TURING

WHITE PAPER

Maximizing Your LLM ROI: The True Costs of Poor Strategy, Evaluation, and Training

Table of Contents

Intro

With the rapid adoption of generative artificial intelligence (genAI), business leaders are often unsure where to focus their efforts to maximize their return on investment (ROI). When done well, AI has the potential to transform productivity and improve bottom-line performance.

Goldman Sachs projects that genAI could drive an additional $7 trillion in economic value over the next decade.

However, if leaders are unsure where to invest resources, AI investments may miss the mark.

The importance of LLM evaluation in training

Investing in training large language models (LLMs) is a key step in an AI project's success. However, this investment in training requires a systematic model evaluation. Otherwise, training may not yield effective LLM improvements. The case study below shows how missing this step can cause significant problems.

What happens when model training goes wrong?

For one large, fast-growing eCommerce business, artificial intelligence held the potential to streamline the entire customer service process. The company decided to develop a customer service chatbot solution, along with an underlying LLM, that could provide natural responses to customer questions. To support service agents, the firm developed a solution that could quickly surface information to increase the accuracy and efficiency of the agents' customer interactions. These value discovery projects would allow the company to demonstrate the ROI on selected priority use cases, and then decide how to scale based on these results.

As a pilot project, the initiative was put on a tight budget. Like many leaders who embraced genAI at scale for the first time, executives were skeptical that AI could make a true difference in their business. They were also worried about data provenance and controlling costs until their initial LLM investments showed a clear ROI. As a result, the leadership team thought they could save money on LLM evaluation and training. Because they didn't have previous LLM experience, the strategy they crafted had significant gaps that set the project on a path toward a poor outcome.

The AI customer service rollout was a success—at least initially. Users were eager to engage with chatbots, which were now available 24/7/365 and allowed them to interact on any device of their choice. In the short term, the project improved

customer satisfaction and reduced the daily workload that customer service support agents faced. At the same time, the support tool for the customer service agents saw some preliminary wins in helping the service team obtain vital information more quickly and reducing the time it took to resolve a customer question—a critical key performance indicator (KPI) for the business.

Over time, however, it became apparent that the current LLM had limitations. While it had passed the organization's internal evaluation benchmarks, the company was seeing troubling signs of poor quality in the model. Dissatisfied chatbot users complained that it was unable to understand certain queries and usage began to plummet. Customer service agents started to use these AI tools less because the tools were providing incorrect information to selected questions. With lackluster performance, dissatisfied users, and unusable results for certain use cases, the team reached a critical conclusion: The model was creating significant downstream costs and issues. They needed to find a solution.

Model retraining, fine-tuning, or enhancement was necessary because the model was failing, but that introduced another unknown set of variables into the process. These variables included how much this process would cost, how long it would take, and how additional training investments might impact users. The project's poorly crafted strategy had become a substantial stumbling block, especially as the team lacked the expertise and resources to solve this in-house. They needed insight on how to find the right partner.

Strategic training investment gives the smartest return for LLMs

It's understandable that organizations want to be conscious of costs and investments when implementing new technologies. However, training large language models is one area that shouldn't be skimped on. In fact, strategically investing in training can maximize the value of genAI in general and a specific LLM in particular. With the right partner, this can be a streamlined, cost-effective process.

That's true whether organizations are deploying a new model or upgrading the performance of an existing LLM.

For organizations launching a foundation model, considering how to assess the value of training is key. For organizations with an existing AI model that's not meeting expectations, there are options that can expand capabilities, reduce inaccuracies, and improve performance. Many executives navigating critical AI decisions need help to better understand the role that LLM training plays in capturing the value of genAI models.

For example, McKinsey notes that looking at investment through two lenses—cost savings and productivity improvements—can provide a framework for understanding the value of AI-related investments. For leaders, it's important to consider the following questions: What are the costs of poor LLM training? What's involved in turning around the performance of a lagging LLM? How can leaders think about LLM training investments to drive the overall ROI of their AI initiatives, or even avoid poor investments in the first place?

