

Watsonx.ai Assessment Framework Final Presentation

Aaliyah Y., Andres M., Jingruo C., Natalia J., Xiaoyu F., Yawen T., Zixiang M.



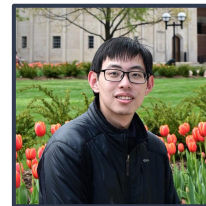
Meet The Team



Jingruo (Gina) Chen
Researcher



Natalia Jordan
MLOps Specialist



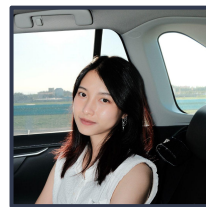
Zixiang Meng
Foundation Model Specialist



Yawen Tan
Researcher



Andres Murillo
SCRUM Leader



Aaliyah Yang
Foundation Model Specialist



Xiaoyu (Amber) Fan
Researcher

Executive Summary

We developed comprehensive frameworks to assess IBM's enterprise generative AI platform, focusing on identifying its strengths and weaknesses to pinpoint its competitive advantage. We focus on four critical use cases essential for enterprise AI: **Customer Service Chatbot**, **Business Intelligence**, **Code Assistance**, **HR Management**. We designed specific test cases for each use case, executed assessments, and collected data to evaluate the platform against competitors.

Our findings reveal that in **Customer Service Chatbot**, Watson X is robust, yet could benefit from simplified training processes. In **Business Intelligence**, it shows a solid foundation but needs expanded integration with IBM's Planning Analytics and enhanced features for data visualization and AI insights. For **Code Assistance**, while Watson X scores high in maintainability, it could improve in code style and QA through exposure to diverse problems and feedback during training. Lastly, in **HRM**, adjusting training to improve keyword precision, reduce bias, and enhance language will optimize chatbot efficacy. These recommendations aim to refine Watson X's functionality and extend its applicability across diverse user needs, laying a foundation for future enhancements and broader user engagement.

Case Studies and Product Testing Results

Summary of Results

1. **Customer Service Chatbot**: Watsonx performance well comprehensively, and we provide recommendation on easier training.
2. **Business Intelligence**: Watsonx.ai has a solid foundation, but should expand integration options, especially with IBM's Planning Analytics, enhance data visualization, improve transparency of AI-generated insights, and strengthen collaboration features. Introducing an AI insight feature like Zia Insights could aid comprehensibility. Moreover, develop a flexible AI chat that accommodates diverse input types.
3. **Code Assistant**: Watsonx perform better on maintainability, but scores lower on code style and QA performance. It can be improved by being exposed to a more diverse problem set, and getting code style feedback during the training process.
4. **HR**: Chatbots perform differently based on three metrics. Overall, Watson prompt lab can be trained to improve the keyword precision, reduce gender-encoded words and increase language attractiveness for a job description.

Customer Service Chatbot

Customer Service Chatbot

Test platforms

watsonx

 **yellow.ai**



INTERCOM



Amazon Lex

Overall Comparison

User-Friendliness

Intercom and IBM Watson are particularly user-friendly.

Training Efficiency

Amazon Lex and Yellow.AI have more complicated setups due to specific requirements and occasional technical issues, respectively.

Customization and Flexibility

IBM Watson and Yellow.AI offer significant customization options.

Documentation and Support

All platforms generally offer clear documentation.

Recommendations

Interface Simplification

Enhanced Integration Process

Clarification of Features

Improved Search Visibility

Customer Service Chatbot

watsonx

yellow.ai



INTERCOM



Amazon Lex

Recommendation

User Friendliness

4

3

5

3

Ease of Training

4

3

5

2

Customization & Flexibility

4

5

2

3

Clarity of Help & Documentation

5

5

5

5

Interface Simplification

Enhanced Integration Process

Clarification of Features

Improved Search Visibility

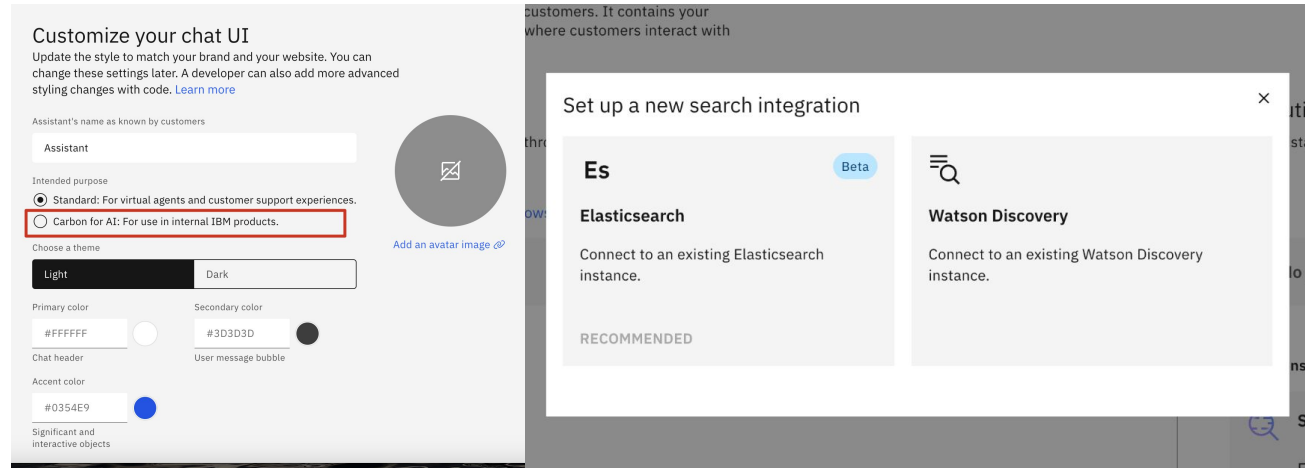
Customer Service Chatbot

User-friendliness

Clean and intuitive user interface,
easy for users to navigate
and manage their projects,
logical separations between
different features and
functionalities.

Recommendation

Examples



Interface Simplification: Further simplify the training interface to make it more accessible for beginners who are unfamiliar with terms like "carbon for AI".

Customer Service Chatbot

User-friendliness

Clean and intuitive user interface, easy for users to navigate and manage their projects, logical separations between different features and functionalities.

Recommendation

Interface Simplification: Further simplify the training interface to make it more accessible for beginners who are unfamiliar with terms like "carbon for AI".

Customize your chat UI

Update the style to match your brand and your website. You can change these settings later. A developer can also add more advanced styling changes with code. [Learn more](#)

Assistant's name as known by customers

Assistant

Intended purpose

- ☒ Standard: For virtual agents and customer support experiences.
- ☐ Carbon for AI: For use in internal IBM products.

Choose a theme

Light

Dark

Primary color

#FFFFFF

Secondary color

#3D3D3D

Chat header

User message bubble

Accent color

#0354E9

Significant and interactive objects



Add an avatar image [@](#)

Vague term: what is Carbon AI function?

Customer Service Chatbot

User-friendliness

Clean and intuitive user interface
easy for users to navigate
manage their projects
logical separations between
features and functions

Recommendations

Interface Simplification

simplify the training interface
make it more accessible

beginners who are unfamiliar with
terms like "carbon for AI"

customers. It contains your
where customers interact with

Set up a new search integration

Es

Beta

Elasticsearch

Connect to an existing Elasticsearch
instance.

RECOMMENDED



Watson Discovery

Connect to an existing Watson Discovery
instance.

Vague term: what is the difference between Elasticsearch
and Watson Discovery?

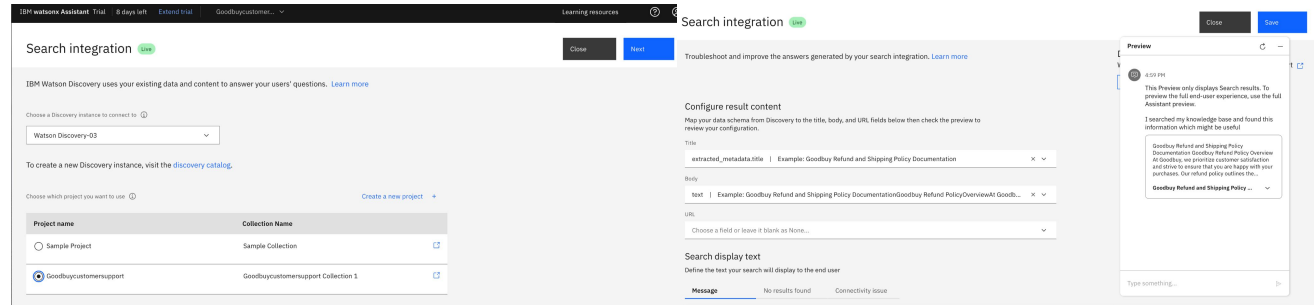
Customer Service Chatbot

Ease of Training

Evaluate how easily new intents, content, and updates can be managed within the platform

Recommendation

Examples



Enhanced Integration Process: **Streamline the integration** between Watson Discovery and Watson Assistant to eliminate the need to switch between platforms during bot creation.

