Measuring AI/LLM Capabilities and Progress

Your Name

June 2025

Why Measure AI/LLM Capabilities?

- **Rapid progress** in Large Language Models (LLMs) and AI systems.
- **Need for robust, transparent benchmarks** to track abilities and risks.
- **Inform policy, safety, and deployment decisions.**
- **Ensure accountability and trust** in AI systems used by the public and organizations.
- **Identify and mitigate biases and harmful behaviors** that may emerge in AI outputs.
- **Guide investment and research priorities** by highlighting strengths and weaknesses.
- **Support ethical AI development** through measurable standards and continuous improvement.
- **Facilitate fair comparisons** between different models and approaches.

Frameworks for Measuring Progress

- **Task Completion Benchmarks:** Assess ability to perform complex, multi-step tasks; widely used for model comparison and validation.
- **Leaderboards:** Public, standardized comparisons across models; encourage transparency and drive innovation.
- **Prediction Markets:** Aggregate expert and crowd forecasts on AI milestones; leverage collective intelligence for progress tracking.
- **International Reports:** Comprehensive, cross-institutional tracking of AI progress and risks; provide global perspective and policy guidance.
- **Standardized Evaluation Datasets:** Curated datasets for reproducible testing; ensure fair and consistent model assessment.
- **User Feedback and Real-World Deployment:** Collect empirical data from real applications; highlight practical performance and user satisfaction.
- **Longitudinal Studies:** Track AI system improvements over time; identify trends and inflection points in capability growth.
- **Expert Panels and Peer Reviews:** Structured assessments by domain specialists; offer nuanced insights beyond automated metrics.

Task Completion Benchmarks

- **Long-Horizon Task Evaluation:** METR's research measures LLMs' ability to complete extended, multi-step tasks, highlighting current limitations in reliability and autonomy (METR 2025). Recent results show that frontier models can autonomously complete tasks with a time horizon of about 40 minutes, but not yet work requiring days or weeks (**nature2025**; METR 2025).
- **PowerPoint Task Completion (PPTC):** Evaluates LLMs on multi-turn, multi-modal instructions within PowerPoint, revealing challenges in tool use, error accumulation, and multi-modality (ACL Paper 2024). GPT-4 leads in performance, but all models struggle with non-text operations and long sessions (ACL Paper 2024).
- **Long-Horizon Vision-Language Navigation (LHPR-VLN):** Benchmarks LLMs and specialized agents on multi-stage, complex navigation tasks. Most models fail on short subtasks; only fine-tuned or specialized agents (e.g., MGDM) show partial success, emphasizing the importance of memory and holistic understanding in long tasks (**arxiv-lhpr-vln**).
- **Other Environments:**
  - Robotics: SayCan uses LLMs to generate action sequences for robots.
  - Web Navigation: WebShop assesses LLMs in e-commerce scenarios.
  - Agent-Based Tasks: AgentBench evaluates LLMs as autonomous agents across 8 diverse environments.
- **Key Insights:**
  - Current LLMs excel at short, well-defined tasks but face reliability and error accumulation in long or complex workflows.

Leaderboards

- **Purpose:**
  - Track and compare LLM performance across benchmarks.
  - Provide a standardized way to evaluate model capabilities.
- **Examples:**
  - Gorilla APIBench Leaderboard (G. Team 2025)
  - Aider.chat Leaderboards (A. Team 2025)
  - Hugging Face Open LLM Leaderboard
  - LMSYS Chatbot Arena
- **Benefits:**
  - Promote transparency and reproducibility in AI evaluation.
  - Encourage healthy competition and rapid innovation.
  - Help researchers and practitioners identify state-of-the-art models.
- **Considerations:**
  - Leaderboards may not capture all real-world use cases.
  - Need to ensure benchmarks are robust, diverse, and up-to-date.
  - Risk of overfitting to specific leaderboard metrics.

