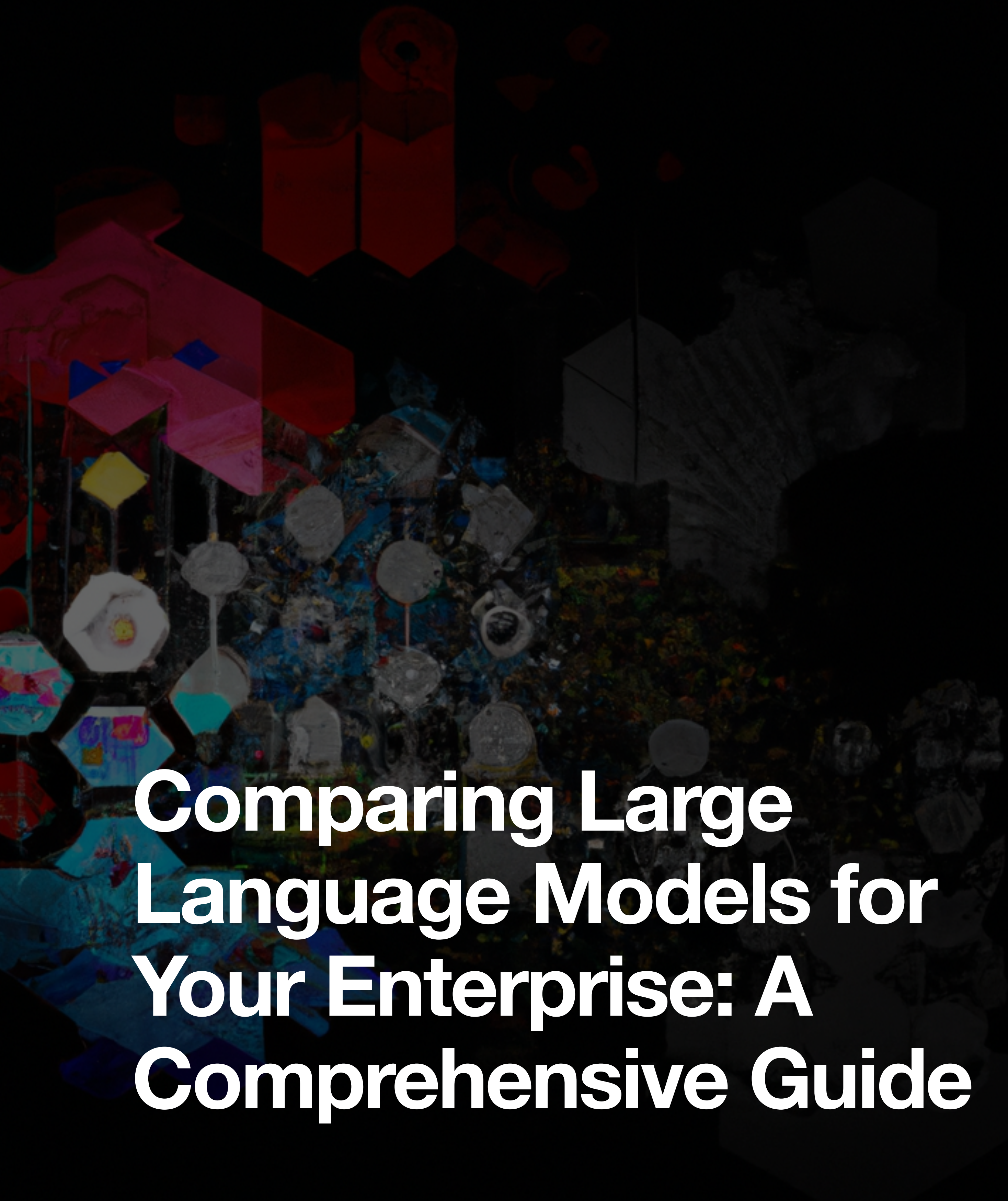
Krista



# Comparing Large Language Models for Your Enterprise: A Comprehensive Guide

# Table of Contents

Introduction 2
Large Language Models are Changing Businesses 2
Typical LLM Use Cases 3
Contact center assistants or AI agents 3
Intelligent customer self-support capabilities 3
Utilizing LLMs for IT Support Management 4
Assisting Employees and Human Resource Professionals 4
Revolutionizing Healthcare 5
How to Choose the Right LLM or LLMs 5
Balancing LLM Capabilities with Costs 6
Testing Three Popular LLMs 7
LLM Test Procedure 7
LLM Test Prompts 8
Questions for LLMs 9
LLM Test Outputs 9
ChatGPT-4 Output 10
ChatGPT-3.5 Output 10
Watsonx Output 11
Summary of Test Results 12
Watsonx Results 12
ChatGPT-4 Results 13
ChatGPT-3.5 Results 13
LLM Innovation Requires Regular Testing 14
How to Conduct Your Own LLM Test 15
Summary 16
Footnotes 16

# Introduction

Generative AI and large language models (LLMs) are transforming human-to-computer interactions opening up groundbreaking possibilities for enhancing both customer and employee experiences. OpenAI's ChatGPT is at the forefront of this revolution. However, it is not the only choice. There are many other generative AI and LLMs available each with their strengths, weaknesses, and costs.

Today, enterprises are grappling with which models to use and how to cost-effectively incorporate them into their existing infrastructures and specific use cases. The purpose of this paper is to provide a comprehensive guide and insightful analysis of how several LLMs fared in a side-by-side test for a knowledge management use case. We tested three different LLMs using a constant document set and collated the results to help you make informed decisions about which LLMs may be best for your business and how to conduct your own test. By reading this paper, you'll have a clear understanding of the capabilities of the LLMs we tested, their strengths, potential shortcomings, and exercises to determine the costs involved. Our results and testing methods can be replicated to assist your organization in selecting and integrating the most suitable LLM to maximize return on investment and transform your customer and employee experiences.

# Large Language Models are Changing Businesses

Large Language Models (LLMs) are fundamentally reshaping the way enterprises operate and interact with their customers. These advanced AI models are capable of understanding and generating human language, enabling businesses to automate and enhance a wide range of processes - from customer support and content creation to data analysis and decision-making. With LLMs, businesses can provide instant and personalized customer service around the clock, generate insightful reports and summarizations from vast amounts of data, and make informed decisions quickly and accurately. Furthermore, LLMs can be trained or fine-tuned to understand specific industries and company jargon, making them highly adaptable across different business contexts. The importance of LLMs lies in their potential to drastically improve operational efficiency, provide unparalleled customer experiences, and give businesses a significant competitive edge. This makes the integration of LLMs into the enterprise technology stack no longer a luxury, but a necessity in today's fast-paced and data-driven business landscape.

