^netomi

Conversational AI Benchmarking and Performance Report

An analysis of intent recognition, accuracy and coverage of leading conversational AI platforms: Google Dialogflow, IBM Watson, Microsoft LUIS, Netomi and RASA.

Executive Summary

AI has arrived, with adoption accelerating tenfold as a result of the COVID-19 Pandemic. Moving past "buzzword" and "shiny new object" status, AI is now successfully being deployed to improve various business processes and automating knowledge work, ultimately creating more efficiency and reducing costs. One of the most prevalent deployments of AI is with customer support, an area where businesses have failed to keep up with evolving customer expectations.

Consumers now expect instantaneous, effortless support 24/7. Even one poor interaction is enough to deter a person from ever doing business with a company again, let alone have any loyalty towards it. Historically, companies have staffed up support teams, but relying on a human-only workforce is no longer sustainable or attainable to scale quickly. The labor shortage has provided an enormous blow to the support industry; an industry which already had a staggering turnover of 45%+.

That's why companies are deploying AI-powered virtual agents across chat, SMS, messaging and email to automatically resolve highly repeatable customer service issues immediately. These AI agents offer many benefits: lower resolution time and costs, and improved customer and agent experience.

Coupled with the contagion from COVID-19, organizations are stymied by globalization in an increasingly competitive global market where they need to adapt quickly and rethink customer experience (CX) for many different stakeholders - employees, suppliers, and all the different people that interact with your organizations. The need to increase operational excellence 360 degrees is paramount now more than ever.

When deploying conversational AI agents for customer support, it's critical that companies realize how the AI will perform against customer expectations as poor experiences could actually cause more friction and harm. Customer experience is more critical today than ever before, and therefore ensuring AI agents handle every interaction properly directly impacts a company's bottom line and ability to compete in an increasingly turbulent environment.

In this natural language understanding (NLU) benchmarking report, we measure how the most prominent conversational AI platforms perform across a few key metrics: Coverage, Accuracy, Out of Scope Accuracy and Balanced Accuracy. The report reveals a lot of disparity in AI performance even amongst the most established platforms, signaling that even as the market matures, the end user experience is still greatly differentiated based on the underlying AI platform a company deploys.

Our report reveals that based on accuracy, out-of-scope accuracy and balanced accuracy, Netomi outperforms all other platforms in the correctness of responses sent to users. When it is not confident in its ability to understand a user's intent, the Netomi AI takes the best course of action and escalates a user to a human agent to minimize user frustration instead of offering an incorrect or irrelevant response. In a nod to the company's omnichannel focus, Netomi is also more accurate on both chat and email.

^netomi

Table of Contents

Methodology

We conducted natural language understanding (NLU) testing to analyze Netomi's AI performance and effectiveness against Google Dialogflow, IBM Watson, Microsoft LUIS and RASA. The 14 datasets in this study represented real-world scenarios and considered multiple combinations of divergent factors. Each test set had distinctive properties around entities and volume of training samples. We assessed the ability of an AI to correctly answer use cases for businesses of all sizes (small to enterprise). The datasets spanned core consumer-facing industries: retail, travel, gaming and telecommunications. Each dataset included common utterances of core ort questions that would be highly likely scenarios for an AI to come across if deployed to automate customer service queries in that industry.

Metrics Definition

The confusion matrix is based on the number of actual and predicted intent classes.

		Predicted Class	
		Positive (In-Domain)	**Negative** (Out-of-Scope/Unclassified)
Actual Class	**Positive** (In-Domain)	True Positive (TP) (Correct intent predicted for trained intents)	False Positive (FP) (Incorrect intent predicted for trained intents)
	Negative (Out-of-Scope/ Unclassified)	False Negative (FN) (Incorrect intent predicted for untrained intents)	True Negative (TN) (No intent predicted for untrained intents)

Table 1: Confusion Matrix

The following metrics were used to evaluate model performance.