Prediction Markets for AI Progress

- Platforms like Metaculus and Polymarket aggregate forecasts on AI milestones (Metaculus 2025; Polymarket 2025).
- Useful for synthesizing expert and crowd expectations about future capabilities.
- Complement empirical benchmarks with probabilistic insights.
- Recent surge in activity: Major platforms have seen notable increases in trading volume and engagement, especially around high-profile events and technological milestones (Polymarket 2025; Metaculus 2025).
- Metaculus specializes in long-term, technology-focused questions, enabling nuanced tracking of progress in areas like quantum computing and advanced AI systems (Metaculus 2025).
- Polymarket and PredictIt demonstrate how prediction markets can reflect real-time shifts in collective expectations, sometimes diverging from traditional expert consensus (Polymarket 2025; PredictIt 2025).
- AI-powered information aggregation is enhancing prediction markets, allowing for finer-grained, real-time analysis and more targeted event creation (**hackernoon**).
- Prediction markets can help identify emerging trends, inform policy, and guide strategic investments in AI by revealing where consensus and uncertainty lie.

International Tracking and Reports

- **International AI Safety Report:**
  - Annual, multi-stakeholder assessment of AI progress, risks, and governance (I. A. S. R. Team 2025).
  - Includes expert insights from academia, industry, and civil society.
- **TrackingAI.org:**
  - Centralized resource for tracking AI system capabilities and benchmarks (T. Team 2025).
  - Features interactive dashboards and regular updates.
- **Emerging Initiatives:**
  - Regional and international AI observatories (e.g., EU AI Observatory).
  - Collaborative databases for sharing best practices and incident reports.
- **Key Benefits:**
  - Facilitate global coordination and evidence-based policy.
  - Increase transparency and accountability in AI development.
  - Support proactive risk management and regulatory adaptation.

Challenges in Measuring AI Capabilities

- **Multi-modality and real-world complexity remain difficult to benchmark** (ACL Paper 2024).
  - Integrating text, images, audio, and video introduces interdependencies that are hard to isolate and measure.
  - Real-world scenarios often involve ambiguous or incomplete information, making standardized evaluation challenging.
- **Error accumulation in long-horizon tasks.**
  - As AI systems perform longer sequences of actions or reasoning steps, small errors can compound, leading to significant inaccuracies.
  - This makes it difficult to assess reliability over extended interactions or complex workflows.
- **Subjective tasks (e.g., aesthetics) are hard to evaluate automatically.**
  - Human judgment is often required for tasks involving creativity, style, or subjective quality.
  - Automated metrics may fail to capture nuances that are obvious to humans.
- **Need for continual updates as models and tasks evolve.**
  - Benchmarks quickly become outdated as new models and capabilities emerge.
  - Continuous adaptation of evaluation frameworks is necessary to keep pace with technological progress.
- **Generalization across domains remains a key challenge.**
  - Models often perform well on specific benchmarks but struggle to generalize to unseen or novel situations.
  - Ensuring robustness and adaptability in diverse environments is an ongoing research problem.

Conclusion

- Measuring AI/LLM capabilities is essential for safe and effective deployment.
- Combination of benchmarks, leaderboards, prediction markets, and international reports provides a holistic view.
- Ongoing research is needed to address emerging challenges and ensure robust evaluation.
- Collaboration among academia, industry, and policymakers is crucial for advancing evaluation methods.
- Transparency in AI assessment processes builds public trust and supports informed decision-making.
- Future directions should consider ethical implications and societal impact alongside technical performance.

Thank You

Thank you for your attention! Questions?

References I

ACL Paper, Author(s) of (2024). "PowerPoint Task Completion: Evaluating LLMs on Multi-Turn, Multi-Modal Instructions". In: *Findings of the Association for Computational Linguistics*. ACL 2024. URL: https://aclanthology.org/2024.findings-acl.514.pdf.

Metaculus (2025). *Metaculus: AI Prediction Markets*. URL: https://www.metaculus.com/questions/?topic=ai.

METR (2025). *Measuring AI Ability to Complete Long Tasks*. URL: https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.

Polymarket (2025). *Polymarket: AI Markets*. URL: https://polymarket.com/markets/all/ai.

PredictIt (2025). *PredictIt Markets*. URL: https://www.predictit.org/.

Team, Aider (2025). *Aider.chat Leaderboards*. URL: https://aider.chat/docs/leaderboards/.

Team, Gorilla (2025). *Gorilla APIBench Leaderboard*. URL: https://gorilla.cs.berkeley.edu/leaderboard.html.

References II

 Team, International AI Safety Report (2025). *International AI Safety Report*. Tech. rep. International AI Safety Report. URL: https://arxiv.org/abs/2501.17805.

 Team, TrackingAI (2025). *TrackingAI.org*. URL: https://trackingai.org/home.