Customer Service Chatbot

Environments

Live environment ⚙️

Use the live environment for deployment to customers. It contains your published content and channel integrations where customers interact with your assistant.

Channels

Your customers interact with your assistant through different communication platforms.

Channels [Browse](#)

Web chat

Set up a new search integration ✕

Es Beta

Elasticsearch

Connect to an existing Elasticsearch instance.

Watson Discovery

Connect to an existing Watson Discovery instance.

RECOMMENDED

Search

Extend what your assistant can answer by searching your existing documents

[Add](#)

Search function: last step in training & not visible to new users

during bot creation.

Customer Service Chatbot

IBM watsonx Assistant

Trial

8 days left

[Extend trial](#)

Goodbuycustomer... ▾

[Learning resources](#)

?

⌵

Search integration

Live

Close

Next

IBM Watson Discovery uses your existing data and content to answer your users' questions. [Learn more](#)

Choose a Discovery instance to connect to ⓘ

Watson Discovery-03

▾

To create a new Discovery instance, visit the [discovery catalog](#).

Choose which project you want to use ⓘ [Create a new project](#) +

Project name	Collection Name
<input type="radio"/> Sample Project	Sample Collection ↗
<input checked="" type="radio"/> Goodbuycustomersupport	Goodbuycustomersupport Collection 1 ↗

Watson Assistant to eliminate the need to switch between platforms during bot creation.

Search function: Need to create the IBM Discovery project first in another tab

Customer Service Chatbot

☒ Select project type

☐ Select data source

☐ Connect to data

☐ Configure collection

What type of project are you working on?


Project name

Untitled project 1

Project type

Conversational Search

Conversational Search



Supply answers to a virtual agent built with IBM Watson Assistant.

Watson Assistant to eliminate the need to switch between platforms during bot creation.

Search function: Need to create the IBM Discovery project first in another tab

Customer Service Chatbot

Search integration Live

Troubleshoot and improve the answers generated by your search integration. [Learn more](#)

Configure result content

Map your data schema from Discovery to the title, body, and URL fields below then check the preview to review your configuration.

Title

extracted_metadata.title | Example: Goodbuy Refund and Shipping Policy Documentation x v

Body

text | Example: Goodbuy Refund and Shipping Policy DocumentationGoodbuy Refund PolicyOverviewAt Goodb... x v

URL

Choose a field or leave it blank as None... v

Search display text

Define the text your search will display to the end user

Message

No results found

Connectivity issue

Preview

4:59 PM

This Preview only displays Search results. To preview the full end-user experience, use the full Assistant preview.

I searched my knowledge base and found this information which might be useful

Goodbuy Refund and Shipping Policy Documentation Goodbuy Refund Policy Overview

At Goodbuy, we prioritize customer satisfaction and strive to ensure that you are happy with your purchases. Our refund policy outlines the...

Goodbuy Refund and Shipping Policy ... v

Type something... v

need to switch between platforms during bot creation.

Search function: Returns the whole doc instead of answers generated based on the documents

Customer Service Chatbot

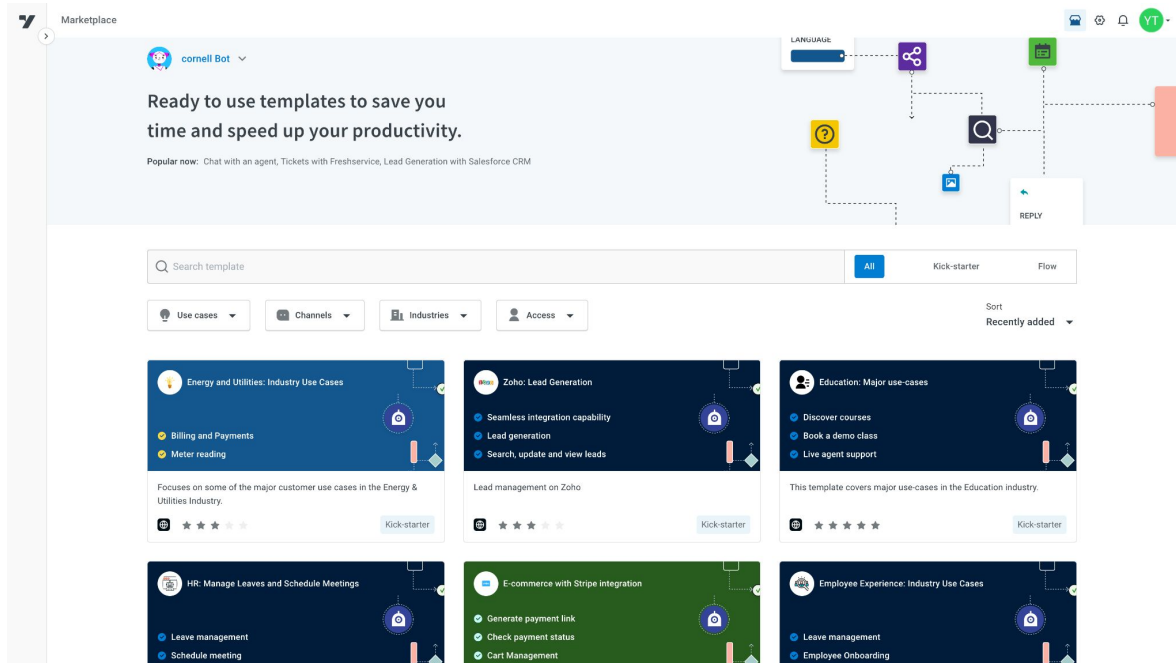
Customization and Flexibility

Evaluate the platform's capabilities in terms of customization to meet specific needs or adapt to new requirements.

Recommendation

Template Marketplace Access:
Develop a Template Marketplace on IBM Watson to simplify the chatbot creation process.

Example

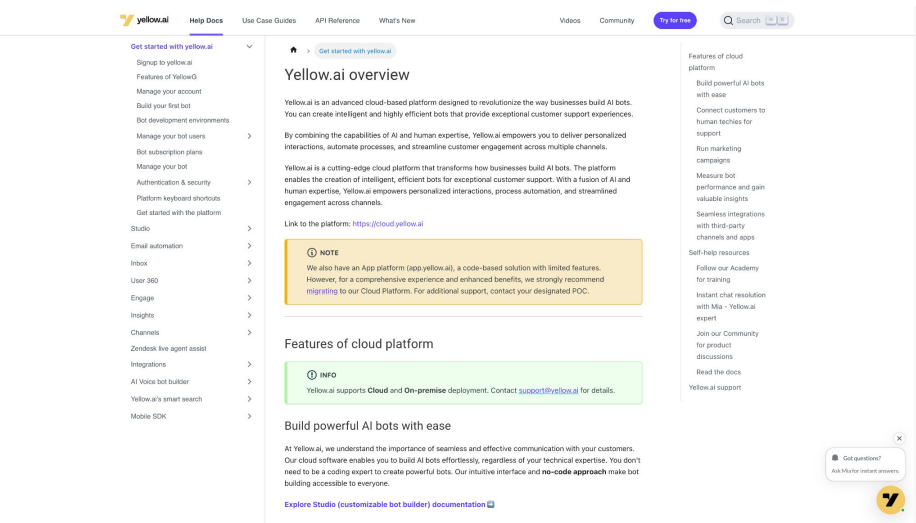
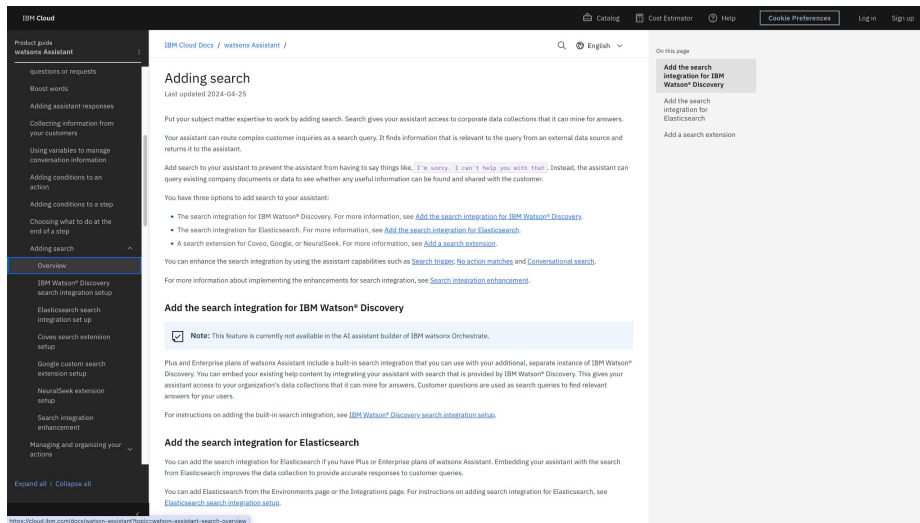


Customer Service Chatbot

Clarity of Help and Documentation

Provide details on the availability, clarity, and comprehensiveness of the documentation and help resources provided by the platform.