Metric (Internal)	Statistical Term(s)	Description	Formula
Accuracy	Precision or Positive Predictive Value (PPV)	Ratio of correct positive predictions out of all positive predictions made, i.e., higher accuracy means more predictions from AI are accepted by user, thus causing less user frustration from wrong replies	TP/ (TP + FP)
Coverage	Recall or True Positive Rate or Sensitivity (TPR)	Ratio of correct positive predictions out of all actual positives, i.e., higher coverage results in more value delivered by AI, preventing human handoff for trained intents	TP/ (TP + FN)
Out-of-Scope (OOS Accuracy)	True Negative Rate or Specificity or Selectivity (TNR)	Ratio of correct negative predictions out of all actual negatives, i.e., higher out-of-scope accuracy leads to reduced user frustration since the bot avoids giving wrong replies where it is not trained	TN/ (FP+TN)
Balanced Accuracy	Balanced Accuracy	Average of coverage (TPR) and out-of-scope accuracy (TNR), i.e., higher Balanced Accuracy leads to more value delivered and less user frustration as the bot is providing more correct and relevant responses	(TPR + TNR)/2

1. Netomi has the highest accuracy compared to other AI platforms

In conversational AI, accuracy (or precision or positive predictive value) relates to the total number of correct replies out of all replies for the trained topics. Higher accuracy means that the AI is responding accurately, therefore causing less user frustration than if the bot provided an incorrect or irrelevant response.

This metric also acknowledges that the AI recognizes what it is not trained on, and therefore can take the correct course of action instead of providing an irrelevant response. When a bot answers incorrectly, user frustration grows. If a bot is not trained on a particular topic, instead of receiving an incorrect answer, a user would prefer to be directly handed off to a human agent in order to get their question or issue resolved.

In the Conversational AI Benchmarking Report, Netomi has the highest accuracy (85.17%), followed by IBM Watson (73.20%), Google Dialogflow (71.16%), RASA (68.56%) and Microsoft LUIS (61.79%). When we analyze performance per dataset, Netomi outperforms Google Dialogflow and Microsoft LUIS on all benchmark datasets, and IBM Watson and RASA on 13 out of 14 datasets.

Netomi has the highest accuracy in retail, gaming, travel, and telecom

Netomi's AI has the highest positive predictions out of all positives made across every industry included in this study. In retail, we tested 2,618 queries. Within this set, Netomi's AI is 87.08% accurate, compared to Google Dialogflow (75.88%), IBM Watson (75.18%), RASA (66.97%) and Microsoft LUIS (63.23%).

For the 4,263 common travel queries we tested, Netomi is 81.38% accurate, which is up to 26.33% points (pp) higher than other AIs in the study. IBM Watson performs at 66.49% accuracy, RASA at 67.66% accuracy, Google Dialogflow at 64.24% and Microdoft LUIS at 55.04%.

For the 1,185 gaming queries in the study, Netomi is 86.13% accurate, outperforming other platforms by up to 46.64% points (pp). IBM Watson is 77.61% accurate, Google Dialogflow 71.12% accurate, Microsoft LUIS 45.45% accurate and RASA 39.49% accurate.

In the telecom datasets, which included 1,194 queries, Netomi is accurate 87.74% of the time, followed by RASA (78.54%), IBM Watson (78.50%), Google Dialogflow (74.79%) and Microsoft LUIS (72.89%). This shows that Netomi outperforms other leading platforms in telecom by up to 14.85% points (pp).

Netomi is the most accurate AI on email and chat

Conversational AI performance often differs by channel as the ability to accurately decipher a user's intent is more difficult on certain channels. Email messages are typically harder for AI to parse as they tend to be longer messages with multiple intents, while chat messages are typically shorter and more succinct where it is easier for a machine to recognize an intent. While most companies have started leveraging AI for chat, increasingly companies are starting to leverage AI to automate customer service resolutions with email as well.

In the Conversational AI Benchmarking Report, Netomi has the highest accuracy across chat (21.16% points (pp) higher than Microsoft LUIS) and email (25.58% points (pp) higher than Microsoft LUIS).