# Typical LLM Use Cases

LLMs are vastly versatile, providing functionality that can be leveraged across a multitude of business domains. Here are a few common use cases:

## Contact center assistants or AI agents

LLMs can be deployed inside virtual assistants for contact center agents, offering real-time help in sifting through and summarizing vast amounts of data, navigating multiple systems, and generating insightful responses to customer queries and emails. With their ability to process and understand human language, these models can quickly generate responses after retrieving relevant information from various sources including CRM systems, databases, and knowledge bases. They can draft initial responses based on this information, which the agents can then refine and personalize before sending. This not only speeds up response times but also ensures consistent, accurate, and personalized service, leading to improved customer satisfaction.

A forward-thinking North American eCommerce and retail chain integrated LLMs and automation with Krista to improve customer service.[1] This store has an enormous parts inventory and therefore has very complex orders to repair both antique and modern transportation equipment. Customer service reps were forced to navigate several order management and ERP systems to service customers. Many times responses were delayed more than 24 hours causing customers to become frustrated by long wait times. Operational leaders sought AI solutions and found Krista could integrate with ERP and order management to automate processes and use LLMs to help generate intelligent email responses to customers, delivering immediate and accurate customer communication. The result? A significant reduction in operating costs as the company moved more tasks to AI software. But the benefits didn't stop at cost savings. This modernized approach has dramatically improved customer service quality, delivering faster resolution times and earning positive customer satisfaction ratings.

## Intelligent customer self-support capabilities

In the realm of self-service, LLMs significantly upgrade the capabilities of automated customer support portals and chatbots. These advanced models can understand and respond to a wide array of customer queries, ranging from simple questions about a product or service to more complex troubleshooting scenarios. By analyzing the context and intent of customer inquiries, they can provide accurate and detailed responses, or suggest step-by-step solutions to resolve issues when prompted correctly. Integrated with an organization's knowledge base and customer support systems, LLMs are capable of generating personalizing responses based on a customer's history and preferences. This intelligent self-support capability not only expedites issue resolution but also makes the customer experience smooth and effortless. As a result, it notably enhances customer satisfaction, while reducing the load on live agents and support resources.

## Utilizing LLMs for IT Support Management

LLMs can alleviate a lot of stress in IT support management (ITSM). These AI-powered assistants help serve as 'Virtual Engineers', capable of revolutionizing the way ITSM teams operate and support requests are handled.

Traditionally, ITSM personnel have been tasked with managing and resolving a vast array of IT support requests. From minor issues like forgotten passwords to more complex scenarios like system outages, and system access the demand on IT support teams is constant. Plus, the sheer volume of support requests and the need for round-the-clock availability stresses teams and causes burnout.

Enter LLMs. With their ability to understand and generate human language, LLMs, along with automation software, can help summarize data and simplify a significant part of ITSM operations. They can understand the nature of IT support requests, sift through relevant knowledge bases, and generate detailed, step-by-step solutions to resolve the issues. If the issue is complex and needs human intervention, automation can escalate the request to the appropriate IT personnel or manager to assist or approve the next best action.

A compelling example of this is seen in the case of Zimperium, a Dallas-based enterprise software company specializing in mobile security.[2] Zimperium had a complex and time-consuming software deployment process that involved rigorous approval cycles, decision gates, and high-touch coordination between multiple teams. Zimperium incorporated LLMs using Krista's AI-powered intelligent automation platform, enabling engineers to simply "Ask Krista" for a business outcome. Krista handles all subsequent tasks, such as creating IT tickets, sending notifications, ensuring adherence to development cycles, and reviewing exceptions.

The integration of Krista not only streamlined Zimperium's deployment process but also led to substantial cost savings. A procedure that previously required over four hours was reduced to mere minutes, resulting in a 400% ROI and freeing engineers and developers to focus on core tasks. This is a testament to the transformative power of LLMs and generative AI.

## Assisting Employees and Human Resource Professionals

Generative AI and integrated LLMs are revolutionizing human resource (HR) operations driving greater efficiency and transforming employee experiences. LLMs empower HR assistants enabling them to provide comprehensive, instant employee support and reducing the load on HR teams. One of the primary benefits of HR assistants is their ability to streamline complex processes like onboarding, changing benefits, or scheduling time off through natural, conversational interactions.

For instance, new hires can engage with an AI-powered HR assistant to walk them through the intricate onboarding process, providing real-time answers to their questions and guiding them step-by-step in a personalized, user-friendly manner.

Similarly, existing employees can use the system to navigate through the nuances of changing their benefits or scheduling time off based on availability and other constraints found in the HR, scheduling, or production schedules. This not only makes such processes more accessible and easy to understand but also ensures they are completed correctly and efficiently.

## Revolutionizing Healthcare

LLMs can power virtual agents to provide self-service wellness programs, delivering personalized, accessible, and convenient healthcare advice to patients or employees. This allows for improved patient and employee engagement and adherence to wellness programs, enhancing health outcomes and well-being.

Integrating LLMs and generative AI in healthcare, especially with current NLP, and speech-to-text capabilities can revolutionize processes, outcomes, and experiences for all parties involved - back-office staff, healthcare providers, and patients, making it a win-win proposition for the entire healthcare ecosystem.

A leading healthcare application is helping customers shift from "Healthcare to Wellcare", centralizing patients and physicians in an interconnected healthcare experience.[3] The application leverages advanced technology to incentivize positive behavioral changes while maintaining stringent security measures.

The healthcare app integrates the full capacities of an AI/ML solution, including access to a public catalog of connectors, Language Translation Services, and Intelligent Document Understanding, to offer patients and healthcare providers an innovative approach to health and wellness management.

Through its mobile interface, the app enables the capture of images of Aadhar cards (a unique identifier for Indian citizens). These images are processed by the integrated Intelligent Document Understanding feature to extract and parse relevant identity information, which is then automatically updated into the Health Wallet feature within the app.

The initial phase of this project focuses on developing a solution that incorporates LLMs, a Chatbot, integration with EHR/EMR systems, and Intelligent Document Understanding.