Examples



Business Intelligence

Product Testing Results

Assessment Metrics	watsonx	 Zoho Analytics		
Connectivity & Integration Capabilities	4	5	3	3
Visualization, Interactivity, & Customization	4	5	2	4
AI-Generated Insights & Transparency	3	4	2	5
User Experience, Comprehension & Collaboration	4	5	2	4

Recommendations

Seamless Data Integration and Connectivity

Powerful and Interactive Data Visualization

Transparent AI-Generated Insights and forecasting

Intuitive User Experience and Robust Collaboration

Data Connectivity and Integration Capabilities

Use Case Description

Evaluate the platform's ability to connect to various data sources. Additionally, consider maximum volume of data per project and accepted file types

Significance

1. **Comprehensive Data Access**
2. **Real Time Insights**
3. **Data Quality and Consistency**
4. **Scalability and Flexibility**

Overall: Saves time while conducting analysis

Evaluation & Recommendation

Watsonx Evaluation:

Efficiently handles large data volumes (10TB) and formats. Extensive pre-built connectors and low-code / no-code options for integrating diverse data sources. Takes 40+ third party connection options.

Recommendations:

Despite strong range of connectivity options, Watson can further expand its integration capabilities to match Zoho (60+) and Salesforce (3000+)

Product Ranking



Zoho Analytics

watsonx



Data Connectivity **Watsonx.ai** Integration Capabilities

Connect to a data source

[Supported connectors](#)

Create a new connection or connect to a service

New connection

Deployed services

Provider

- ☐ IBM
- ☐ Third-party

Compatible services

- ☐ Catalogs
- ☐ Data Quality Rules
- ☐ Data Replication
- ☐ DataStage
- ☐ Metadata Enrichment
- ☐ Metadata import
- ☐ Watson Query
- ☐ Watson Studio

Find connectors

All connectors

- | | | | |
|---------------------------|-----------------------|--------------------------------|--------------------|
| Amazon RDS for MySQL | Google BigQuery | IBM Db2 on Cloud | OData |
| Amazon RDS for Oracle | Google Cloud Pub/Sub | IBM Db2 Warehouse | ODBC |
| Amazon RDS for PostgreSQL | Google Cloud Storage | IBM Informix | Oracle |
| Amazon Redshift | Greenplum | IBM Match 360 | Oracle (optimized) |
| Amazon S3 | HTTP | IBM MQ | PostgreSQL |
| Apache Cassandra | IBM Cloud Data Engine | IBM Netezza Performance Server | Presto |

Cancel

Back

Select

Overall: Saves time while conducting analysis



Category

All

Most Popular

Recently Added

Files & Feeds

Databases

Zoho Apps

Sales/CRM

Marketing

Finance

Human Resources

IT & Help Desk

ERP

Project Management

E-Commerce

Social Media

Survey/Forms

Custom Solutions

All



Files



Feeds/URLs



Cloud Storage/Drive



Cloud Databases



Local Databases



Azure Data Lake



Zoho Analytics Workspace



Elastic Search



Zoho DataPrep



Zoho CRM



Zoho SalesIQ



Google Analytics 4 (GA4)



Bigin



Microsoft Dynamics CRM



Teamwork CRM



Google Universal Analytics (UA)



Google Analytics 4 (GA4)



Google Ads



Facebook Ads



Facebook Pages



Recommendation

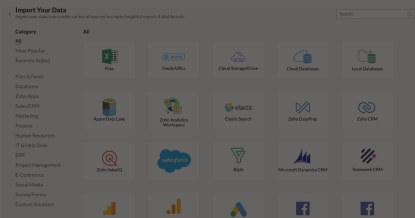
Seamless Connectivity:

Offer extensive pre-built connectors and low-code/no-code options for integrating diverse data sources. Efficiently handle large data volumes and formats.

[SalesForce AppExchange →](#)

← Zoho Analytics

Examples



Jitterbit | Your Solution for Salesforce I...

by Jitterbit, Inc.
★★★★★ 4.88 (68)

Jitterbit Harmony is a leading Enterprise iPaaS that helps accelerate innovation by combining the power of APIs and integration. Integrate Salesforce with any SaaS, on-prem or cloud app, and easily compose innovative...

Integration



DocuSign eSignature for Salesforce: Th...

by DocuSign, Inc.
★★★★★ 4.55 (4639)

Send, sign, track and save agreements in Salesforce with the most downloaded electronic signature app on the AppExchange. Get started with a free 30-day trial by clicking "Get It Now"

Customer Service



Metazoa Spotlight - Org Health and Se...

by Metazoa
★★★★★ 4.54 (13)

Intelligent Assistant is your administrative thinking partner. Engineer prompts and share them with your team. Work with metadata assets in your org. Best practices for using Generative AI language models...

IT & Administration Admin & Developer Tools



DocuVault-Document Management f...

by BIGWORKS
★★★★★ 5 (43)

DocuVault is a secure file storage and document management App that lets users upload and access large files from within Salesforce. Files are stored in Amazon S3 (cloud storage platform) and can be accessed securely...

Document Management



Vonage for Service Cloud Voice and ...

by Vonage
★★★★★ 4.92 (975)

Deliver exceptional agent and customer experiences with a fully integrated Salesforce and contact center solution that seamlessly unifies voice, AI (conversational, virtual assistant, voicebot, chatbot), digital channels, & CRM...

Customer Service



QuickBooks Integration with Salesforce...

by BREADWINNER INTEGRATIONS INC.
★★★★★ 4.99 (77)

Breadwinner for QuickBooks offers a powerful two-way integration solution that easily connects Salesforce and QuickBooks Online. Gain access to live critical data by using Breadwinner's intuitive and flexible integration.

Productivity Integration



Triggr

by Triggr



Softsquare - Lightning DataTable Dev

by Softsquare

★★★★★ 4.37 (49)



FinancialForce Accounting & Financia...

by Certinia

★★★★★ 4.6 (150)

Data Visualization, Interactivity & Connectivity

Use Case Description

Assess the variety and quality of data visualization options (charts, graphs, dashboards, etc.). Explore interactive features for drilling down, filtering, and slicing data, fine-tune insights to business requirements.

Significance

1. Enhanced Understanding
2. Improved Decision-Making
3. Faster Insights
4. Customization and Personalization

Overall: Reduce time spent on visuals and opportunity to see previously hidden trends in data

Evaluation & Recommendation

Watsonx Evaluation:

Provides a wide array of aesthetic visualizations to choose from. 30+ customizable options including AI-based recommendations depending on the dataset / variables.

Recommendations:

To enhance the data visualization and interactivity experience, WatsonX.ai could introduce more interactive features like:

- Drilling down
- Filtering
- Slicing data

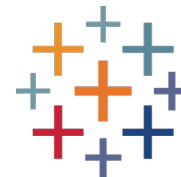
Additionally, providing greater flexibility in selecting variables for axes could address the user feedback received during testing.

Product Ranking



Zoho Analytics

watsonx



Data Visualization, Interactivity & Connectivity

Use Case Description

Watsonx.ai

Assess the variety and quality of data visualization options (charts, graphs, dashboards, etc.). Explore interactive

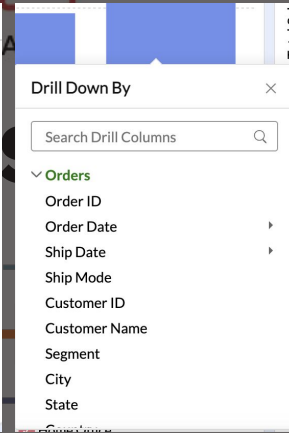
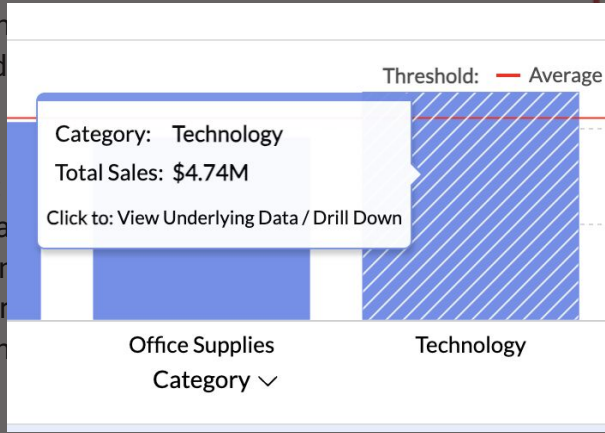
Evaluation & Recommendation

Watsonx Evaluation:

Provides a wide array of aesthetic visualizations to choose from. 30+



Zoho Analytics



Overall: Reduce time spent on visuals and opportunity to see previously hidden trends in data

Additionally, providing greater flexibility in selecting variables for axes could address the user feedback received during testing.