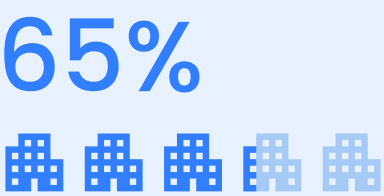
In the case study above, the team behind the customer service initiative brought in an outside partner to get an independent, third-party evaluation of their model, revealing performance gaps and improvement opportunities. By selectively introducing new data and working at the intersection of artificial and human intelligence, they were able to dramatically improve performance. Ultimately, the team found that these investments in training, development, and model enhancement were critical in reshaping how they approach future projects.

Let's examine the role LLM training plays in long-term value realization for AI projects, and how to make the right strategic decisions as part of an organization's overall AI implementation road map.

What makes a genAI model work, and where can it go wrong?

Introducing generative AI with large language models can offer dramatic improvements to processes, productivity, and an organization's ability to get more value from their data. The rapid adoption of genAI is showing this at scale.

A recent McKinsey study reported that 65% of respondent organizations use generative AI in some way, up from just one-third the year before. In the same study, 63% of respondents noted that inaccuracy and quality issues were their top concerns.



This highlights a critical issue: What makes AI large language models work well, and where do many organizations go wrong?

Building an effective genAI model has many variables, including how clearly the use case for the model was defined, how effectively the data was sourced and curated, and whether a robust AI technology infrastructure is needed to support performance. In addition, it's critical to look at the resources that went into defining what successful outputs looked like, how much training a model received, and how objectively model performance was evaluated.

Training can mean the quality of curated data used to train the model or the caliber of human feedback received from developers, researchers, trainers, and users.

What you're trying to accomplish with your model impacts model evaluation, what goes into making it work well, and how you might structure it. For example, an AI coding assistant or copilot requires broad training data in order for it to produce clean, accurate, and well-reasoned code in a language like C++ or Python. A real-time analytics and predictive tool that's designed to help a think tank identify emerging trends may be more influenced by the breadth of external data sources that it can ingest through agentic workflows. While the use cases and inputs vary, the underlying theme is the same: It's essential to make the right investments in the right data, training, development, and evaluation early on so that a genAI model works as intended and produces increasingly useful results.

Factors that influence how well an LLM works include:

Use case definition
How well was the use case for the model defined, and how is that reflected in the coding and reasoning that drives the model?

Source data
Data is key, whether you're using an existing public LLM or building a model based on proprietary data. Data quality, sourcing, and management are the critical inputs that drive a model's performance. Beyond data hygiene, relevance is critical when training datasets. Other considerations may include ensuring broad datasets are in place if required—for example, using agentic workflows that incorporate APIs when necessary.

Coding and reasoning
The quality of the coding and reasoning that provides the model's underlying architecture and parameters is crucial. Critically evaluating these factors and looking for opportunities for improvement can significantly impact LLM performance.

Domain knowledge
Are the training datasets, experts defining the model parameters, and all others involved in applying training methodologies to the LLM reflective of relevant industry or use case expertise?

Model evaluation
As the LLM was tested and rolled out, was structured model evaluation performed? Was the evaluation completed in-house or by an experienced third-party source? Did the criteria used in the evaluation look at areas such as performance, fluency, accuracy, and relevance?

User experience
Is the interface easy to use? When issues occur, factors such as response time and error recovery timelines should also be considered.

Training
A variety of training methodologies, including data-driven training and human feedback, can help create and refine models. The methods used and the expertise and talent behind the training all impact model quality—whether LLM training occurred initially or as part of an effort to refine the model.

A host of factors influence how well genAI models perform, but there's an underlying theme that ultimately determines the coherence, relevancy, accuracy, fluency, and safety of an ideal model. A genAI model stands the best chance of delivering significant long-term value if an organization has been thoughtful at each step of the process, applied best practices, and incorporated initial training or ongoing evaluation and refinement.

How an AI research organization used LLM training to transform its model

When an organization integrates an LLM into its business, it must be able to demonstrate ROI.

As companies move from the value discovery phase of trial projects to the value realization of genAI tools for critical use cases, they may find that improvements are needed to remediate a poorly performing model or that enhancements are required to expand the model's capabilities.

