CASE STUDY

TESTING VIRTUAL AGENTS (VA) WHEN OUTPUTS ARE GENERATED BY LARGE LANGUAGE MODELS (LLM).

A HOLISTIC APPROACH →

TABLE OF CONTENTS

CASE & GOAL

A company creates AI agents for customer support. Large Language Models bring a variety of opportunities, but also challenges with them. In order to push the company's mission forward, the AI agents need to provide customers with personalized and accurate answers and ultimately become indistinguishable from a human.

However, the current testing approach is insufficient for testing LLM generated outputs and their quality. The goal is to identify problems, come up with solutions and develop a product roadmap for the new approach.

PROBLEMS

The output speed and resource utilization of the LLMs can be assessed, however the qualitative output is difficult to track. A lack of consistent quality in the LLMs output has adverse effects on our customers and their user base:

CUSTOMER MIGHT MISS OUT ON SALES

When the LLM provides incorrect or misleading information, the customer satisfaction and conversion can drop as a result

REPUTATIONAL DAMAGE

The end consumer might take offense, because of biased answers or the use of disrespectful language, due to cultural differences.

LOSS OF TRUST & ENGAGEMENT

When answers lack are generic, the end consumer and customer might lose trust and stop engaging with the system

LOW END CONSUMER SATISFACTION

When inquiries aren’t answered in a satisfactory way, customers become frustrated and might not purchase again.

SOLUTIONS

Taking a holistic approach between establishing new testing routines and continuously improving the LLM.

MANUAL TESTS & RATING

1. By Cognigy staff
2. By external paid testers
3. By implementing a simple rating system (thumbs up/down) in the customer’s VA

PROGRAMMATIC TESTING

1. Generate and run questions based on dataset and have own or external LLM assess the **answer’s correctness** and whether it’s made up
2. Generate and run questions, re-generate response and evaluate **accuracy** of content

establish average score for each solution, track scores, identify low-ranked answers and their cause, use data to train LLM

TRAINING LLM ON

1. Datasets of different languages and cultures to enable culturally nuanced answers, if possible
2. Industry specific data set to enable better context capturing and context-based answers
3. Inclusive and unbiased language

PRIORITIZATION

The LLM optimizations are an on-going effort and don’t solve the testing issue itself. Hence they’ve been omitted.

	Rating each 1-10	IMPACT	CONFIDENCE	EASE	TOTAL	
MANUAL	Cognigy Staff	4	6	5	16	4
	External Paid Testers	5	6	4	15	5
	Customer Rating System	10	7	6	23	3
AUTO	Correctness Test	9	8	8	25	1
	Re-Generation Test	8	8	8	24	2

DOWNSIDES / RISKS

1. **CUSTOMER RATING SYSTEM** • GDPR, data access, customer agreement, vast amount of data
2. **ANSWER ACCURACY TEST** • Self-check leaves room for error
3. **ANSWER CORRECTNESS TEST** • Self-check leaves room for error, security risk through third party assessment
4. **COGNIGY STAFF** • Too experienced at prompting, potentially biased, only possible for sampling
5. **PAID USERS** • Need for payment, hard to obtain meaningful data set (across industries, cultures,...)

METRICS

How can we measure the success of our prioritized solutions?

RATING SYSTEM FOR END CONSUMER [RS]

1. Rating in % o. score
2. Customer specific rating

RE-GENERATION TEST [GT]