Data Visualization, Interactivi

Desired Visualization ->

Use Case Description

Assess the variety and quality of data visualization options (charts, graphs, dashboards, etc.). Explore interactive features for drilling down, filtering, and slicing data, fine-tune insights to

Evaluation & Recommendation

Watsonx Evaluation:

Provides a wide array of aesthetic visualizations to choose from. 30+ customizable options including AI-based recommendations depending on the

Comparison of yearly sales from 2012-2015



CHART TYPE

Scatter plot



Line



Multi-series



Histogram



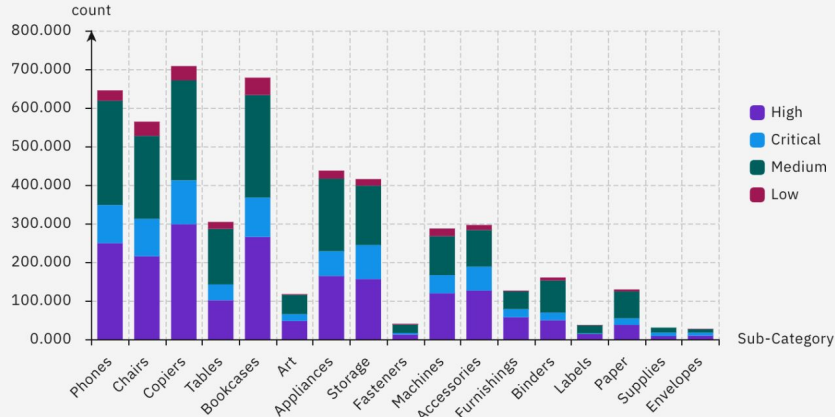
Population ...



Q-Q plot



Pie



Data

Category* ⓘ

- ☒ Sub-Category
- ☐ Order ID
- ☐ Ship Mode
- ☐ Customer ID
- ☐ Customer Name

- ☐ Count unique
- ☐ Sum
- ☐ Minimum
- ☐ Maximum
- ☐ Mean

watsonx



<- Watsonx Axis Options



AI-Generated Insights and Forecast Transparency

Use Case Description

Evaluate the accuracy of generated insights in comparison to human-generated insights (done by Vikas Kumar Appani on data.world) and the platforms transparency and explainability of the AI-generated insights.

Significance

1. **Trust and Credibility**
2. **Identify Biases and Errors**
3. **Compliance and Accountability**

Overall: Boost forecasting capabilities while also understanding model behavior and assumptions

Evaluation & Recommendation

Watson Evaluation:

AI capabilities provide recommendations for how to visualize data. Overall Watson scored well on the accuracy test however could improve with transparency and performance metrics.

Recommendations:

Transparent AI-Generated Insights:

Generate accurate, explainable AI insights with clear visualizations and forecasting. Provide transparency through model evaluation, performance metrics, and visibility into underlying data/algorithms.

Product Ranking



Zoho Analytics



watsonx



AI-Generated Insights and Forecast Transparency

Description

Evaluate the accuracy of generated insights in comparison to human-generated insights and the platforms transparency and explainability of the AI-generated insights.

Recommendation

Transparent AI-Generated Insights
Generate accurate, explainable insights with clear visualizations and recommendations. Provide transparency through evaluation, performance metrics, and visibility into underlying data.

Einstein Prediction



Predicted Outcome



494.46

Actual Outcome



2892.51

Residual



-2398.05

Percent



-82.91%

Top factors

↓ -374.63

Quantity is 5 and Sub-Category is Phones

↓ -145.50

Sub-Category is Phones and Profit is -96.54

↑ 143.54

Quantity is 5 and Category is Technology

Examples

Einstein Prediction



Predicted Outcome



494.46

Actual Outcome



2892.51

Residual



-2398.05

Percent



-82.91%

Top factors

↓ -374.63

Quantity is 5 and Sub-Category is Phones

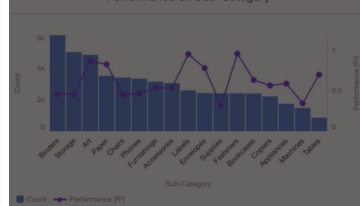
↓ -145.50

Sub-Category is Phones and Profit is -96.54

↑ 143.54

Quantity is 5 and Category is Technology

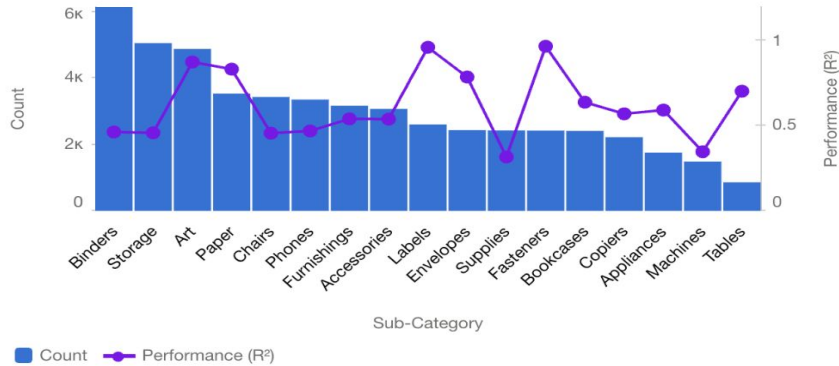
Performance of: Sub-Category



Use the graph to compare values by performance and row count. Performance is usually similar across values and outcomes. Any disparities can be due to relative differences in row count. For example, if a value is underperforming the others, it can be due to a lower row count.

SalesForce Einstein

Performance of: Sub-Category



Use the graph to compare values by performance and row count. Performance is usually similar across values and outcomes. Any dissimilarities can be due to relative differences in row count. For example, if a value is underperforming the others, it can be due to a lower row count.

Done

← Sub-Category

Alert Performance Settings

See how well the model works for each value in Sub-Category.

Value	Performance ⓘ
Binders	0.46
Storage	0.46
Art	0.88
Paper	0.83
Chairs	0.46
Phones	0.47
Furnishings	0.54
Accessories	0.54
Labels	0.96
Envelopes	0.79
Supplies	0.32
Fasteners	0.97
Bookcases	0.64
Copiers	0.57
Appliances	0.59
Machines	0.35
Tables	0.70

Row Count Analysis ⓘ

DATA ALERT

Outliers

Multicollinearity

High Correlation

Multicollinearity

Multicollinearity

Multicollinearity

Multicollinearity

High Cardinality

Multicollinearity

AI-Generated Insights and Forecast Transparency



Transparent AI-Generated insights:

Zia AI generated ↑

Provide transparency through model evaluation, performance metrics, and visibility into underlying data/algorithms.



Human generated ↓



User Experience, Comprehension, & Collaboration

Use Case Description

Provide details on the availability, clarity, and comprehensiveness of the documentation and help resources provided by the platform. Evaluate collaboration capabilities and comprehensibility of insights. Note this was evaluated using user testing.

Significance

1. User Adoption
2. Increased Productivity
3. Accessibility

Overall: A strong user experience and comprehension allows for a streamlined work experience

Evaluation & Recommendation

Watson Evaluation:

Solid overall layout with a "board" for visualizations, datasets, and variable summarizations. Ability to export insights as CSV or PDF.

Recommendations:

Limited collaboration features, with the ability to add WatsonX.ai users as collaborators but no external sharing or team communication options.

Prioritize user-friendliness, customization options, and collaboration features like insight sharing and team communication. Offer comprehensive support resources, like webinars and personalized demos

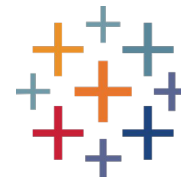
Product Ranking



Zoho Analytics




watsonx



User Experience, Comprehension, & Collaboration


Add users as collaborators

Users



Start typing in the search field to get a list of users

Selected users



No users selected
After you select users from the list they will appear here.

Cancel

Add

Intui
Colla

Prior
options, and collaboration features like insight
sharing and team communication. Offer
comprehensive support resources, like webinars
and personalized demos



4

Copy the URL below to access this Slideshow

<https://show.zoho.com/show/present/VwkFOQjg1MTg0MzgxODoxNzE0MjQ1Mjc2NTQ3>

Copy

360° insights on your company sales

Deals Leads Salesperson P

Deals Overview

Filters

Timeline Filter

Last 12 Mon

Deals Conversion Apr 2



Deals Closed in Apr 202

7 ↓

Mar 2023: 158

Avg Deal Size Won

\$73.68K

Avg Sales Cycle in Days

21.12

Deals Summary

Enter email addresses or group names

Pick Users / Groups



lea.torres



sarah.c



Marketing



Sales

Apply Permissions & Filters - Read Only, Allow Commenting

☒ Send Invitation Mail - [Edit Message](#)

Existing Shared Details

Share

Cancel

Search Contacts

> Shared Contacts (2)

> Zoho Contacts (171)

> Google Contacts (0)

> Office 365 Contacts (0)

> Groups (2)

> Organization Contacts (13)



lea.torres

lea.torres@zylker.com



sarah.c

sarah.c@zylker.com



john.mj

john.m@zylker.com



Jason Henderson

jason.h@zylker.com



oliverm

oliverm@zylker.com



Share Reports and Dashboards

Copy the URL below to access this Slideshow

<https://show.zoho.com/show/present/VwKFOOjg1MTg0MzgxODoxNzE0MjQ1MjcNTQ3>

Copy

Provid
clarity
docum
provi
colla
comp

Intui
Colla

Prior
optio
shari

comprehensive support resources, like webinars
and personalized demos

with Zoho Show

Slideshow Preview

[Edit with Zoho Show](#)