When one U.S.-based AI research organization wanted to transform its LLM performance, it evaluated how fine-tuning, training, and development could help achieve these goals. AI initiatives had made significant contributions to the organization's efficiency and ability to make informed business decisions, but it was clear that enhancing AI model capabilities with real-time data analysis would offer significant competitive advantages. The organization's goal was to improve accuracy, adaptability, and personalized insights to empower users with even stronger results. The firm partnered with an organization that specializes in training large language models to make this vision a reality.

As the two firms worked together to evaluate the most urgent use cases, the foundation model state, and desired business and technical outcomes, they developed a plan to improve the model. Integrating multiple APIs with the existing AI model infrastructure would enable the model to offer real-time data fetching and analysis capabilities from diverse sources, yielding more accurate and useful results.

By introducing multimodality—or the ability to incorporate different data types including text, images, video, and more as inputs and outputs—genAI can deliver deeper insights and even produce data outputs in a variety of forms.

The team conducted an extensive analysis of the existing model and established clear KPIs for measuring improvements. With an expansion road map in place, they worked on identifying and integrating various data sources, including APIs from different models and services. From there, they seamlessly integrated genAI and LLMs into the existing AI model architecture to expand the capabilities. While this alone offered measurable improvements, the team continued to dive deeper to deliver even more significant value by adding real-time analysis. They configured the AI model to operate in real-time, with coding improvements and testing to support the processing and analysis of data from over 200 distinct APIs. With access to broader data sources and improved real-time analysis capabilities, intricate data synergies evolved into workflows that can stack multiple APIs and offer faster, smarter, and more nuanced analysis. A team of developers and research experts continuously tested and optimized the model for accurate, effective performance that aligned with the project's initial expectations.

The organization ultimately created a genAI system that's capable of continuously monitoring incoming data for patterns, anomalies, and trends. Alerts were created so that real-time alarms notify key researchers and decision-makers of events or trends as they emerge, providing an important competitive advantage. At the end of the project, this dynamic transformation enabled the enhanced model to handle 1,600 unique use cases, analyze real-time data to offer key insights, and provide greater agility to make fast, informed business decisions.

By investing in additional training and model enhancement, the organization turned its AI model into an important competitive advantage that allows it to stay agile and make strategic decisions based on real-time insights and emerging trends.

How to boost the performance of your LLM for a genAI application

If an organization has invested in generative AI, it's possible to improve a model that's performing poorly or expand its capabilities for broader use cases.

Fine-tuning, model training, and enhancement can help bridge the gap from a model's current state to an ideal future state. Let's take a closer look at different elements that may be included in the improvement process, depending upon the challenges facing a specific organization and LLM.

Improving model performance begins with third-party model evaluation. A common challenge organizations face is that their evaluation and benchmarks are conducted internally. Teams conducting these checks may lack expertise, objective perspective, and industry best practices. A reputable third party should ideally conduct model evaluation as a checks and balances system that can avoid internal bias or tunnel vision. An external third party's expertise and insights can help identify a model's strengths and weaknesses by comparing it to industry best practices. From there, the team can identify gaps, determine if additional data or human training is required, and outline opportunities for strategically expanding the LLM's capabilities.

Once an evaluation has uncovered an LLM's challenges, there are a wide variety of options available to improve model performance, including:

Multimodality

Many of the familiar genAI models today are text-based. However, through strategic multimodality implementation, it's possible to introduce the ability to accept inputs and produce outputs in other formats, such as visuals, video, voice, and more. Multimodality can help increase a model's usefulness and add a dynamic, agile element to enable users to create outputs that can be applicable to a range of different use cases.

Coding

For genAI models designed to act as coding assistants or co-pilots, improving the underlying code and reasoning can help. In some cases, a broader set of training data related to the language in question can be introduced to help the LLM develop usable code for a variety of problems and situations. In others, reasoning improvements can help more closely map the code produced to desired business outcomes as well as enhance problem-solving capabilities within the model to address coding challenges.