# How to Choose the Right LLM or LLMs

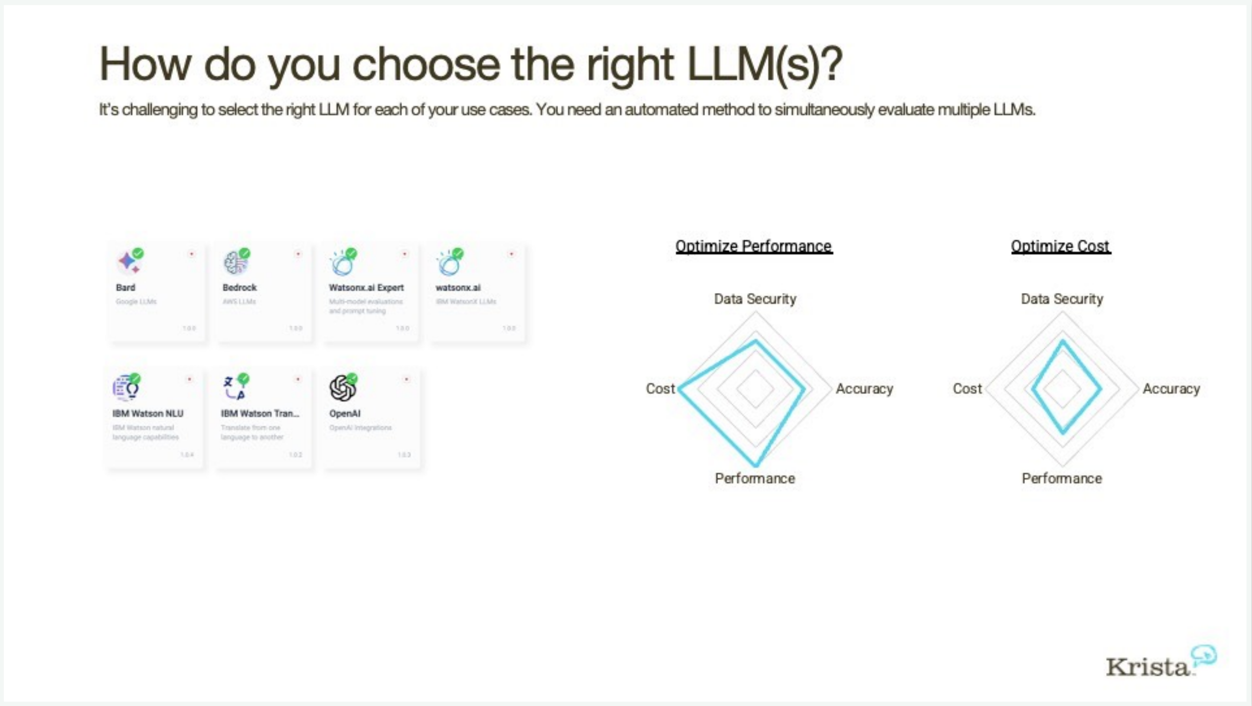
### Why one size doesn't fit all

OpenAI has indeed emerged as a market leader with its advanced LLMs, captivating attention with its groundbreaking capabilities and potential applications. However, it's important to remember that "one size does not fit all" when it comes to implementing

LLMs into your company. OpenAI's offerings, while powerful, may not always be the best fit for all use cases or budgets.

# Balancing LLM Capabilities with Costs

Cost matters. And when it comes to choosing the right model, the price can be an influential factor as any. But you should think beyond just the upfront costs to consider data security, accuracy, and performance as well. It's not merely about choosing the 'biggest' or 'best' model available because that may be too expensive for your use case. For instance, comparing the output costs of two models from OpenAI demonstrates that ChatGPT-4 is thirty times more expensive than ChatGPT-3.5.[4] Building processes with an LLM that is thirty times more expensive could prove costly as your employees and customers increase usage. A less expensive model could prove more accurate for your use cases or more secure so you will need to conduct tests to find the right fit per use case. However, it is important to note that while cost may be a factor, it should not be the only one. LLMs frequently handle specific tasks and use cases better than others so picking the right one is essential to ensure optimal outcomes. For instance, in our testing, we found one LLM performed better when generating content from tabular data than others, while others generated more concise content. Therefore, you need to consider both "cost" and "capability" when selecting an LLM and to do so you will need an automated method to constantly evaluate LLM capabilities since they innovate so quickly.



How do you choose the right LLM(s)?

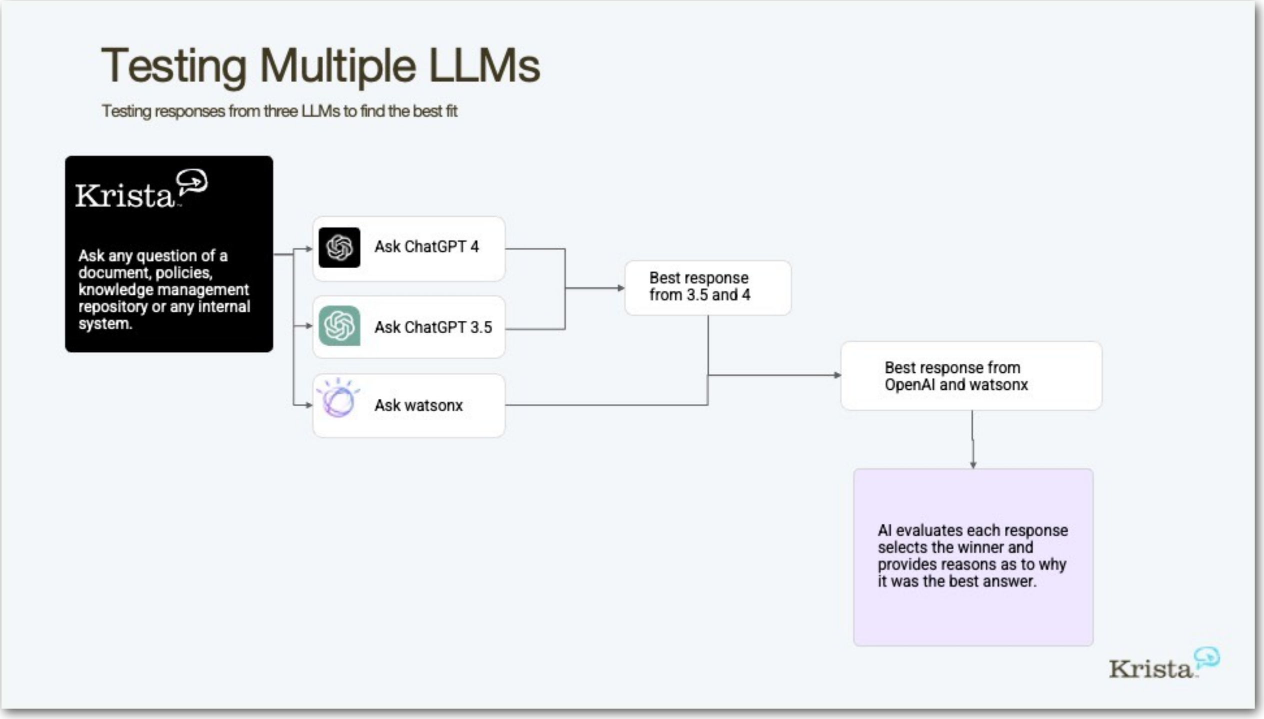
It's challenging to select the right LLM for each of your use cases. You need an automated method to simultaneously evaluate multiple LLMs.

