TURING

Your Path from LLM Evaluation to AI Acceleration

Whether you're seeking the right model to enhance your business or benchmarking your LLM against competitors, the evaluation process can help. Gather insights that turn into real performance gains.

What is LLM evaluation?

If you are looking to purchase an LLM

Assess a model's performance and effectiveness in fundamental tasks and use a variety of techniques to determine the best model for your needs.

If you are looking to improve your LLM

Evaluate an LLM's current performance against competitors. Identify where the model needs improvement and create a roadmap for training and refinement.

Which LLM evaluation path is right for you?

Purchasing an LLM

Evaluation questions:

What's your goal for the LLM?

What's your current data strategy and technology stack?

To develop insights:

What model is best for your specific use case?

What paths are available to fine-tune that model?

Improving an LLM

Evaluation questions:

How does your model measure key metrics like perplexity, BLEU, ROUGE, and F1 score?

How do you perform on benchmarks like GLUE, DeepEval, and OpenAI Evals?

To holistically improve model:

Based on your LLM's goals, use case, and domain, what specific areas need improvement?

What techniques will most effectively retrain your model, from creating new datasets to reinforcement learning from human feedback?

Assessing LLM performance

Evaluation insights inform a unique roadmap to fine-tune an LLM you're using or optimize an LLM you're building.

Purchasing an LLM

Improving an LLM

Tailored evaluation approach

A healthcare company seeks a domain-specific LLM to speed up FDA compliance tasks and improve productivity. **Routine audits take 3-5 days** on-site to evaluate the model's complex inspection workflows, plus preparation and follow-up activities.

An AI company, focused on benchmarking its model's **reasoning and problem-solving capabilities** against competitors, wants to evaluate its proprietary datasets and improve cognitive performance. Complex audits may take weeks, depending on the findings and corrective actions.

Establishing needs and fact-finding

Evaluation considers domain needs and use cases to find the right model. Evaluators tailor assessment metrics to focus on regulatory compliance, accuracy in medical language, ability to **simplify inspection workflows, and productivity (up to 30%).**

Evaluation reveals reproducible patterns of failure and behavioral generalization weaknesses slowing down the model's reasoning and high-level cognitive capabilities, requiring **new datasets and RHLF to address gaps.**

Addressing identified weaknesses

A fine-tuning roadmap for the application includes **curating domain-specific datasets** to strengthen its understanding of medical terminology and nuance, **driving greater accuracy during FDA audits and supporting complex cognitive tasks.**

Model improvements were supported by creating high-quality proprietary datasets trained by **effective SFT and RLHF techniques to enhance reasoning and problem-solving** capabilities. Model's performance was continuously benchmarked against competitors, ensuring improvements across language generation, summarization, and high-level cognitive tasks.

Results: LLM evaluations maximize ROI value

Implementing an LLM

Improving an LLM

Potential costs of no evaluation

Sunk costs of having to purchase a different model after realizing fine-tuning costs were too high or simply unachievable for the model chosen.

Focusing training throughput on the wrong areas and seeing minimal performance gains.

Results from evaluation

↓22%

Time spent managing FDA compliance reduced 22%.

+18%

Curated datasets boosted model training, improving benchmark score by 18%.

Results you can expect when fine-tuning an LLM:

Developing an effective UI to simplify engagement

Domain knowledge allowed for testing against edge cases

Results you can expect when retraining an existing LLM:

Increased accuracy as model trained on targeted datasets

Diversified inputs and outputs through multimodality for genAI

Get a free 5-minute AI maturity assessment to benchmark your organization's readiness.

https://go.turing.com/ai-maturity/assessment

TURING