: Automatic evaluation method for open-domain conversations with LLMs.
- **HELM, Chatbot Arena and others** : Some previously mentioned benchmarks, also possess the ability for automatic Evaluation.

Automatic Evaluation

- **Accuracy:** is a measure of how correct a model is on a given task. The concept of accuracy may vary in different scenarios. (Exact Match, F1 score, and ROUGE score.)
- **Calibrations:** pertains to the degree of agreement between the confidence level of the model output and the actual prediction accuracy. (Expected Calibration Error (ECE), Area Under the Curve (AUC))
- **Fairness:** whether the model treats different groups consistently. (Demographic Parity Difference (DPD), Equalized Odds Difference (EOD))
- **Robustness:** evaluates the performance of a model in the face of various challenging inputs, including adversarial attacks, changes in data distribution, noise, etc.

Human Evaluation

- When automatic means not enough.
- Experts, researchers or users evaluate answers.
- It is important that assessors are methodologically trained and familiar with the task.

Human Evaluation

- **Accuracy:** assesses the precision and correctness of the generated text.
- **Relevance:** tests how well the text addresses the given context or query.
- **Fluency:** assesses the language model's ability to produce content maintaining a consistent tone and style
- **Transparency:** assessing how well the model communicates its thought processes.
- **Safety:** examines the language model's ability to avoid producing content that may be inappropriate, offensive, or harmful.
- **Human alignment:** assesses the degree to which the language model's output aligns with human values, preferences, and expectations.

Success And Failure Cases

Success Cases

- LLMs are skilled at generating fluent text.
- LLMs excel in language understanding tasks.
- LLMs perform well in arithmetic and logical reasoning.
- LLMs handle context to generate coherent responses.
- LLMs achieve strong results in various NLP tasks.

Failure Cases

- Very complex reasoning often leads to mistakes.
- LLMs' Performance is weak in non-latin languages.
- LLMs can produce biased and harmful content.
- LLMs may generate fake or incorrect information.
- LLMs have limitations in incorporating real-time information.
- LLMs are vulnerable to carefully crafted prompts.

Grand challenges

Grand Challenges

1. **Designing true AGI benchmarks - beyond current human-centric testing**
2. **Developing complete behavioral evaluation in open environments**
3. **Ensuring robustness against diverse real-world inputs**
4. **Creating dynamic evaluation systems that evolve with LLM capabilities**

A Survey on Evaluation of Large Language Models

https://arxiv.org/pdf/2307.03109