# Testing Three Popular LLMs

Choosing the right LLM for your company can be a challenge. To provide a comprehensive perspective, we decided to evaluate three popular LLMs, each with their unique technical capabilities and levels of sophistication. This comparative analysis aims to highlight the outputs of each model to formulate a side-by-side comparison and then have AI judge the best output and explain why it chose that particular answer. Understanding why a specific answer is the best fit will help you determine the best LLM for your use case. If you need short answers, then one model may be better for you. If you need greater detail and context then another may be a better fit. The goal is to guide you in how to select the most appropriate LLM that aligns with your specific use cases.

# LLM Test Procedure

To objectively evaluate the performance of the three selected Large Language Models, we employed a thorough and consistent testing strategy using our automation platform, Krista. Krista is an AI-led intelligent automation platform used to integrate and orchestrate business processes across your people, systems, and any AI services. Using Krista greatly simplifies an LLM test since she possesses an easy-to-use conversational interface and each of the LLMs is already integrated into the platform.



Testing Multiple LLMs

Testing responses from three LLMs to find the best fit

Krista

Ask any question of a document, policies, knowledge management repository or any internal system.

Ask ChatGPT 4

Ask ChatGPT 3.5

Ask watsonx

Best response from 3.5 and 4

Best response from OpenAI and watsonx

AI evaluates each response selects the winner and provides reasons as to why it was the best answer.

Krista

Figure 1

This test we conducted mimics an employee, HR assistant, or knowledge management use case mentioned above. This is one of the first use cases enterprises evaluate since they already possess lots of documents for policies, procedures, products, and services. In lieu of conducting the test on proprietary data, we chose to run the test against a public document. Doing so allows you to read the source document, our inputs and prompts, the LLM outputs, and the output summaries so you can objectively evaluate our analysis. The standardized corpus document we used for this test is McKinsey & Co.'s, 'The state of AI in 2023: Generative AI's breakout year' report.[5]

In our comparative analysis, we chose to put three renowned LLMs to the test - OpenAI ChatGPT-3.5, OpenAI ChatGPT-4, and watsonx.[4,6] ChatGPT-3.5 and ChatGPT-4, products of OpenAI, have been garnering attention for their advanced language understanding capabilities, but they differ in terms of cost and scale. Watsonx, on the other hand, is an offering by IBM providing access to IBM-selected open source, third-party, and a family of IBM-trained foundation models of different sizes and architectures.[7] Our objective was to assess these models in the same environment and with the same inputs, thereby providing a fair and balanced comparison to help enterprises make informed decisions.

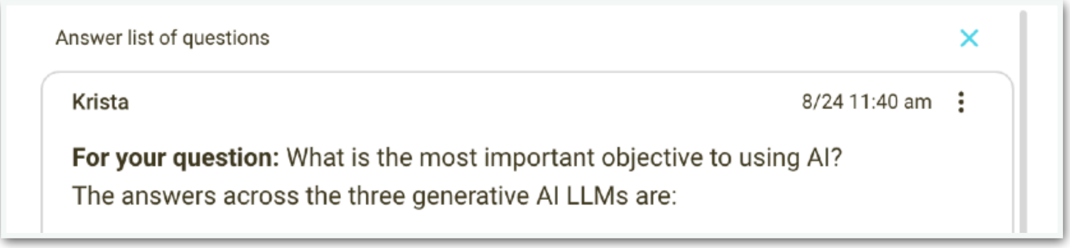
The test is set up in a bracket-style comparison similar to a tournament. First, we prompt ChatGPT-3.5, ChatGPT-4, and watsonx with the same data using Krista. Then, Krista judges the outputs from ChatGPT-3.5 and ChatGPT-4 and determines a first-round winner. The winning output of the first round is then compared to the watsonx output and Krista again evaluates the outputs and determines the best output or the second-round winner. (Figure 1) After the second round winner is identified, Krista informs on why this output is the best one based on all of the circumstances.

# LLM Test Prompts

The testing framework we devised centered around ten distinctive prompts or queries, each constructed to prove the LLMs' abilities in language comprehension, context decoding, and response formulation. We constructed the prompts to determine which LLM excels at summarization tasks, which can effectively interpret tables and data to derive valuable conclusions, and whether AI can stack rank data from the source document. The specific questions and prompts we utilized in our experiment are outlined below.

# Questions for LLMs

1. How many people say they've had at least some exposure to generative AI?
2. What are the most common use cases?
3. What are the main concerns of using AI?
4. What is the most important objective to using AI?
5. How will workforces change after implementing AI?
6. Which department is most likely to shrink?
7. Are companies expecting to reskill workers?
8. How many companies have adopted AI?
9. What benefits are companies realizing?
10. Can AI help increase revenues?



**For your question:** What is the most important objective to using AI?
The answers across the three generative AI LLMs are:

