Training Course on Evaluating and Benchmarking LLM Performance

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

Large Language Models (LLMs) are rapidly transforming industries, offering unprecedented capabilities in natural language processing and content generation. However, harnessing their full potential requires a deep understanding of their performance, limitations, and how to effectively evaluate and benchmark them. Training Course on Evaluating & Benchmarking LLM Performance: Metrics and Methodologies for Assessing Generative Models delves into the critical metrics and methodologies essential for rigorously assessing generative models, ensuring optimal deployment and maximizing business value in real-world applications. Participants will gain practical skills to identify key performance indicators, implement robust evaluation frameworks, and confidently select and fine-tune LLMs for diverse use cases.

The ability to accurately evaluate and benchmark LLMs is paramount for data scientists, AI engineers, and product managers navigating the complex landscape of generative AI. This course addresses the critical need for a structured approach to model assessment, moving beyond superficial metrics to encompass nuanced aspects like factual accuracy, bias detection, safety, and human-in-the-loop evaluation. By mastering these techniques, organizations can mitigate risks, enhance model reliability, and unlock the true potential of LLMs to drive innovation and competitive advantage.

Programme Curriculum

Training Course on Evaluating & Benchmarking LLM Performance: Metrics and Methodologies for Assessing Generative Models

Introduction

Large Language Models (LLMs) are rapidly transforming industries, offering unprecedented capabilities in natural language processing and content generation. However, harnessing their full

potential requires a deep understanding of their performance, limitations, and how to effectively evaluate and benchmark them. Training Course on Evaluating & Benchmarking LLM Performance: Metrics and Methodologies for Assessing Generative Models delves into the critical metrics and methodologies essential for rigorously assessing generative models, ensuring optimal deployment and maximizing business value in real-world applications. Participants will gain practical skills to identify key performance indicators, implement robust evaluation frameworks, and confidently select and fine-tune LLMs for diverse use cases.

The ability to accurately evaluate and benchmark LLMs is paramount for data scientists, AI engineers, and product managers navigating the complex landscape of generative AI. This course addresses the critical need for a structured approach to model assessment, moving beyond superficial metrics to encompass nuanced aspects like factual accuracy, bias detection, safety, and human-in-the-loop evaluation. By mastering these techniques, organizations can mitigate risks, enhance model reliability, and unlock the true potential of LLMs to drive innovation and competitive advantage.

Course Duration

10 days

Course Objectives

1. Understand and implement comprehensive frameworks for Large Language Model evaluation and benchmarking generative AI models.
2. Learn to select and apply the most relevant LLM performance metrics, including perplexity, BLEU, ROUGE, F1-score, and semantic similarity.
3. Develop strategies for bias detection and mitigation in LLMs to ensure ethical AI and fairness .
4. Evaluate LLMs for safety, adversarial robustness, and hallucination detection in real-world scenarios.
5. Design and execute human evaluation protocols for subjective and nuanced LLM outputs.
6. Compare and contrast the performance of various LLM architectures (e.g., Transformer, GPT, Llama) across diverse tasks.
7. Tailor LLM evaluation strategies to specific business use cases, from chatbots to content generation.
8. Differentiate and apply intrinsic and extrinsic evaluation methods for a holistic assessment of LLM capabilities.
9. Utilize cutting-edge automated LLM evaluation tools and platforms for efficient assessment.
10. Measure and optimize LLM inference speed and cost efficiency for production deployment.
11. Conduct adversarial testing and red teaming to uncover vulnerabilities and improve LLM resilience.
12. Learn to create and curate domain-specific benchmarking datasets for tailored LLM assessment.
13. Explore the latest LLM evaluation trends, model selection criteria, and Responsible AI practices.

Organizational Benefits

- Accurately assess LLM performance to ensure maximum return on **generative AI** projects.
- Minimize the risks of **biased or unsafe AI outputs** by implementing robust evaluation processes, fostering greater trust in AI systems.

- Make informed decisions on selecting and deploying the **best-performing LLMs** for specific organizational needs.
- Streamline **LLM development and deployment pipelines** through efficient evaluation methodologies.
- Leverage **superior LLM performance** to drive innovation in product development, customer service, and content creation.
- Empower teams with the data and insights needed to make strategic decisions about **AI model governance** and improvement.
- Deploy highly effective and reliable LLMs for applications like chatbots and personalized content, leading to enhanced customer satisfaction.
- Identify and address performance bottlenecks, leading to more efficient resource allocation and **cost reduction** in LLM operations.