Slideshows 1

- Created By Natalia Jordan

[Fullscreen](#)

Copy the URL below to access this Slideshow

<https://show.zoho.com/show/present/WkFOOjg1MTg0MzgxoDoxNzE0MjQ1MjcNTQ3>

[Copy](#)

Provide de
clarity, and
documenta
provided by
collaborati
comprehen

Intuitive U
Collaborat

Prioritize u
options, an
sharing an

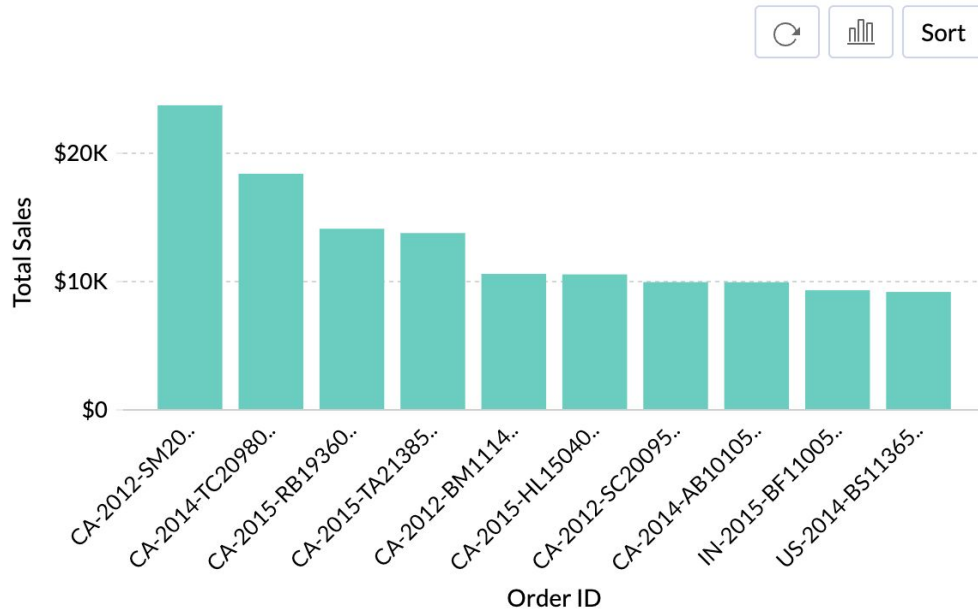
comprehensive support resources, like webinars
and personalized demos

Copy the URL below to access this Slideshow

<https://show.zoho.com/show/present/WkFOOjg1MTg0MzgxoDoxNzE0MjQ1MjcNTQ3>

[Copy](#)

Top 10 Order ID by Sales



Code Assistance

Recap of Midpoint Presentation

Key Competitors

- Watsonx (Granite 13b)
- ChatGPT
- Claude3
- Code Whisper

Significance

- Streamlines Coding Process
- Reduce Manual Coding Efforts
- Consist Coding Standards
- Efficient and Accessible Coding

Evaluation Matrix

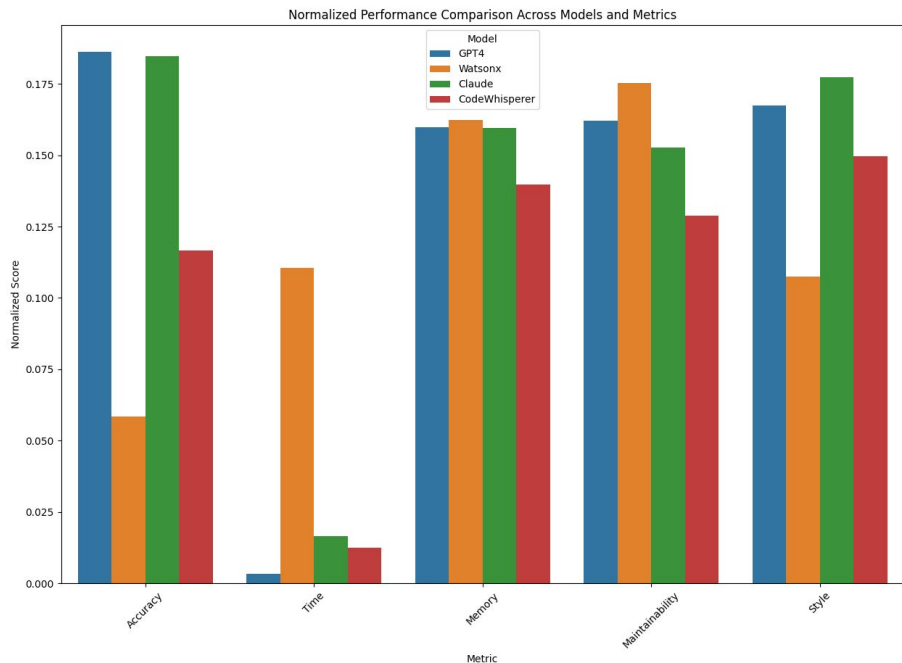
- Performance
 - Correctness
 - Execution time
 - Memory usage
- Maintainability
 - Lines of Code
 - Halstead Volume
 - Cyclomatic Complexity
- Code Style

Test cases

- 10 different Leetcode samples that cover different algorithms and struction with median and difficult levels. [\(link here\)](#)

Code Assistance

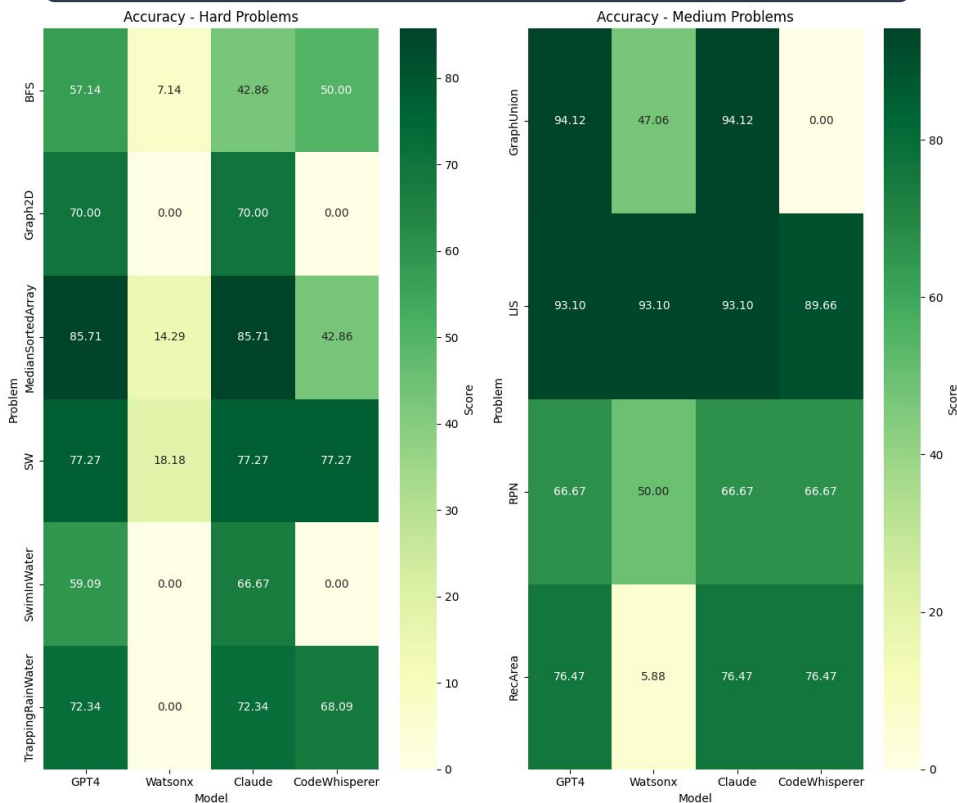
Performance



- **Accuracy:**
 - Claude and GPT show similar levels of accuracy.
 - Watsons lags behind the others.
- **Time:**
 - Watsons shows a significantly higher run time, indicating slower performance on tasks.
 - The other three models are clustered closely together with much lower time required, suggesting faster task execution.
- **Memory:**
 - Relatively balanced between the models.
 - CodeWhisperer has a slight advantage.
- **Maintainability:**
 - Watsonx has the edge over the other models.
 - CodeWhisperer has lower maintainability scores.
- **Style:**
 - Claude leads in style.
 - Watsons far behind, indicating potential areas for improvement in code style or readability.