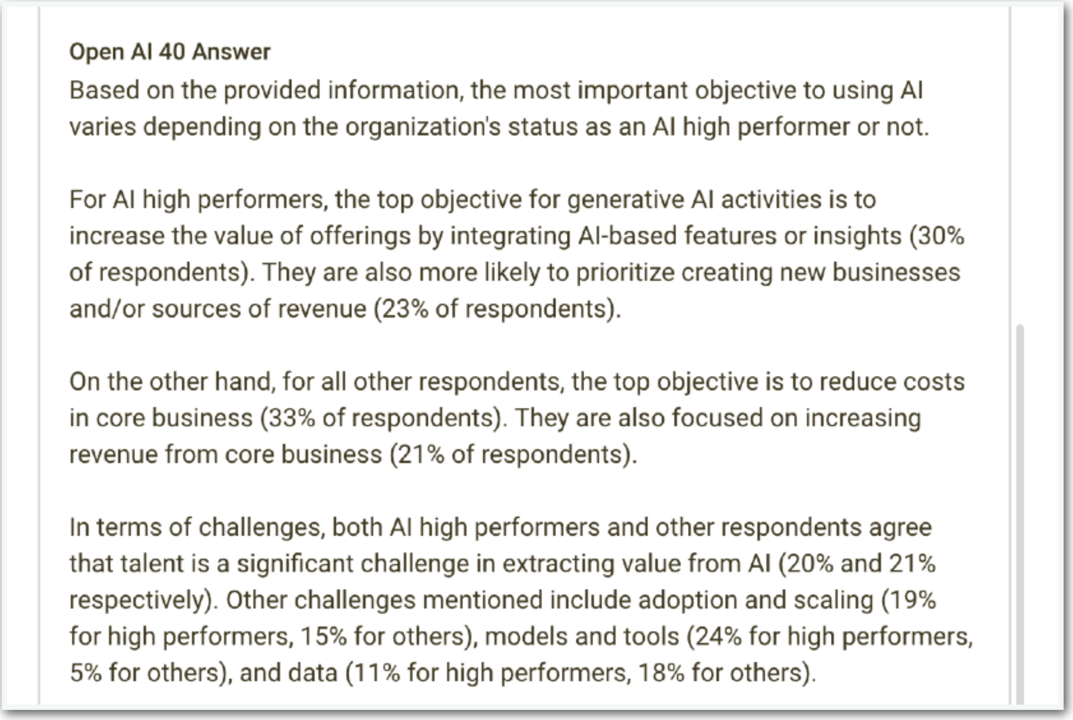
Krista prompt

# LLM Test Outputs

Our LLM test comprises several critical steps, each designed to assess the capability of the chosen language models. In the first step, we used our intelligent automation platform, Krista, to prompt each model with a series of ten questions. Here is a screenshot of the Krista interface showing one of these prompts.

The second step is to compare the responses from each LLM. These outputs, captured in their entirety here, provide a clear view of how each model interpreted and responded to the prompts. The responses varied in detail and context, reflecting the unique understanding and generation capabilities of each model.

## ChatGPT-4 Output



**Open AI 40 Answer**

Based on the provided information, the most important objective to using AI varies depending on the organization's status as an AI high performer or not.

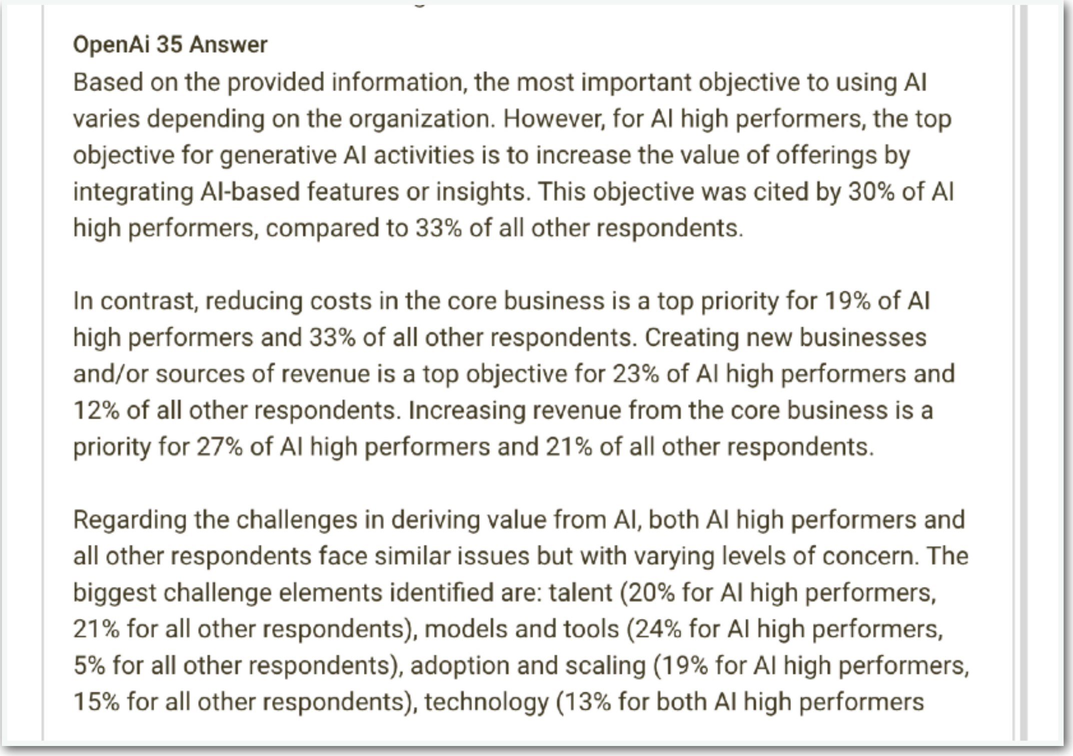
For AI high performers, the top objective for generative AI activities is to increase the value of offerings by integrating AI-based features or insights (30% of respondents). They are also more likely to prioritize creating new businesses and/or sources of revenue (23% of respondents).

On the other hand, for all other respondents, the top objective is to reduce costs in core business (33% of respondents). They are also focused on increasing revenue from core business (21% of respondents).

In terms of challenges, both AI high performers and other respondents agree that talent is a significant challenge in extracting value from AI (20% and 21% respectively). Other challenges mentioned include adoption and scaling (19% for high performers, 15% for others), models and tools (24% for high performers, 5% for others), and data (11% for high performers, 18% for others).