Target Audience

1. Data Scientists.
2. AI/ML Engineers.
3. Product Managers (AI/ML Focus)
4. AI Researchers
5. Quality Assurance (QA) Engineers.
6. Solution Architects.
7. Data Ethicists and Governance Specialists
8. Technical Leads and Project Managers

Course Outline

Module 1: Introduction to LLM Evaluation Fundamentals

- Understanding the LLM Landscape: Overview of popular LLMs (GPT, Llama, Gemini, Claude) and their diverse applications.
- Why Evaluate LLMs? The critical need for systematic assessment beyond qualitative observation.
- Challenges in LLM Evaluation: Addressing the non-deterministic nature, open-ended responses, and subjective quality.
- Key Evaluation Dimensions: Introduction to accuracy, fluency, coherence, relevance, safety, and bias.
- **Case Study:** Analyzing early public perception issues with un-evaluated LLM chatbots and the business impact.

Module 2: Intrinsic Metrics for LLM Performance

- Perplexity and Log-Likelihood: Measuring how well a language model predicts a sample of text.
- BLEU Score: Evaluating machine translation and text generation against reference translations.
- ROUGE Score: Assessing summarization quality by comparing generated text to reference summaries.
- F1-Score, Precision, and Recall: Applying classification metrics to structured LLM outputs.
- **Case Study:** Comparing different summarization models using ROUGE scores for news article condensation.

Module 3: Extrinsic Metrics and Task-Specific Evaluation

- Task-Oriented Evaluation: Assessing LLM performance in specific downstream applications (e.g., question answering, sentiment analysis).
- Human Evaluation Protocols: Designing clear guidelines and rubrics for human annotators to score LLM outputs.
- A/B Testing and User Feedback: Incorporating real-world user interactions and feedback for live model performance assessment.
- Goal-Oriented Dialogue Metrics: Evaluating conversational AI for task completion, engagement, and user satisfaction.
- **Case Study:** Evaluating an LLM-powered customer service chatbot based on resolution rates and customer satisfaction scores (CSAT).

Module 4: Benchmarking Methodologies and Datasets

- Standardized LLM Benchmarks: Exploring popular benchmarks like MMLU, HELM, GLUE, and SuperGLUE.
- Creating Custom Benchmarking Datasets: Strategies for developing relevant and representative datasets for specific domains.
- Cross-Model Comparison: Techniques for comparing different LLMs on the same tasks and metrics.
- Reproducibility and Robustness in Benchmarking: Ensuring consistent and reliable evaluation results.
- **Case Study:** Benchmarking open-source vs. proprietary LLMs for a financial text analysis task using a curated dataset.

Module 5: Evaluating Factual Accuracy and Hallucinations

- Defining Hallucinations: Understanding the phenomenon of LLMs generating factually incorrect or nonsensical information.
- Metrics for Factual Consistency: Using techniques like fact-checking APIs and knowledge graph validation.
- Automated Hallucination Detection: Exploring methods like self-consistency checks and contradiction detection.
- Human-in-the-Loop for Factual Review: Incorporating human oversight for high-stakes factual generation.
- **Case Study:** Developing a system to identify and reduce factual inaccuracies in an LLM-generated medical diagnosis report.

Module 6: Bias and Fairness in LLM Evaluation

- Sources of Bias in LLMs: Understanding how training data and model architecture can lead to biased outputs.
- Measuring Bias: Using demographic parity, equalized odds, and other fairness metrics.
- Bias Mitigation Strategies: Techniques like data re-weighting, adversarial debiasing, and in-context learning.
- Ethical Considerations in AI: Discussing the societal impact of biased LLMs and the importance of responsible AI development.
- **Case Study:** Analyzing gender and racial bias in an LLM used for resume screening and implementing debiasing techniques.

Module 7: LLM Safety and Adversarial Robustness

- Defining AI Safety: Identifying harmful, unethical, or dangerous LLM outputs.

- Red Teaming LLMs: Proactive identification of model vulnerabilities and failure modes through adversarial prompting.
- Guardrails and Content Moderation: Implementing mechanisms to prevent the generation of harmful content.
- Robustness to Adversarial Attacks: Testing LLMs against subtle input perturbations designed to elicit undesirable behavior.
- **Case Study:** Simulating prompt injection attacks on a public-facing chatbot and implementing robust defenses.