Netomi has less than a 2% point difference in accuracy on email and chat. This compares to IBM Watson which has a 7% point difference in performance between these channels. This underscores the flexibility of the Netomi platform to successfully scale across channels and still deliver a highly accurate performance.

Metric	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Accuracy	71.16%	73.20%	61.79%	85.17%	68.56%

Industry	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Retail	75.88%	75.18%	63.23%	87.08%	66.97%
Telecom	74.79%	78.50%	72.89%	87.74%	78.54%
Travel	64.24%	66.49%	55.04%	81.38%	67.66%
Gaming	72.12%	77.61%	45.45%	86.13%	39.49%

Channel	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Chat	69.59%	69.75%	62.88%	84.04%	67.75%
Email	72.70%	76.65%	60.71%	86.29%	69.36%

2. RASA has the highest overall coverage rates

Coverage, or Recall or True Positive Rate or Sensitivity (TPR), is the ratio of correct positive predictions out of all predictions. In other words, coverage is a measure of how well an AI is able to identify the actual topics. An AI with higher coverage rates results in lower escalation to human agents for trained intents, ultimately reducing customer support costs. In the study, RASA has the highest coverage (69.61%), followed by Google Dialogflow (69.45%), IBM Watson (66.80%), Microsoft LUIS (58.80%) and Netomi (44.47%).

It's important to note that lower coverage rates may indicate that an AI platform is risk-averse and could have higher confidence thresholds set. AI platforms with lower coverage have been trained to escalate a user to a human agent or take another course of action rather than take the chance of providing the wrong response, resulting in a suboptimal user experience if they're not confident in their classification. There's a trade-off between sensitivity (coverage) and specificity (out-of-scope accuracy). AIs with higher sensitivity have lower specificity, and vice versa.

Coverage rate performance across industries

It's not surprising that coverage rates varied, sometimes quite significantly, based on the industry. In retail, Google Dialogflow has the highest coverage rates at 72.56%, while RASA has the highest coverage rates (78.22%) in telecom. For travel queries, RASA has the highest coverage rates (68.70%), followed closely by Google Dialogflow (63.97%) and IBM Watson (61.33%). Coverage rates in the gaming industry were the lowest compared to other industries for every platform except IBM Watson which had 63.53% coverage.

Coverage rate performance per channel

When we look at the coverage rates across all platforms, RASA has the highest on chat at 71.59%, followed by Google Dialogflow (67.56%) and IBM Watson (65.31%). On Email, Google Dialogflow has the most coverage at 71.53%, followed by IBM Watson (68.28%) and RASA (67.64%).

Metric	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Coverage	69.45%	66.80%	58.80%	44.47%	69.61%

Industry	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Retail	72.56%	66.40%	62.53%	55.97%	69.90%
Telecom	77.61%	74.85%	68.36%	45.81%	78.22%
Travel	63.97%	61.33%	52.00%	40.51%	68.70%
Gaming	51.86%	63.53%	39.67%	12.88%	38.63%

Channel	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Chat	67.56%	65.31%	59.12%	50.97%	71.59%
Email	71.35%	68.28%	58.48%	37.97%	67.64%

3. Netomi is the most likely to identify topics it's not trained on, resulting in less user frustration

Out of Scope Accuracy, or True Negative Rate or Specificity (TNR), is the ratio of correct negative predictions out of all negatives. If an AI has low out-of-scope accuracy, user frustration increases as the bot does not give the correct response or take the best course of action when it has not been trained on a topic. On the contrary, AI with high out-of-scope accuracy decreases user frustration as it understands which topics it is not trained on and follows the appropriate behavior (i.e. escalates a user query to a human agent or directs them to another channel).