ChatGPT-4 Response

## ChatGPT-3.5 Output



**OpenAi 35 Answer**

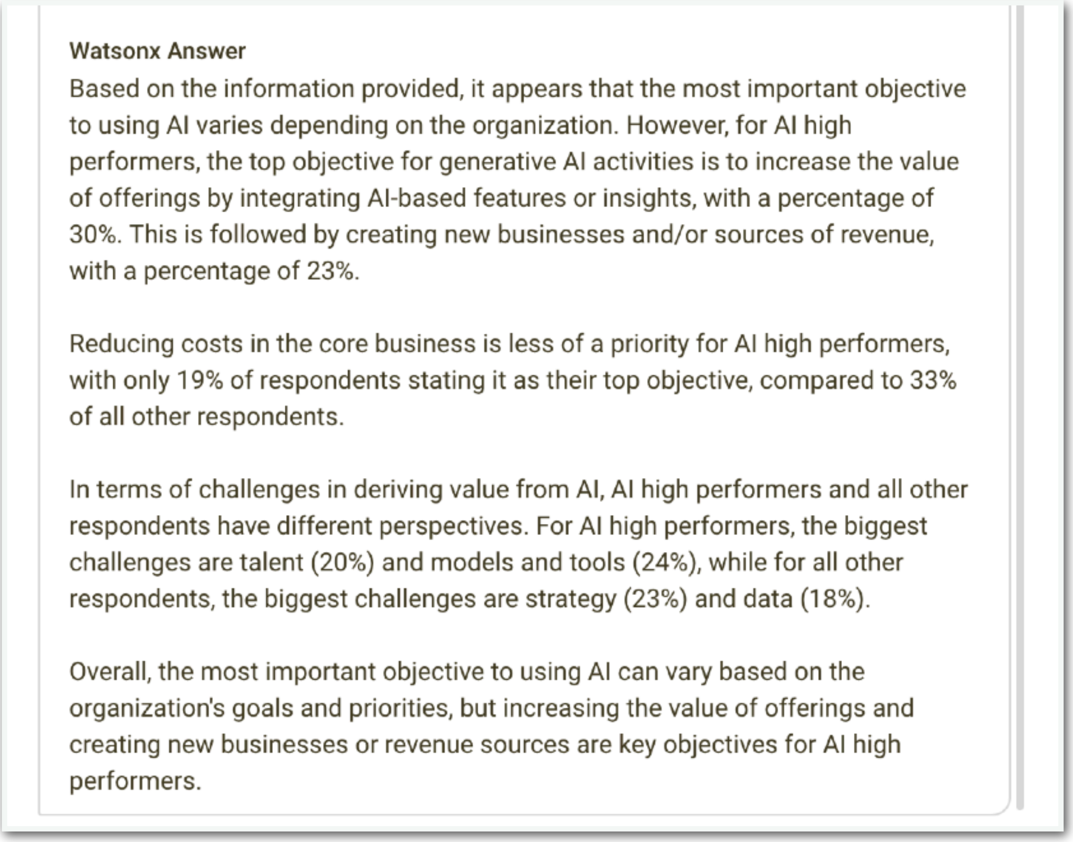
Based on the provided information, the most important objective to using AI varies depending on the organization. However, for AI high performers, the top objective for generative AI activities is to increase the value of offerings by integrating AI-based features or insights. This objective was cited by 30% of AI high performers, compared to 33% of all other respondents.

In contrast, reducing costs in the core business is a top priority for 19% of AI high performers and 33% of all other respondents. Creating new businesses and/or sources of revenue is a top objective for 23% of AI high performers and 12% of all other respondents. Increasing revenue from the core business is a priority for 27% of AI high performers and 21% of all other respondents.

Regarding the challenges in deriving value from AI, both AI high performers and all other respondents face similar issues but with varying levels of concern. The biggest challenge elements identified are: talent (20% for AI high performers, 21% for all other respondents), models and tools (24% for AI high performers, 5% for all other respondents), adoption and scaling (19% for AI high performers, 15% for all other respondents), technology (13% for both AI high performers

ChatGPT-3.5 Response

# Watsonx Output



**Watsonx Answer**

Based on the information provided, it appears that the most important objective to using AI varies depending on the organization. However, for AI high performers, the top objective for generative AI activities is to increase the value of offerings by integrating AI-based features or insights, with a percentage of 30%. This is followed by creating new businesses and/or sources of revenue, with a percentage of 23%.

Reducing costs in the core business is less of a priority for AI high performers, with only 19% of respondents stating it as their top objective, compared to 33% of all other respondents.

In terms of challenges in deriving value from AI, AI high performers and all other respondents have different perspectives. For AI high performers, the biggest challenges are talent (20%) and models and tools (24%), while for all other respondents, the biggest challenges are strategy (23%) and data (18%).

Overall, the most important objective to using AI can vary based on the organization's goals and priorities, but increasing the value of offerings and creating new businesses or revenue sources are key objectives for AI high performers.

watsonx Output

Finally, the third step involves the assessment and identification of the best output. Krista judges each output evaluating the quality, relevance, and completeness of each LLM's response. The reasons for selection are carefully explained, focusing not only on the accuracy of the answer but also on the context, detail, and clarity provided by each model. The winning output of each round is then used for further comparison, culminating in the selection of the overall best output.

To provide a full understanding of the process, all of the complete test responses and reasons an output was selected as the best are available in a Google Sheet linked in the footnotes.[8] This comprehensive document contains all of the output data and illuminates how each of the LLMs responded to the same prompts against the same source document and the reasons why each generated response was selected as the best.

# Summary of Test Results

After conducting our comprehensive analysis, we ran a ten-question test ten times for a total of one hundred opportunities for an LLM to claim victory. The results revealed that both ChatGPT-4 and watsonx secured a win in 42 instances each, demonstrating a similar level of competence and effectiveness. ChatGPT-3.5 lagged behind, having won 16 times (Figure 2).