Module 8: Human-in-the-Loop (HITL) Evaluation

- Role of Human Evaluation: Why human judgment remains crucial for nuanced and subjective LLM outputs.
- Designing Effective Annotation Tasks: Crafting clear instructions, rubrics, and quality control measures for human annotators.
- Crowdsourcing vs. Expert Annotators: Choosing the right human evaluation strategy based on task complexity and budget.
- Inter-Annotator Agreement: Measuring consistency among human evaluators.
- **Case Study:** Conducting a large-scale human evaluation campaign to assess the creativity and coherence of an LLM-generated marketing copy.

Module 9: Automated LLM Evaluation Tools and Platforms

- Overview of Open-Source Tools: Exploring libraries like Hugging Face evaluate, RAGAS, and DeepEval.
- Commercial LLM Evaluation Platforms: Reviewing features of platforms like Galileo, TruLens, and LangFuse.
- Setting up Evaluation Pipelines: Integrating evaluation tools into CI/CD workflows for continuous assessment.
- Customizing Evaluation Metrics: Extending existing tools or developing bespoke metrics for unique use cases.
- **Case Study:** Implementing an automated evaluation pipeline for a Retrieval-Augmented Generation (RAG) system using RAGAS to measure faithfulness and contextual precision.

Module 10: Performance Optimization: Throughput, Latency, and Cost

- Understanding LLM Inference: The computational demands of running LLMs in production.
- Measuring Throughput and Latency: Benchmarking inference speed and efficiency across different hardware and model configurations.
- Cost Analysis of LLM Usage: Calculating the operational costs of large-scale LLM deployment.
- Optimization Techniques: Exploring quantization, pruning, distillation, and efficient serving frameworks (e.g., vLLM).
- **Case Study:** Optimizing an LLM deployment for reduced latency and cost in a real-time recommendation engine.

Module 11: Evaluating Multimodal LLMs

- Introduction to Multimodal LLMs: Understanding models that process and generate across text, image, and audio.
- Metrics for Multimodal Tasks: Extending traditional metrics to evaluate cross-modal understanding and generation.

- Human Evaluation for Multimodal Outputs: Challenges and best practices for assessing multimodal content.
- Benchmarking Multimodal Capabilities: Exploring specialized benchmarks for image captioning, visual question answering, etc.
- **Case Study:** Evaluating a multimodal LLM for its ability to generate descriptive image captions, assessing both textual quality and visual relevance.

Module 12: LLM Evaluation in Production and MLOps

- Continuous Monitoring of LLM Performance: Setting up alerts and dashboards for tracking model drift and performance degradation in live environments.
- A/B Testing in Production: Conducting live experiments to compare different LLM versions or strategies.
- Feedback Loops for Model Improvement: Integrating user feedback and production data back into the development cycle.
- Observability for LLM Applications: Tools and techniques for understanding the internal workings and outputs of deployed LLMs.
- **Case Study:** Implementing an MLOps pipeline for an LLM-powered content generation system, monitoring for content quality and user engagement.

Module 13: Advanced Topics in LLM Evaluation

- Evaluating LLM Agents: Assessing the performance of autonomous AI agents built with LLMs.
- Long-Context Window Evaluation: Challenges and strategies for evaluating LLMs with very long input sequences.
- Robustness to Distribution Shift: Testing LLMs on data that deviates from the training distribution.
- Counterfactual Reasoning Evaluation: Assessing LLM ability to handle hypothetical scenarios.
- **Case Study:** Evaluating an LLM agent designed to automate complex multi-step workflows, focusing on task completion and error handling.

Module 14: Responsible AI and LLM Governance

- Establishing AI Governance Frameworks: Developing policies and procedures for ethical and responsible LLM deployment.
- Regulatory Compliance: Understanding emerging regulations and standards for AI systems.
- Explainable AI (XAI) for LLMs: Techniques for interpreting and understanding LLM decisions and outputs.
- Data Privacy and Security in LLM Development: Best practices for handling sensitive data during training and inference.
- **Case Study:** Developing a governance framework for

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
07 Sep 2026	18 Sep 2026	Physical	$3,000

Start Date	End Date	Location	Price
14 Sep 2026	25 Sep 2026	Physical	$3,000
21 Sep 2026	02 Oct 2026	Physical	$3,000
28 Sep 2026	09 Oct 2026	Physical	$3,000
05 Oct 2026	16 Oct 2026	Physical	$3,000
12 Oct 2026	23 Oct 2026	Physical	$3,000
19 Oct 2026	30 Oct 2026	Physical	$3,000
26 Oct 2026	06 Nov 2026	Physical	$3,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com