Code Assistance

Data



Rating: 1-5



watsonx



Correctness

4

2

4

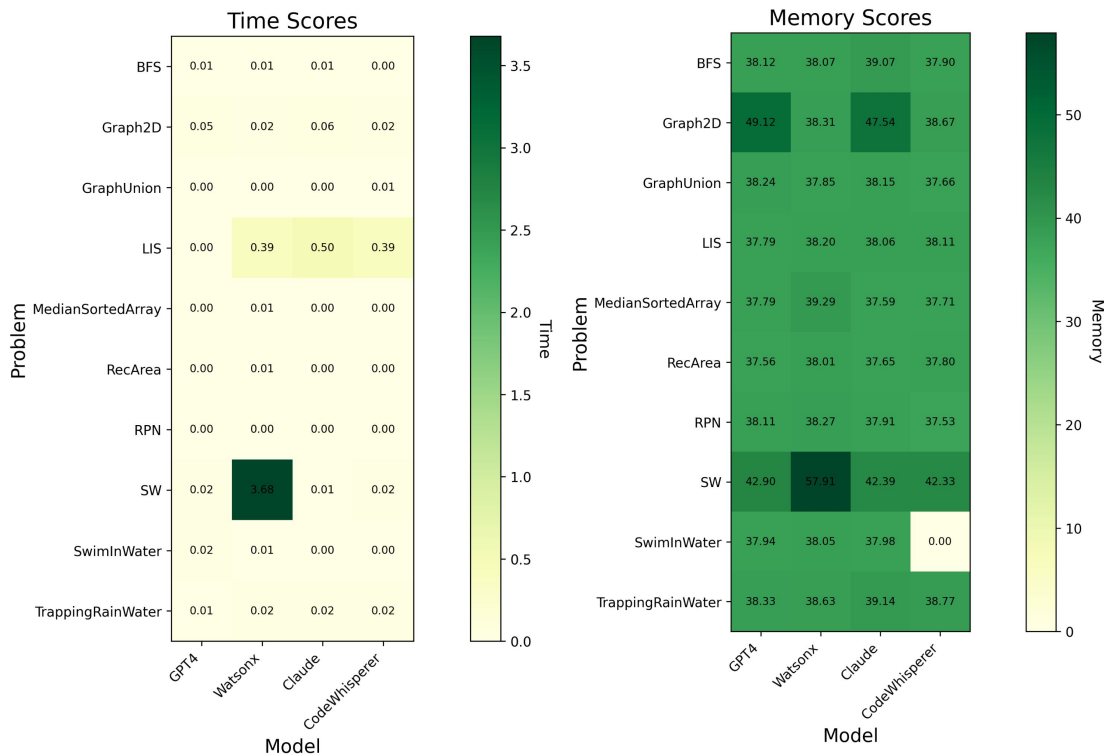
3

Insights

- Accuracy:** Claude and GPT-4 generally perform well in accuracy, with Claude excelling in GraphUnion problem and GPT-4 maintaining consistent scores across problems.
- Problem Complexity Handling:** Watsonx seems to struggle with more complex problems, especially those involving advanced data structures or algorithms.

Code Assistance

Data



Rating: 1-5



watsonx



Time

5

4

5

5

Memory

5

5

5

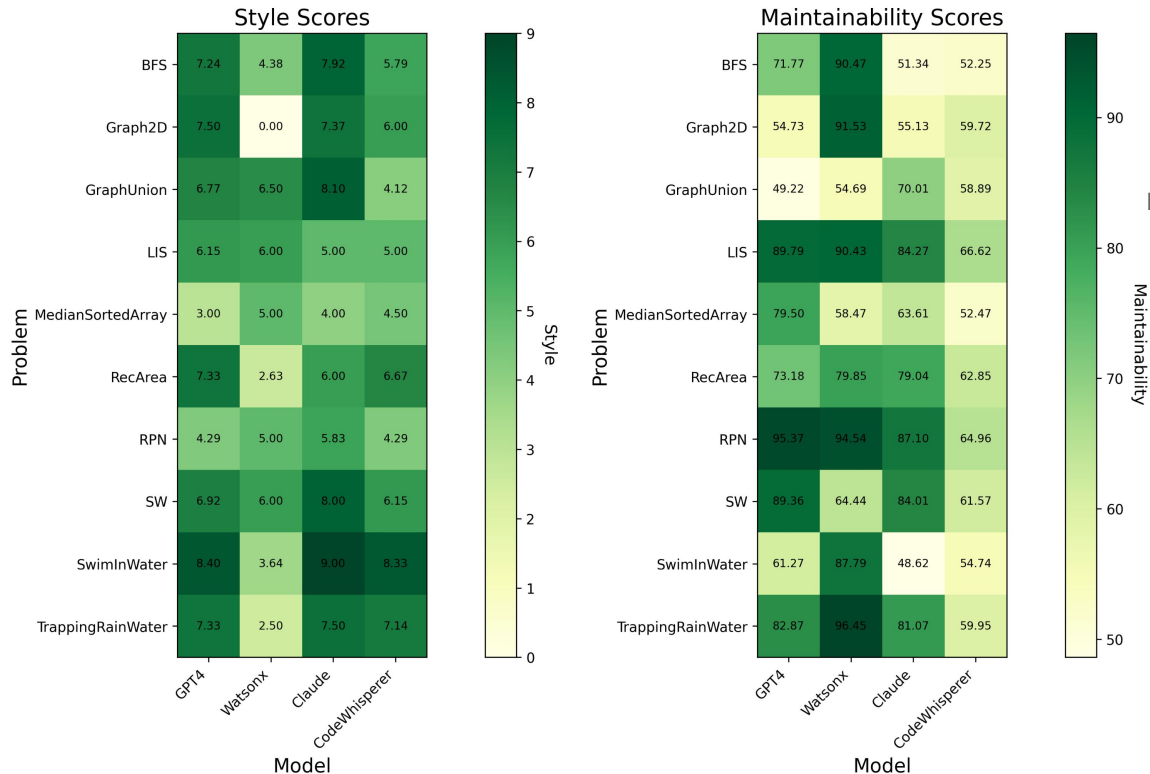
4

Insights

- Time:** Watsonx takes notably longer to solve the SW problem compared to the other models, which demonstrate slower execution times across the board.
- Memory:** Watsonx's memory usage spikes for the SW problem (hardest one), while other models are more consistent across different problems.

Code Assistance

Data



Rating: 1-5



watsonx






Style	4	2	4	3
Maintainability	3	4	3	2

Insights

- Style:** Watsonx's style scores are relatively low compared to other models, indicating that the generated code may lack some stylistic considerations.
- Maintainability:** Watsonx scores higher than other models for most of the time. This suggests that the code generated by Watsonx is generally more readable and easier to maintain.

Suggestions

	Correctness	Time	Memory	Code Style	Maintainability	Overall
watsonx	2	4	5	2	4	17
 Amazon CodeWhisperer	3	5	4	2	2	16
 Claude ANTHROPIC	4	5	5	3	3	20
 OpenAI ChatGPT	4	5	5	4	3	21

- **Expand Training Data and Problem Coverage:**
 - Watsons may have difficulty with more complex problems, according to the lower scores for certain **problem-metric combinations**.
 - Expose Watsonx to a more **diverse set of coding problems**, covering a wider range of algorithms, data structures, and problem domains.
- **Enhance Code Generation Quality:**
 - Watsonx's style scores are relatively low compared to other models, may lack some **stylistic considerations**.
 - **Incorporate coding style guidelines** and best practices into the training process to improve the stylistic quality of the generated code.
 - **Leverage code quality** analysis tools and metrics to provide feedback during the training process and refine the generated code.
- **Interpretability and Explainability:**
 - While not directly visible in the visualizations, improving Watsons' ability to provide explanations or rationales for its generated solutions could be valuable.
 - Enhancing interpretability and explainability could not only improve trust in the model but also facilitate debugging and further development.

Chatbot as HRM Assistant

Chatbot as HRM Assistant

Description

Chatbot as HRM assistant automate employee recruiting and selecting by developing a precise job description, saving time compared to manually writing a job description strategy.

Watsonx Prompt Lab main IBM product for testing.

Test platforms

watsonx



Overall Comparison

Relevance of the Position/Job Description

All Chatbots include most keywords in the job description with ChatGPT being the most flexible and Bing being the most precise.

Inclusivity & Bias

All Chatbots perform well (≤ 4 words misuse) with Gemini performing better than the other three.

Engagement & Appeal

Gemini and Bing perform well due to proper length, clear structure, vivid language and emojis. Chat GPT has bigger score variation due to short length.

Chatbot as HRM Assistant

Rank Products in order with 4 (rank 1st), 3 (2nd), 2 (3rd), 1 (4th) points

watsonx



Bing

Gemini

Relevance of the Description

1

3

4

2

Inclusivity & Bias

2

3

2

4

Engagement & Appeal

2

2

3

4

Recommendations

Precise keywords at the assigned position

Data training focused on promoting gender-neutral language


Clearer structure and more descriptive language

Chatbot as HRM Assistant

Relevance of the Description

Evaluate if the job description accurately reflects the role responsibilities and the prompt language requirements.

Relevance Points

watsonx  Bing 

1	3	4	2
---	---	---	---

Data Analysis

watsonx  Bing 

	watsonx	OpenAI ChatGPT	Bing	Gemini
Required Keywords	14/16	16/16	16/16	15/16
Follow the Order	Yes	No	Yes	No
Additional Section	Yes	Yes	No	No

Recommendation

Generated job description include more precise keywords specified in the prompt language.