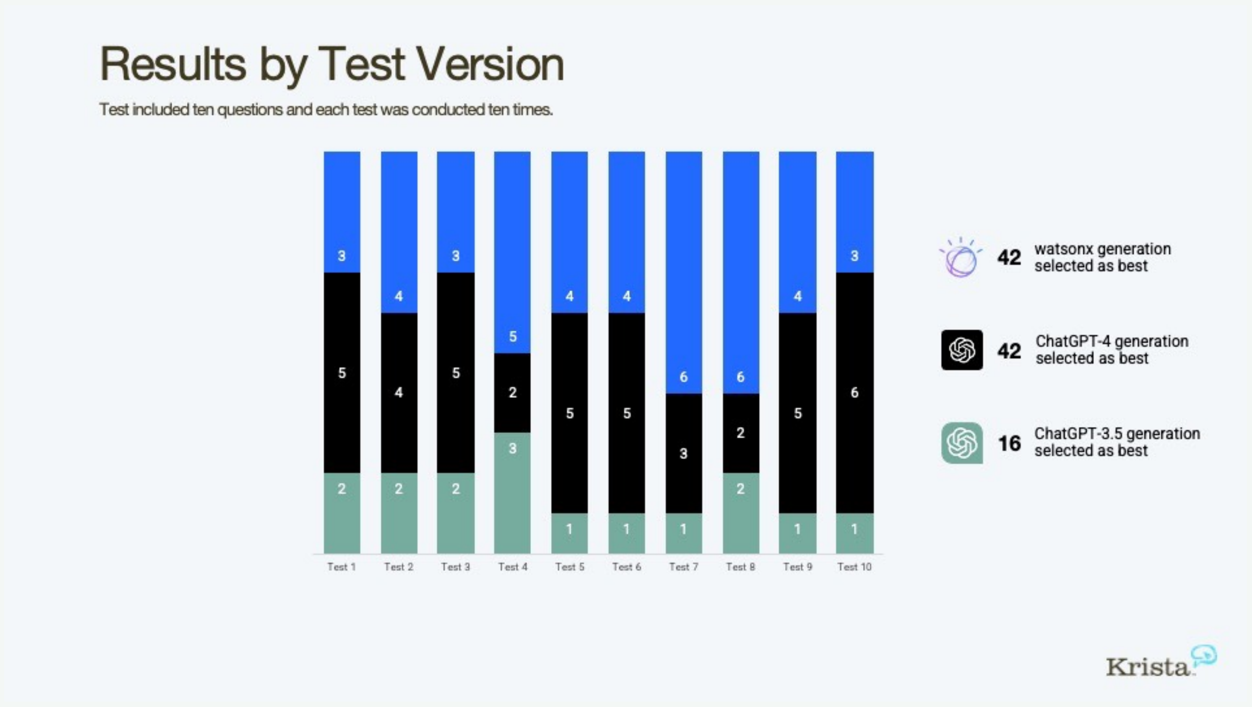
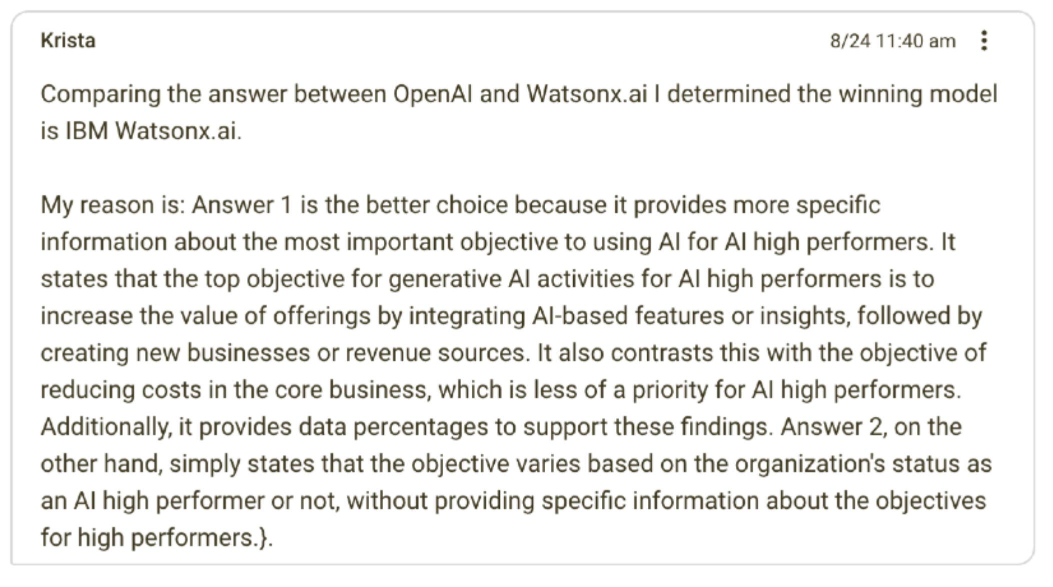


Figure 2

## Watsonx Results

Watsonx displayed superior performance in answering key questions that included chart data. These questions included "What are the main concerns of using AI?", "What is the most important objective to using AI?", and "How will workforces change after implementing AI?" (Figure 3) In these instances, watsonx demonstrated an edge in its ability to extrapolate and interpret data from the charts more effectively than both ChatGPT-3.5 and ChatGPT-4. This capability is highly important as it highlights watsonx's proficiency in understanding complex data structures and extracting accurate information. This competence in handling data-driven queries makes watsonx a strong contender for use cases that demand precise interpretation and communication of quantitative data, a feature that can critically influence decision-making processes in an enterprise setting.



Krista 8/24 11:40 am

Comparing the answer between OpenAI and Watsonx.ai I determined the winning model is IBM Watsonx.ai.

My reason is: Answer 1 is the better choice because it provides more specific information about the most important objective to using AI for AI high performers. It states that the top objective for generative AI activities for AI high performers is to increase the value of offerings by integrating AI-based features or insights, followed by creating new businesses or revenue sources. It also contrasts this with the objective of reducing costs in the core business, which is less of a priority for AI high performers. Additionally, it provides data percentages to support these findings. Answer 2, on the other hand, simply states that the objective varies based on the organization's status as an AI high performer or not, without providing specific information about the objectives for high performers.}.

Reason Why LLM Response is the Best Output, Answer 1 = watsonx.ai

## ChatGPT-4 Results

ChatGPT-4 demonstrated exceptional performance in responding to questions such as "What are the most common use cases?", "How many companies have adopted AI?", and "Which department is most likely to shrink?” (Figure 3) For the first question, ChatGPT-4 was selected as it provided a comprehensive understanding of the most common use cases of generative AI. Mentioning the increased investment in AI due to its promise and the fact that it has made its way to boardroom discussions, ChatGPT-4 provided valuable insights into the specific business areas where AI is proliferating, like marketing, product development, and service operations.

For the second question, ChatGPT-4 performed better than the other models by providing more extensive details about the AI adoption rate among companies.

Finally, in response to the question about which department is likely to shrink due to AI adoption, ChatGPT-4 again came out on top by offering specific and thorough information. It indicated that service operations is the department most likely to experience a decrease in workforce size due to generative AI's prevalence. Through these examples, ChatGPT-4 exhibited a remarkable ability to provide detailed, relevant, and contextually rich responses, making it an effective tool for enterprises seeking to leverage AI's potential.

## ChatGPT-3.5 Results

In the comparison of responses, ChatGPT-3.5 outperforms watsonx due to its ability to provide precise, data-driven, and comprehensive insights. It excels in its capacity to offer succinct and accurate summaries, capturing essential nuances that watsonx sometimes overlooks. For instance, ChatGPT-3.5 identifies service operations as the department most likely to experience downsizing, a vital detail neglected by watsonx in one of the test runs. It also provides robust evidence of AI's potential to boost organizational revenues, presenting specific growth rates, which watsonx fails to do. Furthermore, the superiority of ChatGPT-3.5 lies in its comprehensive explanations, utilizing specific data and comparisons between AI high performers and other organizations that are absent in watsonx's response. Overall, the strength of ChatGPT 3.5 lies in its data-driven precision, its attention to detail, and its ability to paint a complete and accurate picture, thus making it a viable choice to use in your organization.