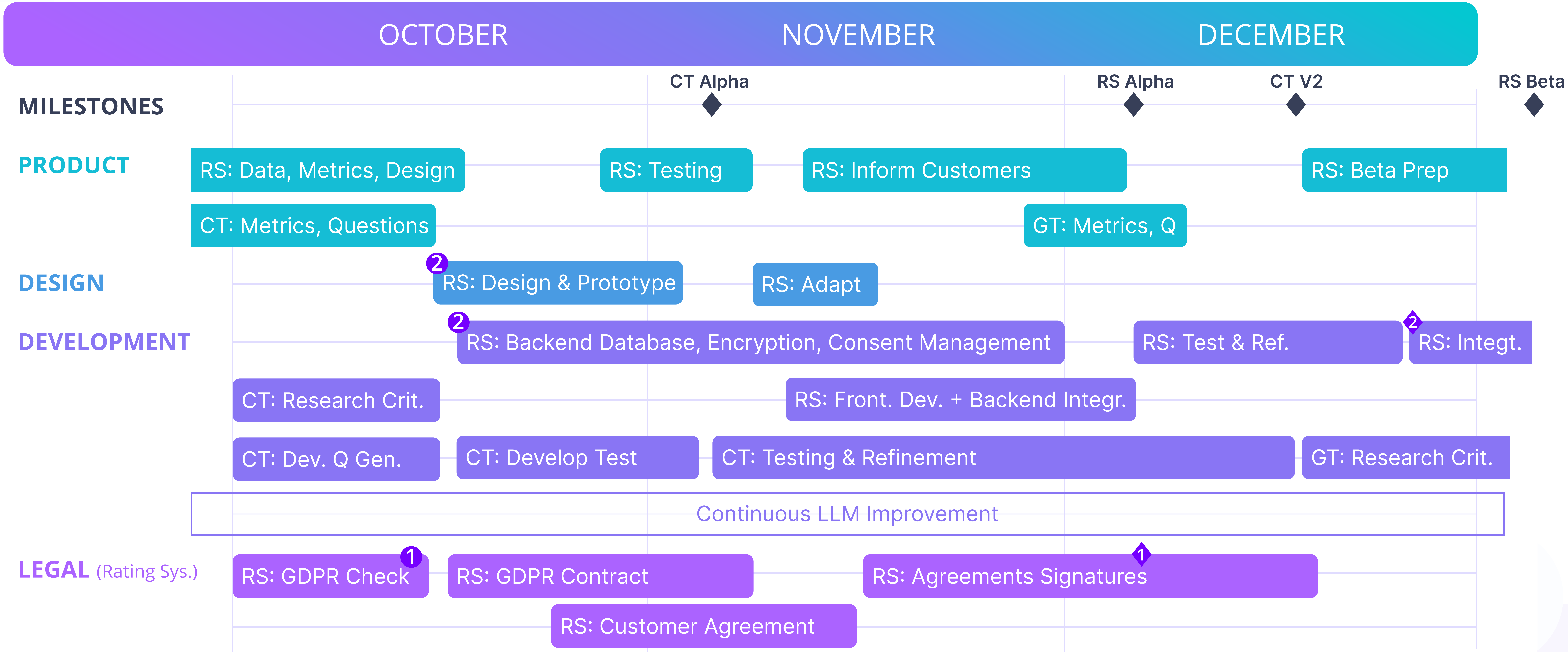
1. Ø Accuracy in %
2. % of contradictory answers

CORRECTNESS TEST [CT]

1. % of correct vs incorrect
2. % of made up answers

Differentiated across different languages and industries

DELIVERY PLAN



OCTOBER NOVEMBER DECEMBER

MILESTONES

CT Alpha

RS Alpha

CT V2

RS Beta

PRODUCT

RS: Data, Metrics, Design

RS: Testing

RS: Inform Customers

RS: Beta Prep

CT: Metrics, Questions

GT: Metrics, Q

DESIGN

2 RS: Design & Prototype

RS: Adapt

DEVELOPMENT

2 RS: Backend Database, Encryption, Consent Management

RS: Test & Ref.

2 RS: Integt.

CT: Research Crit.

RS: Front. Dev. + Backend Integr.

CT: Dev. Q Gen.

CT: Develop Test

CT: Testing & Refinement

GT: Research Crit.

Continuous LLM Improvement

LEGAL (Rating Sys.)

RS: GDPR Check 1

RS: GDPR Contract

RS: Agreements Signatures 1

RS: Customer Agreement

USER STORY: #1 RATING SYSTEM

As a product manager, I want to have a clear understanding of the end consumer’s satisfaction and gain qualitative insights into the LLMs performance, so I can identify negative experiences and their causes and improve the LLM to provide more accurate output for the VA.

Acceptance Criteria:

- After their conversation with the VA, users encounter a feedback dialogue titled "How was your experience?" featuring thumbs up and down icons to indicate their answers
- Users can skip the rating screen if desired
- A "Thank you" message displays once users submit their ratings
- Customer data is collected, including customer_ID [business customer], industry, conversation_history, and rating [positive or negative].
- The collected data is analyzed to create both customer-specific and total positive vs. negative interaction ratings as percentages.
- The ratings and associated customer data are stored securely in a SQL database, adhering to data privacy regulations.
- An API endpoint is established to retrieve and utilize the ratings and customer data for analysis and improvements.

CONCLUSION

In order to ensure great quality output, the most effective and efficient way is to:

1. Collect customer data through rating system
2. Build self-assessment systems for the LLM
3. Utilize data to identify weaknesses
4. Specifically improve those areas of the LLM