Reasoning

Many AI models rely on reasoning to create tasks, such as creating a menu or providing a trip itinerary. Coding can be used as the basis of reasoning improvements to help assistants better navigate steps such as planning, tooling, and workflow.

Factuality

While AI "hallucinations" have captured headlines, the reality is a model that lacks access to quality data and the right feedback may produce inaccurate results. Factuality improvements may include data cleanup, source expansion, and human feedback cycles to ensure that the outputs produced are accurate and grounded in facts.

Agents and function calling

Agentic workflows incorporate APIs to automate specific tasks. These calls can include private APIs, such as Salesforce or ServiceNow, or access to public APIs, like governmental data sources. By improving sourcing, streamlining integration, or improving supporting workflows, accuracy can be increased and overall performance improved.

Domain expertise

When organizations develop LLMs and tools for highly specific industry use cases, domain expertise is crucial. For example, creating an effective model for financial trading is going to be vastly different from drug discovery in pharmaceuticals.

In addition to these use cases, a vast array of options can also help systematically improve LLM performance. For example, models can be enhanced to be more useful in specific settings by introducing supervised fine-tuning (SFT), with feedback provided by experienced industry professionals. SFT involves directly training the LLM with a carefully curated set of data that's been labeled and annotated to better inform the model. Reinforcement learning from human feedback (RLHF) is a reward-based learning process that uses human feedback to iteratively improve genAI model performance. Direct preference optimization (DPO) is another way to improve models, using direct human feedback to select the best text response or image choice while bypassing feedback loops.

The current state of the project, relevant business objectives, and available resources serve as guides for businesses looking to understand how to best train large language models for specific cases. Ultimately, these factors can come into play with base model training, improving an existing model in a specific area, or a holistic transformation designed to dramatically enhance existing overall model performance.

Building a cohesive LLM training process

With many potential solutions available to improve model performance, it's helpful to understand how each element can come together into a cohesive approach. Product, people, and processes make up the trifecta of elements that help create an LLM training and enhancement approach that's repeatable, scalable, timely, and cost-effective.

For example, Turing follows this clearly defined process.

During the initial project conversations, the Turing team brings its deep LLM strategic and technical expertise to understand your unique needs and objectives. With these insights, Turing crafts a tailored plan to evaluate, train, and enhance your LLM. Once the general LLM training road map is approved, Turing scopes, budgets, and builds timelines to execute the training. The process begins with a model evaluation phase to systematically understand the LLM's current state and create a dynamic work plan to address specific issues.

Talent from the client side—including executive stakeholders, researchers, and data analysts—partners with Turing's team. The LLM team consists of machine learning engineers, data scientists, and natural language processing experts who design, train, and optimize the model. The team also includes software engineers for infrastructure and tool development, cloud computing experts for managing computational resources, and ethics and bias experts to ensure fairness and accountability. Depending on the project, domain experts and computational linguists may also contribute. Turing's extensive on-demand talent platform allows for the rapid onboarding of hundreds of LLM trainers in weeks versus months.

LLM trainers start with a series of defined tasks on secured endpoints. Teams are closely managed, and their work is constantly evaluated for quality and progress. As tasks are completed or new priorities emerge, trainers are reallocated to ensure progress and thoroughness.

The management program includes weekly self-improvement solutions, ensuring resources are used efficiently and timelines are met.

If you bring in an external partner for base training on a model built with public data, the assembled team identifies areas of improvement. This may involve introducing multimodality for greater value or enhancing reasoning capabilities for better outputs. Prioritization is guided by feedback from all stakeholders, and the team works through tasks to completion.

For fine-tuning an existing model for a new use case, the process is slightly different. After onboarding LLM trainers, the team creates evaluative data for testing, using human expertise as the benchmark. Through the SFT process, new data is introduced, refined, and evaluated to build the necessary expertise. Subsequently, RLHF cycles assess the performance, generate new data requirements, and introduce technical improvements. These cycles continue until project completion, and often become part of an ongoing model improvement process.

While every project has unique requirements, core models, and desired outcomes, a strong approach that delivers improvements every time is one that's built on proven people, processes, and products.