Try to include the keywords at the assigned position.

Chatbot as HRM Assistant

Inclusivity & Bias

Check for any biased language or exclusionary terms that might discourage certain demographic groups from applying.

Data Analysis

Number of words that can be improved

watsonx



Bing

Gemini

4

3

4

1

Recommendation

Data training includes examples of gender-neutral language and guidelines on how to avoid gender bias in job descriptions.

Inclusivity Points

watsonx



Bing

Gemini

2

3

2

4

Examples

2

Gender-coded words

Words/phrases that have a masculine or feminine connotation

Chatbot as HRM Assistant

Engagement & Appeal

Assess whether the job description is engaging and appeals to the target audience.

watsonx  Bing 

2	2	3	4
---	---	---	---

Data Analysis

Rate the attractiveness of the company overview from 1 to 5

	Watson	ChatGPT	Gemini	Bing
Scores from surveys	3, 2, 2, 3, 3, 3	4, 3, 1, 4, 2, 2	4, 4, 5, 5, 5, 5	5, 4, 4, 4, 4, 4

Examples

“I give Chat GPT 1 score because the company overview part is too short”

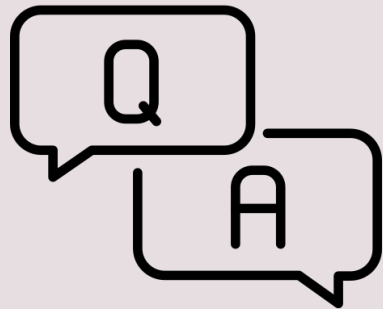
“I think the company overview generated by Gemini is vivid and clear”

Recommendation

Train the prompt lab to incorporate more vivid and descriptive language to make the company overview more engaging and compelling

consider structuring it using clear bullet points

THANK YOU



Questions & Feedback

APPENDIX

Recap of Midpoint Presentation

Description

Problem: Customer service can't answer fast enough when too many customers need help at once. Also, sometimes give wrong answers, or don't understand what's needed.

Solution: AI-powered assistants that can handle a range of customer service tasks, such as answering FAQs, guiding users through processes, managing simple transactions, and escalating complex issues to human agents.

We will use watsonx Assistant as main IBM product for testing.

Benefits

Manage high volumes of customer inquiries

Reduce labor costs & 24/7 Availability

Consistency in Responses

Integration with Other Services

Comparable Free Trials



Amazon Lex



yellow.ai

INTERCOM

Evaluation Metrics

First Contact Resolution

Average Handling Time

Ease of training

Recap of Midpoint Presentation

Evaluation Metrics

First Contact Resolution (FCR): capability of understanding customer issues and providing solutions without additional follow-ups.

Average Handling Time (AHT): Shorter handling times -> assistant is prompt in providing solutions -> higher customer satisfaction.

Ease of training: Ability to update and manage intents and content without significant effort -> adapt to new products, services, or changes in customer behavior.

Justification

[How to Improve First Call Resolution With Conversational AI | Cognigy](#)

[Chat Report Metrics explained | yellow.ai](#)

Test Methods

Start: Train the chatbots using one set of documentations

Input: Generate customer inquiry situations

Output: Conversation with the customer to resolve the inquiry

How to evaluate:

1. FCR: Calculate the percentage of issues resolved in the first contact.
2. AHT: Calculate the average handling time of each interaction (total time of one conversation/n of interactions)
3. Rate the ease of use of the training interface, the clarity of documentation, and any challenges faced.

Recap of Midpoint Presentation

Description

Chatbot as HRM assistant automate employee recruiting and selecting by developing a precise job description, saving time compared to manually writing a job description strategy. **We will use Watson Prompt Lab as the main IBM product for testing.**

Benefits

Improved Efficiency

Consistency in Language
Formatting

Optimization for Keywords

Reduce Bias

Comparable Free Trials



OpenAI
ChatGPT

Gemini



Evaluation Metrics

Relevance of Position

Inclusivity & Bias

Engagement & Appeal

Recap of Midpoint Presentation

Evaluation Metrics

Relevance of the Job Description: Evaluate if the job description accurately reflects the role responsibilities and the prompt language requirements.

Inclusivity & Bias: Check for any biased language or exclusionary terms that might discourage certain demographic groups from applying.

Engagement & Appeal: Assess whether the job description is engaging and appeals to the target audience.

Justification

[Some common Chatbot Evaluations metrics](#)

[Some common issues in current job descriptions](#)

Chatbot facilitate job description generation by [reducing bias](#) and [improve efficiency](#)

Test Methods

Input: Some Key requirements of the job description

Output: The detailed job description with corresponding sections

How to evaluate:

1. Calculate the number of appearing key words like skills and experience
2. Use [Inclusive language Checker](#) to test Inclusivity
3. Surveys to ask about the attractiveness and company culture in job description

Research Methodology

User Experience

Goal:

Ensure intuitive, efficient user interactions

Why:

Direct impact on adoption and productivity

Example:

Task completion time, user satisfaction ratings

[Designing Trustworthy AI: A User Experience \(UX\) Framework at RSA Conference 2020, Carol Smith, Carnegie Mellon University, 10 Usability Heuristics for User Interface Design](#)

Academic

Goal:

Align with cutting-edge research and methodologies

Why:

Foundation for robust and advanced solutions

Example:

Model accuracy benchmarks
[Stanford 2023 AI index](#)

Institution

Goal:

Meet organizational and industry-specific standards

Why:

Ensures scalability and adherence to policy

Example:

Compliance with regulatory requirements
[Organisation for Economic Co-operation and Development's AI principles](#)

Assessment Framework

1	Accuracy	<p>The degree to which an AI system's outputs or decisions are correct based on the given input data.</p> <p>Evidence: According to the Trust and Artificial Intelligence guidelines from the U.S. Department of Commerce and National Institute of Standards and Technology, accuracy is ranked first among their research in terms of GenAI's trustworthiness features.</p>
2	Transparency and explainability	<p>The extent to which an AI system's actions or decisions can be understood by humans.</p> <p>Evidence: According to the OECD's AI principles, AI Actors should commit to transparency and responsible disclosure regarding AI systems.</p>
3	Robustness/ Security/ Safety	<p>Evaluate the chatbot's ability to correctly understand the user's intent and provide relevant responses.</p> <p>Evidence: According to the OECD's AI principles, AI systems should be robust, secure, and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose an unreasonable safety risk.</p>
4	Accountability	<p>Analyze the average number of steps or messages it takes for the chatbot to resolve a query. Fewer steps might indicate a more efficient process.</p> <p>Evidence: According to the OECD's AI principles, AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art.</p>

Assessment Framework

5	Privacy	<p>Evaluate how easily the chatbot can be customized or trained to suit specific needs or adapt to new requirements.</p> <p>Evidence: According to the paper titled <i>Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation</i> by Bernd Carsten Stahl and David Wright, the European General Data Protection Regulation (GDPR) (https://gdpr-info.eu/) explicitly addresses the impact of smart information systems. Among the novel features relevant to smart information systems are breach notifications, hefty financial penalties, data protection impact assessments, privacy by design, and the so-called right to be forgotten.</p>
6	Reliability	<p>For global applications, assess the number of languages supported by the chatbot and its ability to maintain context across languages.</p> <p>Evidence: According to the paper titled <i>In AI We Trust: Ethics, Artificial Intelligence, and Reliability</i>, one can only feel disappointed by AI, because this refers to functional expectations that are not met and, as such, is the appropriate reaction to reliability issues. Reliability is only one factor used to determine whether to trust an agent.</p>
7	Objectivity (Factual Knowledge)	<p>Measure how often users return to use the chatbot, indicating its long-term value to users.</p> <p>Evidence: The Amazon foundational model overview evaluates how well the model encodes factual knowledge. FMEval can measure the model against a custom dataset or use a built-in dataset based on the TREX open-source dataset.</p>
8	Fairness/Inclusivity	<p>Measure how effective and efficient the chatbot integrates with other tools or platforms such as websites, apps, or social media channels.</p> <p>Evidence: According to the paper titled <i>From Reality to World. A Critical Perspective on AI Fairness</i> (https://link.springer.com/article/10.1007/s10551-022-05055-8), Algorithm biases, discrimination, and consequently unfairness have been identified in various AI applications, such as predictive models in justice, facial recognition, search engines, advertising, speech recognition, AI for recruitment, and predictive models in healthcare.</p>

