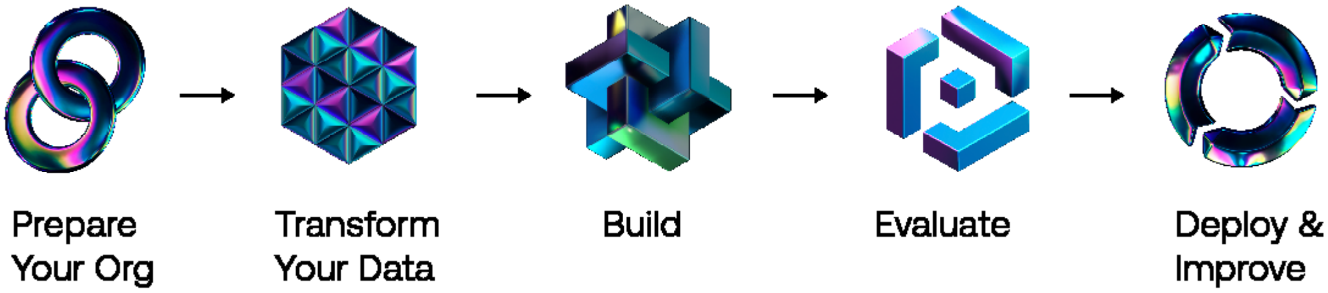
# Blueprint to AI: Model Evaluation Checklist for Enterprise

# Blueprint to AI: Your Generative AI Implementation Checklist



## Table of Contents

2024 has been the year of demanding ROI from Generative AI investments in the enterprise. Companies turned to Retrieval Augmented Generation (RAG) and fine-tuning to help them get more value out of Generative AI and differentiate themselves from their competition. However, many enterprises lack the expertise, tools, and framework needed to build customized Generative AI models and applications at scale.

Scale has delivered custom solutions for leading enterprises across a wide variety of industries and use cases. We have turned that experience into this Blueprint to AI: Your Generative AI Implementation Checklist, with the goal of helping you make 2025 the year deploying Generative AI to production and delivering real business value across your enterprise.

Throughout this guide we will cover the essential Blueprint to AI, providing the steps you need to take to be successful and the questions that you should be asking yourself along the way.

# Evaluate

## STEPS

□ **Create evaluation datasets.** Ideally use a hybrid human and synthetic approach to generate a prompt-response pair evaluation dataset. Use human datasets if you're targeting accuracy or depth. Use automated datasets for breath, coverage, and scale.

□ **Identify a healthy mix of automated and human-in-the loop testing,** evaluation, and monitoring. We suggest hybrid approach. To determine the evaluation methodology needed, use the following chart as a baseline. Then add custom evaluations based on your use case.

| | |
|---|---|
| **Completeness** | Does the response fully answer all explicit aspects of the prompt? Does the response omit any essential information? |
| **Accuracy** | Which response contains the most accurate and reliable information, and which aligns with established facts or evidence? Established facts can be from common knowledge or verified using the provided context. |
| **Depth** | Which response provides a higher level of detail, insight, and nuance? |
| **Relevance** | Which response provides the most useful supporting information and claims in answering the main question or prompt? The supporting information logically defends or clearly illustrates the key points and the central claims made in the response. |
| **Clarity** | Which response is more clearly worded and understandable? |
| **Formatting** | Which response organizes the information in a way that supports the key points with clarity and brevity? |

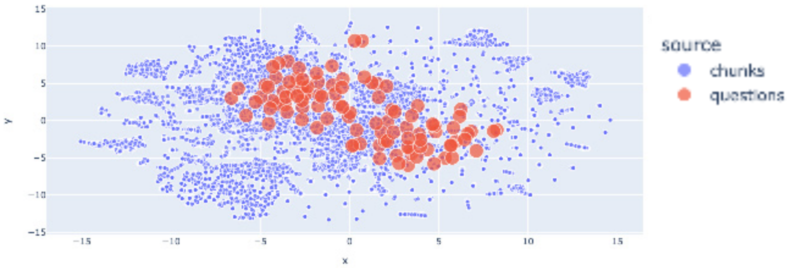
□ **Establish benchmarks and target metrics** for each type of testing.