How much does LLM training cost, and how long does it take?

Once organizations decide to improve the performance of their large language models, the next questions are often: How much does it cost, and how long will the project take? This is decidedly specific to the business and model in question. However, a closer look can offer context for how these come together to shape both budget and project timelines. There are several influencing factors, including:

What is the current state of the LLM?

What existing model evaluation has been completed, and what is required as part of the project scoping?

What performance gaps have been identified, what's the ideal future state or desired goal, and what's involved in getting there?

Is the primary solution to addressing these challenges introducing new data sources, or will the model require significant layers of human feedback to improve outcomes?

What type of talent is required to complete the project? For example, a project that needs science, technology, engineering, and mathematics (STEM) researchers with highly niche knowledge may have a higher cost.

What scale and timeline are you hoping to achieve? Large teams can often move through projects faster, but staffing up to scale can add to the budget.

What internal resources and expertise can you commit to the project to work in tandem with the provider? Consider what happens with your team's overall productivity when you commit these resources to the project instead of other places.

How extensive is the required testing and refinement, and what KPIs serve as success benchmarks? The more advanced and specific, the longer it can take to complete.

It's helpful to have a general idea of the factors that a partner will take into consideration when scoping your project, pricing it, and developing a project timeline. If you have specific project constraints—for example, staying within a particular budget range or delivering within an accelerated timeline—surface those issues early so the prospective partner can outline a pathway to meet those objectives.

How to choose the right partner for your LLM needs and get started

If you've invested in a model and identified the need for improvements, you're at a critical inflection point in your genAI journey. You may have even taken some initial steps, only to find that the results you've generated thus far offer room for improvement. Select a knowledgeable partner who understands all elements of LLMs, has extensive capacity to source data or address development issues, and can also deep dive into your unique business and industry needs to translate genAI into real business results.

When choosing a partner, what are the critical components to consider? Important factors include:

- What's the organization's expertise and have they worked with companies that you recognize?
- Does this organization have experience working with LLMs like yours, whether it's a proprietary model or a fine-tuned model based on public data?
- What does this organization's model evaluation framework and process look like?
- Do they have the capabilities to complete the tasks you're likely to need, whether that's research expertise to determine appropriate datasets or technical capabilities to improve LLM reasoning?

- What's the organization's staffing model, and how might that impact their ability to scale?
- What's the general cost and timeline the organization is able to deliver within and will that work for your business plan?
- Do they have experience as a services provider that can work as both a thought leadership partner and execution vendor to help make your vision a reality?
- What resources, customer testimonials, and other information can you find to better understand their leadership positioning in the LLM space, as well as the impact of their work?
- Do they have a well-documented and managed project process that you feel confident they can deliver?

Once you've identified a prospective vendor, the process will typically begin with a consultation with an expert at the organization to better understand your firm's background.

After the consultation, there's likely to be a model evaluation and discovery phase where the vendor better understands what's currently happening with your model. From there, they'll identify the gaps and improvement opportunities that will shape your LLM training project. The project itself can take weeks to months, depending on the scale. Be sure to get clarity on how they manage the process, communicate with clients, and deal with unexpected changes to scope.

Selecting the right partner can involve some due diligence. However, training your LLM is a serious undertaking that can have profound impacts on your organization. A knowledgeable team can serve as a thought leadership partner that helps develop and implement innovative solutions using best practices with technology and data to close performance gaps.

TURING

Next steps for training your LLM

Early adopters of genAI have the potential to capture significant competitive advantages and innovate and enhance the efficiency of key business processes. However, LLM training is central to long-term, sustainable value creation. Even if you have a poorly performing model, it's possible to take action and transform that outcome. By selecting a world-class LLM training partner, you can secure an impartial model evaluation, streamlined model training process, and results that deliver bottom-line impact for your business.

If you're ready to explore how LLM training can transform a poorly performing model into a competitive advantage, contact Turing today. Speak with a solution expert to help evaluate your needs and learn how our LLM services team can make your AI-accelerated visions a reality.

Learn more at turing.com