In the Conversational AI Benchmarking Report, Netomi has the highest out-of-scope accuracy at 92.45%, signaling that it causes the least amount of user frustration associated with not incorrectly answering a question that it has not been trained on. IBM Watson performs second best at only 52.82%, followed by Google Dialogflow (36.45%), Microsoft LUIS (19.65%) and RASA (10.64%). When we analyze specific datasets, Netomi outperforms Microsoft LUIS, RASA and Google Dialogflow on all benchmark datasets, and IBM Watson on 12 out of 14 datasets.

Out-of-scope accuracy industry analysis

Out-of-scope accuracy for Netomi's AI is the highest across all four industries. Netomi's out-of-scope accuracy is up to 88.73% points (pp) higher than other leading AI platforms for telecom, 80.74% points (pp) for retail, 78.57% points (pp) for gaming and 77.77% points (pp) for travel.
In retail, Netomi performs at 94.15% for out-of-scope accuracy and IBM Watson at 81.00%. RASA scores lowest at only 13.41%. In telecom, Netomi has out-of-scope accuracy of 92.56%, with the closest competitor, IBM Watson, performing at 47.11%. In travel, Netomi has out-of-scope accuracy at 92.35%, with the closest competitor, IBM Watson, performing at 41.11%. In gaming, Netomi has out-of-scope accuracy at 85.71%, while Rasa scored lowest at only 7.14%.

Out-of-scope accuracy channel analysis

The out-of-scope accuracy for AI platforms differs across channels, for some more than others. Netomi performs high on both email and chat (90.99% and 93.91% respectively), while RASA is consistently low across channels: 9.43% on chat and 11.85% on email. IBM Watson had a spread of nearly 15% points, underperforming on email at 45.06% compared to 60.57% on chat.

Metric	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
OOS Accuracy	36.45%	52.82%	19.65%	92.45%	10.64%

Industry	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Retail	57.68%	81.00%	20.80%	94.15%	13.41%
Telecom	23.52%	47.11%	22.93%	92.56%	3.83%
Travel	26.39%	41.11%	15.75%	92.35%	14.58%
Gaming	53.57%	21.43%	21.43%	85.71%	7.14%

Channel	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Chat	44.98%	60.57%	22.70%	93.91%	9.43%
Email	27.92%	45.06%	16.60%	90.99%	11.85%

4. Netomi has the highest balanced accuracy, resulting in the least frustrating experiences

Balanced Accuracy accounts for both coverage rates and out of scope accuracy. It's the measure of value delivered by a conversational AI agent and when Balanced Accuracy is low, user frustration can be expected due to the bot's incorrect or irrelevant responses. For Balanced Accuracy, Netomi scores highest at 68.46%, followed by IBM Watson (59.81%), Google Dialogflow (52.95%), RASA (40.13%) and Microsoft LUIS (39.52%). When we analyze performance of AI coverage across specific datasets, Netomi outperforms Microsoft LUIS and RASA on all benchmark datasets, Google Dialogflow on 13 out of 14 datasets, and Watson on 12 out of 14 datasets.

Balanced Accuracy is higher on chat than email

Netomi's AI has higher balanced accuracy on both chat and email. Netomi scores up to 31.93% points (pp) higher on chat and 26.94% points (pp) higher on email than both Microsoft LUIS and RASA. All companies outperform on chat compared to email.

Netomi has the top performance in retail, telecom and travel while Google DialogFlow tops gaming

Netomi has the highest balanced accuracy in retail, travel and telecom. In these industries, Netomi outperforms other leading AI platforms by up to 33.4% points (pp) in retail, 32.55% points (pp) in travel and 28.17% points (pp) in Telecom. IBM Watson performs second best in these industries.