□ **Leverage your expert employees to test your customized solutions** to make sure they perform well at the specific task.

□ **Audit results from evaluation runs** and use these to iteratively improve your RAG/Fine-Tuning/Prompt Engineering implementations for continuous performance and safety improvement.

□ **Ensure full test coverage.** Your evaluation datasets should cover all of the ways that users will interact with the model for each specific use case.

□ **Monitoring and hill-climbing for continuous improvement.** Monitor how users are actually prompting the model and how the model responds to prompts. Leverage embedding visualizations to ensure coverage of actual user prompts. Then hill-climb to continuously improve the customized model and ensure that the model is able to adequately address the most common prompts.



Sample embedding visualization which shows coverage of user prompts

# Evaluate

## QUESTIONS TO CONSIDER

- → **Do you have a plan to create your prompt-response pair evaluation dataset?** Have you considered a hybrid approach combining human-generated and synthetic data?
- → **Have you established a balanced mix** of automated testing and human-in-the-loop evaluation for your Generative AI solution?
- → **Do you have a methodology chart or framework to guide your evaluation approach** for different aspects of your solution?
- → **Have you set clear benchmarks and target metrics for each type of testing** you plan to conduct?
- → **How are you involving your expert employees in testing** the customized solutions to ensure they perform well for specific tasks?
- → **What process do you have in place to audit results from evaluation runs** and use them to iteratively improve your RAG, fine-tuning, and prompt engineering implementations?
- → **How do you ensure that your evaluation datasets provide full coverage** of all the ways users might interact with the model for each specific use case?
- → **Have you implemented a system to monitor how users are actually prompting the model** and how the model responds in real-world usage?

# Conclusion

We hope this checklist proves helpful on your journey from moving from Generative AI POC to production. By following the steps outlined—from preparing your organization and transforming your data to building, evaluating, and deploying your Generative AI applications —you now have a solid foundation to overcome common hurdles that many enterprises face when developing and scaling custom Generative AI solutions.

Implementing Generative AI at an enterprise scale involves numerous nuances and critical decisions that require specialized expertise. The complexities of readying your organization for GenAI adoption, preparing your data, fine-tuning models, generating evaluation datasets, and optimizing for specific use cases can be challenging. Working with an experienced partner can provide invaluable insights, accelerate your implementation, and help you achieve your business objectives.

Scale has extensive experience in delivering custom Generative AI solutions for leading enterprises across diverse industries and use cases. We're ready to help you take the next step and turn this checklist into action by helping you build, evaluate, and deploy custom Generative AI solutions that drive real business value.

# About Scale

Scale is fueling the generative AI revolution. Built on a foundation of high-quality data and expert insight, Scale powers the world's most advanced models. Our years of deep partnership with every major model builder enables our platform to empower any organization to apply and evaluate AI.

With our Custom Generative AI Solutions for Enterprise, we help leading enterprises build, evaluate, and deploy custom, production-ready Generative AI solutions that drive real business outcomes. Customers choose Scale for our extensive Generative AI expertise, white glove service, proven results, and deep partnerships with model builders such as OpenAI and Meta, cloud service providers such as Azure and AWS, and consulting firms such as BCG and Accenture.

The custom Generative AI solutions that we build are powered by Scale Generative AI Platform, the leading platform to build, test, and optimize Generative AI applications that unlock the value of your data. Custom solutions built on Scale GenAI Platform give you everything you need to accelerate your Generative AI Journey, scale use cases across your organization, and achieve real business value.

To learn more about how Scale can support your Generative AI initiatives, visit scale.com/enterprise/generative-ai-solutions.