Assessment Framework

9	Risk Governance	<p>Management of the potential risks associated with the development, deployment, and ongoing use of foundational models</p> <p>Evidence: The Japan AI regulations (https://www.mofa.go.jp/files/100573473.pdf 2023G7 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems): Governance - Policies and processes for risk assessment, oversight and control throughout the AI lifecycle.</p>
10	Regulation Enforcement	<p>The establishment and application of laws, guidelines, standards, and practices designed to ensure that these technologies are developed and deployed in a safe, ethical, responsible, and transparent manner.</p> <p>Evidence: new Chinese AI regulations (https://time.com/6314790/china-ai-regulation-us/): Regulation Enforcement - Analyzing how strictly oversight agencies like the CAC enforce the rules in practice. Flexible enforcement balances control and growth</p>
11	Trust & Ethical Considerations (Toxicity)	<p>The degree to which the system is free from harmful or unethical outputs and garners user trust.</p> <p>Evidence: “Privacy and surveillance, bias and discrimination, and perhaps the deepest, most difficult philosophical question of the era, the role of human judgment”, said Sandel, - https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/</p>
12	System Efficiency	<p>The performance and resource utilization of these models during training, inference, and deployment.</p> <p>Evidence: Efficient systems maximize the return on investment, handle large volumes of data, provide timely insights, improve user satisfaction, outperform competitors, and allocate computational resources effectively, ultimately driving business growth and success.</p>

Assessment Framework

13	Adaptability of Information/Flexibility	<p>The need for adaptability arises because different users and software contexts require customized approaches for optimal performance and user satisfaction.</p> <p>Evidence: HCI Research for E-Learning: Adaptability and adaptivity to Support Better UserInteraction(https://www.academia.edu/29938663/HCI_Research_for_E_Learning_Adaptability_and_Adaptivity_to_Support_Better_User_Interaction)</p>
14	Intuitive and Innovative Design	<p>Intuitive, user-friendly interface with novel and creative outputs.</p> <p>Evidence: Visuals are essential to establishing good first impressions. (https://www.nngroup.com/articles/aesthetic-minimalist-design/)</p>
15	Customization (User Control)	<p>Users should have control over the generated responses, including options to choose between multiple outputs, modify responses based on preferences, and provide feedback to improve future results. This control enhances the user experience by empowering users to tailor the output to their needs.</p> <p>Evidence: By enabling users to input and modify parameters, the platform can align with unique requirements, provide personalized solutions, and improve outcomes, while empowering users with a sense of control and ownership.</p>
16	Help and Documentation	<p>Definition: Help and documentation for foundational models are critical components that provide users, developers, and researchers with the necessary information to understand, use, and contribute to these models effectively.</p> <p>Evidence: AI technologies can be complex and require specialized knowledge to understand and operate effectively. Clear and comprehensive documentation helps users navigate the platform, understand its capabilities, and leverage its features to their full potential.</p>

Assessment Framework

17	Response effectiveness	Measure the average time taken by the chatbot to respond to user queries. Faster response times generally contribute to a better user experience.
18	User Satisfaction	Employ surveys or feedback tools to gauge user satisfaction. Net Promoter Score (NPS), Customer Satisfaction Score (CSAT), and User Experience (UX) ratings are valuable here.
19	Intent Recognition/ Understanding	Evaluate the chatbot's ability to correctly understand the user's intent and provide relevant responses.
20	Conversation Steps	Analyze the average number of steps or messages it takes for the chatbot to resolve a query. Fewer steps might indicate a more efficient process.

21	Customization and Flexibility	Evaluate how easily the chatbot can be customized or trained to suit specific needs or adapt to new requirements.
22	Language and Multilingual Support	For global applications, assess the number of languages supported by the chatbot and its ability to maintain context across languages.
23	Retention Rate	Measure how often users return to use the chatbot, indicating its long-term value to users.
24	Integration	Measure how effective and efficient the chatbot integrates with other tools or platforms such as websites, apps, or social media channels.

Assessment Framework

25	Scalability	Does the platform allow for scaling the chatbot, adding more complex functionalities over time?
26	Documentation and Support	Are there comprehensive guides, tutorials, and community support to help students learn and troubleshoot?
27	Consistency and Standards	Does the system use the same words/language to describe the same thing?
28	Conversation Steps	User Control and Freedom.

29	Error Prevention	Does the system proactively mitigate issues by either eliminating error-prone conditions or implementing checks to detect them, presenting users with a confirmation option before they proceed with the action?
30	Recognition Rather than Recall	For global applications, assess the number of languages supported by the chatbot and its ability to maintain context across languages.
31	Aesthetic and Minimalist Design	Interfaces should streamline content and visuals to essentials, avoiding extraneous details that can obscure crucial information and ensuring every element supports the user's primary goals.
32	Recognize, Diagnose, and Recover from Errors	Error messages should be clear, jargon-free, and visually prominent, offering straightforward descriptions and solutions for issues.

Customer Service Chatbot

✓ gmail Bot's knowledge

Use your website content and documents as a knowledge source to train the bot.

1 web domain added

0 documents added



Website URL

actum.cx



Save

Documents

+ Upload

FILE NAME

TAGS

No documents uploaded yet. Upload documents.

✓ Control gmail Bot's behavior

Edit your bot's responses during it's interaction with your users

Configured



🔧 Test gmail Bot's Responses

Try out different types of questions gmail Bot can answer

0 reports generated



✓ Live Agent Configuration

Provide your user the last mile support by directing queries to your agents.



Live agent enabled



🔔 Publish & deploy

Take your bot live with the changes you've made

Yet to publish



Customer Service Chatbot

Overview

[Add content](#) **No content** [Set up and go live](#) [Not live](#) [Optimize](#)



Phil from the Customer Education team



To set up Fin, start by adding content 📌

Fin will use AI to automatically generate answers with different kinds of content. Once content is imported, you can test Fin before setting it live.



Import external content

Import content from public URLs, like knowledge bases or websites



Enter the URL of your external support content. We will automatically import all of the pages from the website URL you provide. Provide your external help center homepage link for best results.

[How to add external content to Fin](#)

Add



Use your Intercom Help Center

Let Fin learn from the support content in your Help Center



Use Snippets

Create plain text content specific for Fin, not publicly available



Use Snippets

Create plain text content specific for Fin, not publicly available



Import content from files

Upload PDF files and we'll fetch all the text data inside



- Images are not scraped.
- File size limit is 45 MB.
- Files with multiple text columns, encrypted or password protected PDFs are not supported.



[Learn more about PDF content for Fin](#)

Add



Use Custom Answers

Let Fin learn from your bespoke answers and actions



Use Inbox conversation content

Let Fin learn from conversations handled by your CS reps in the Inbox



Customer Service Chatbot

Welcome to Kommunicate, Jingruo Chen!

Get started with Kommunicate in just 4 steps!

✓ Customize your chat widget

✓ Add a chatbot to automate your customer conversations

Create a bot with Kommunicate's GUI bot builder:



Kompose: GenAI
Powered Bot Builder



OpenAI Powered Bot
Builder



Already have a bot? Integrate with Kommunicate:



Dialogflow ES
[Documentation](#)



Dialogflow CX
[Documentation](#)



Amazon Lex
[Documentation](#)



IBM Watson
[Documentation](#)



Custom bot
[Documentation](#)



✓ Install Kommunicate's chat widget on your website or app

✓ Integrate chatbot with 40+ Channels



Cornell Test

English



Kompose bot builder

Flow Designer

Classic

Search



Welcome Message

Set an initial message for users



Intents

Setup training phrases and bot responses



Small Talk

Setup answers for generic replies



Documents

Train your Bot to answer from uploaded documents



Website Scraper

Use GenAI to train on webpages



Default Fallback

Fallback when bot fails to answer



Documents

We use OpenAI's GPT model to search relevant answer from the uploaded documents.

- Supported file types are .pdf, .docx, and .txt, up to 30MB each. Also, to upload larger volumes of data. [Talk to us](#)
- Password-protected files and files with multiple columns won't be scrapped. [Show more](#)

Click or drag to drop to upload a file

Uploaded Files:

No Documents are uploaded yet

Voice-Enabled Customer Support Framework

Evaluation Metric	Justification	Test Methods
Voice Recognition Accuracy	A high voice recognition accuracy leads to less human employee intervention, freeing workers for more complex tasks. In turn, it would also be a faster process for the customer. Importance of Accurate Speech Recognition in AI Transcription	We would use publicly available speech data that comes with transcripts (groundtruth) as the main input. Once done we would use the metrics such as the following to measure accuracy: Word Accuracy Rate, Word Information Preservation Rate, Word Error Rate, Hallucination Rate, Substitution Rate Assembly.ai
Call Resolution Time	Having a lower call resolution time could lead to higher customer satisfaction as they would have to spend less time on the line. It would also free up call lines and lessen wait times for those on hold. Top AI Speech Applications Nvidia	Develop a written scenario that we could input to watson detailing specific problems and measuring how long it would take to reach the intended solution.
First Call Resolution	One of the most important call center metrics. Measures a call center's performance for resolving customer interactions on the first call or contact, eliminating the need for follow-up contacts. Reduces operating costs, Reduces customers at risk of defection, Improves customer satisfaction. First Call Resolution SQM	Use same written scenario to measure the accuracy and time effectiveness of resolving a customer's issue within the first call (in our case a query). Using data from our other analysis (call resolution time), we can determine resolution time that fall below or above the median. FCR: Total First Call Resolutions / Total Unique Calls