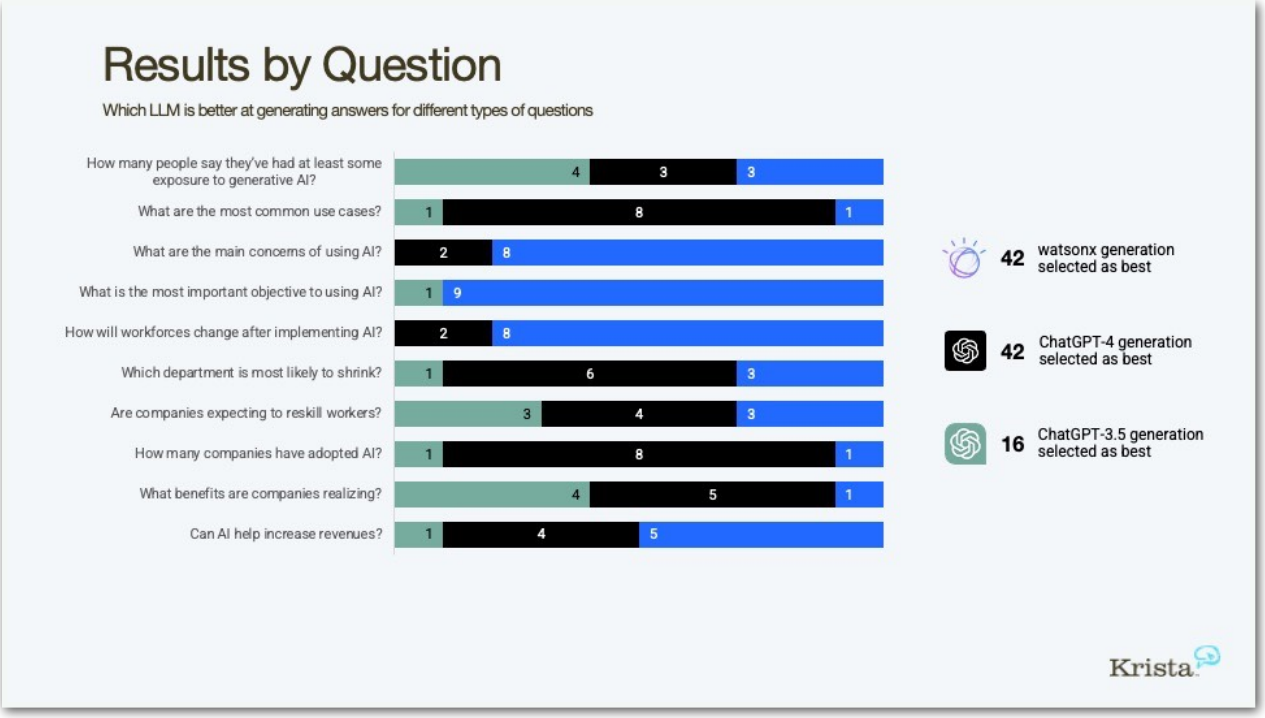


Figure 3

# LLM Innovation Requires Regular Testing

Generative AI is highly innovative and evolves at a rapid pace. Therefore, it is crucial to maintain a regular testing schedule for any LLM used in your processes. The LLMs can either improve or degrade due to constant updates and you can't assume updates will work in your favor. As such, you will need to conduct daily, weekly, or monthly regression tests of the LLM outputs to assess and evaluate model updates based on each of your specific use cases. To cost-effectively manage regular LLM testing, investing in a robust testing framework is essential. The testing framework will not only help you keep track of the evolution of your chosen LLM but also ensure you minimize technical debt when trying to build and design your own LLM tests.

# How to Conduct Your Own LLM Test

To conduct your own Large Language Model evaluation using Krista, the first step is to reach out to us at krista.ai and obtain access to Krista.[9] With Krista, you'll be able to interact with various licensed and open-source LLMs and observe their responses in real time, providing a practical and comprehensive comparison.

Once you have access to Krista, the next step is choosing a suitable dataset for your test. It can either be from your own company or a public document. The dataset serves as the basis for the questions that will be posed to the LLMs. Ideally, the dataset should cover a variety of topics and include different types of data to ensure a wide-ranging test of the models' abilities.

After you've selected your dataset, we at Krista will assist you with crafting the test. This typically involves determining the types of questions to ask and creating a robust test that covers a broad spectrum of capabilities. We aim to test not just the models' ability to generate correct answers, but also their comprehension of context, detail, and clarity.

Next, you will evaluate the responses. Based on the criteria mentioned, you will assess the quality, relevance, and completeness of each LLM's response. Just like the test we conducted, you'll compare the winning output of each round, ultimately selecting the overall best output.

Finally, using our LLM implementation guide, you'll be able to effectively implement the chosen LLMs. The guide provides detailed instructions on how to integrate the model and utilize it to your advantage. By following these steps, you can successfully scale AI in your organization, making informed decisions based on your thorough evaluation and data-backed insights.

# Summary

Our Large Language Model comparison of watsonx, ChatGPT-4, and ChatGPT-3.5 provides a thorough overview for testing any LLM even though we chose only three. Each model demonstrated unique strengths, with watsonx excelling in interpreting tabular data structures, ChatGPT-4 providing detailed, context-rich responses, and ChatGPT-3.5 offering precise, data-driven insights. The results of this comparative analysis highlight the potential for generative AI and LLMs to significantly benefit a wide range of enterprise use cases, from improving decision-making processes to boosting operational efficiency. Additionally, the document can serve as a guide for conducting continuous LLM tests on your own Krista as generative AI continues to advance.

# Footnotes

1. Revolutionizing eCommerce: How AI-driven Automation Transformed a North American Store's Customer Service Operations, Krista Software
2. How Krista Software helped Zimperium speed development and reduce costs with IBM Watson, IBM
3. Leading Healthcare Application Embeds GenAI and Conversational Automation to Revolutionize Employee Wellness Programs, Krista Software
4. OpenAI
5. The state of AI in 2023: Generative AI's breakout year, McKinsey & Company
6. Watsonx, IBM
7. The foundational model for this watsonx test is flan-ul2, Google
8. Krista AI LLM Test September 2023, Krista Software
9. Contact Krista Software

Krista™