Metric	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Balanced Accuracy	52.95%	59.81%	39.23%	68.46%	40.13%

Channel	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Chat	56.27%	62.94%	40.91%	72.44%	40.51%
Email	49.63%	56.67%	37.54%	64.48%	39.75%

Industry	Google DialogFlow	IBM Watson	Microsoft LUIS	Netomi	RASA
Retail	65.12%	73.70%	41.67%	75.06%	41.66%
Telecom	50.56%	60.98%	45.64%	69.19%	41.02%
Travel	45.18%	51.22%	33.88%	66.43%	33.88%
Gaming	52.71%	42.48%	30.55%	49.30%	30.55%

Designing AI-powered Experiences through the Frustration Lens

When we look holistically at different conversational AI performance in the context of the end-user experience, customers are anywhere between 0.6X - 7.44X less frustrated engaging with Netomi-powered bots as compared to other AI platforms. This is because IBM Watson, Microsoft LUIS, RASA and Google Dialogflow are more likely to respond incorrectly to topics that they have not been trained on. With conversational AI agents, there are explicitly trained topics which are within scope, and queries that fall outside of this scope need to be managed correctly. The Netomi AI is anywhere between 11.97% - 23.38% points higher at predicting the right course of action as compared to other leading AI platforms.

Based on accuracy, out of scope accuracy and balanced accuracy, Netomi outperforms all other AI platforms in the correctness of responses sent to end-users.

Best Practices to drive delightful CX with AI

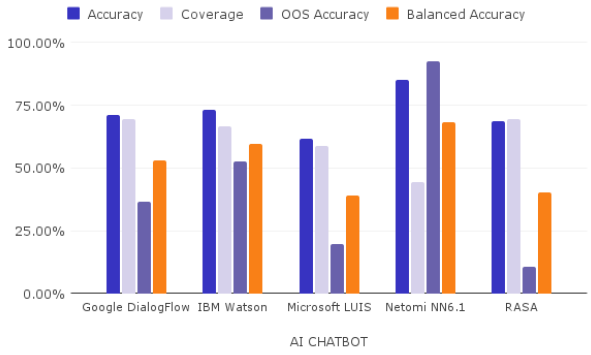
To provide the best possible user experience, companies must minimize the frustration felt by end users. This can be done in a few ways.

First, when determining the use cases to delegate to an AI-powered virtual assistant, select highly repeatable topics that have sufficient data to get started with limited risk. Less frequent or more complex cases should be immediately escalated to a human agent. This allows an AI to learn behind the scenes how a human agent is responding in different circumstances, so it can handle more cases in the future.

AI-powered bots should also take the best course of action based on their confidence level in understanding a user's intent. It's recommended that unless the bot is highly confident, it should not try to respond. When the bot is less confident, there are a few ways forward. A bot can use an explicit intent to confirm its understanding before triggering an entire flow. An explicit intent asks a person to acknowledge the intent was understood, for example: "You want to find an Italian restaurant for dinner, is that correct?" This lets the user correct the bot if needed. You can also use implicit intents which have the bot acknowledge a user and repeat its understanding, allowing a user to correct the bot but not needing input to move forward. An example is: "Let's see what Italian restaurants have a table for dinner. Would you like outside or inside seating?"

Conversational AI deployments need constant optimization to uncover both external market changes and internal operational changes. Just like with human resources - the bot's learning never stops; your bot needs to learn about new utterances (or things people say to it) and needs new pathways to reach a resolution. Regular bot optimization can increase coverage rates, as well as identify emerging topics ripe for explicit training.

Metrics like accuracy, out-of-scope accuracy and balanced accuracy give great insight into the end customer experience. Similar to measuring customer satisfaction following an interaction, these AI performance metrics provide a way to measure if a bot is leading to customer frustration (which could lead to churn) or providing the best possible automated conversation.



About Netomi

Netomi is a provider of AI-first customer experience that creates unprecedented brand access and intimacy in the Relationship Economy. Netomi's Relationship Operating System automatically resolves 80% of customer service inquiries, decreasing resolution time, increasing customer satisfaction and support quality, while reducing costs. The patented, no-code platform works across messaging, chat, email and voice, and understands 100+ languages. Netomi is based in San Francisco and has offices in New York, Toronto and India. Investors include WndrCo, Eldridge and Fin Venture Capital.

For more information, visit https://netomi.com.