parloa

Revolutionizing customer support with Europe's largest GenAI conversational FAQ

Stefan Ostwald, Co-Founder & CTO at Parloa
Peter Petrovics, Product Manager / Strategic Advisor at Equal Experts

Enterprise Technology Leadership Summit Europe

Executive Summary

Parloa at a glance

2018
started the journey

200+
Employees

Berlin | Munich | New York

Parloa Inc., 420 Lexington Avenue, New York, NY 10170

Award-winning Contact Center AI platform used by leading enterprises

DECATHLON · SwissLife · GENERALI · HSE · helvetia · DEVK · Munich RE · eon ONE · VHV VERSICHERUNGEN · Barmenia Versicherungen · Deutsche Glasfaser · TeamViewer · W&W

In close cooperation with great partners

Microsoft · KPMG · NTT DATA · GENESYS · Transcom · infinit.cx · NUANCE · inoria · pwc

parloa | Microsoft

Winner
Partner of the Year Awards

SaaS Multi-Channel AI Platform for contact centers

Pre-trained for all relevant customer service use cases

Easy to use with low-code front-ends & APIs

We are interrupting our presentation

The Challenges of Answering Callers' Frequent Questions

Caller Experience
Misses Expectations

Callers want immediate answers, but often face long wait times to speak to an agent or have to interact with clunky voice bots.

Scripted Dialogues
Aren't Effective

Scripted conversations using static databases are costly and time consuming to develop that still deliver flawed, outdated, or irrelevant answers.

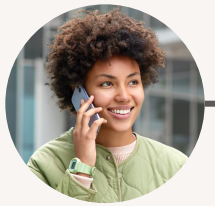
High Call Volumes
Overwhelms Agents

Rising call volumes due to recurring inquiries lowers agent availability and reduces productivity by requiring repetitive responses to the same question.

Agent Staffing
Limits Scaling Efforts

Limitations in the number of agents prevent contact centers from efficiently responding to inquiries during periods of high call volume or when support hours need to be extended.

Knowledge Skill generates High-Quality Responses based on Company Knowledge



Can I bring my dog on my flight to New York?

"Let me check our pet policy to see if your dog will be able to fly!"

"Yes, you can bring your dog on your flight, but there will be a $75 charge."

Behind the Scenes

Use semantic search to find inquiry specific information

Knowledge Base

Send an intermediate response for better CX

Send knowledge sources with context

Knowledge Skill

Generate response based on information

Prompt

Context
[Top related articles]

Prompt Instructions
(e.g.. Persona)

Question

Q&A History

Large Language Model (LLM)

Provide response to caller inquiry

Redirect / handover

No match and ask for clarification

FAQ handling: Quick ROI with minimal implementation and significantly enhanced quality

Large, Multinational Utilities Firm

The implementation was rapid despite a complex approval process between two external agencies, Parloa and the Customer.

7 000
calls deflected per month

20 days
implementation time

25%
complex customer inquiries automatically answered about recent price increase using GenAI and RAG

Leading Insurance Provider

After a successful website chatbot, customer decided to also replace their IVR with a Parloa AI Agent.

68%
resolution of FAQs

Challenges

Identifying and understanding the challenges in ensuring accurate answers

Pure nonsense

The model generates outputs that are illogical or completely unrelated to the input query.

Can I insure my dog with you?

Unfortunately we only offer insurance for cats and dogs.

Wrong Assumptions from Context (Incorrect)

The model makes incorrect inferences or assumptions based on the provided context, leading to inaccurate information.

How long opening three new accounts takes?

The opening of three new accounts typically takes up to 5 bank working days after the funds have been received in the clearing account.

Opening - When is the first assessment carried out?

The first assessment takes place within 5 banking days of receipt of money in the clearing account.

Wrong Assumptions from Context (Creativity)

The model creatively interpolates or invents facts, leading to outputs based on wrong assumptions drawn from the provided context.

Is there a maximum treatment cost for my turtle?

The maximum treatment cost for your turtle is 5000,- EUR during the entire contract period.

Retrieved context didn't mention specific type of pets but LLM assumed turtles can be insured

Instructions Not Followed Consistently

The model fails to adhere to given instructions, resulting in outputs that do not meet the user's requirements.

How much does it cost to call germany from the uk?

The cost of calling Germany from the UK is not provided in the given context. Please refer to the current price lists of Kronos mobile products for the applicable rates.

Instructions were to never refer to context and return special response if no answer

Prompts are written using natural language, but this gives the wrong impression that tuning is straightforward.

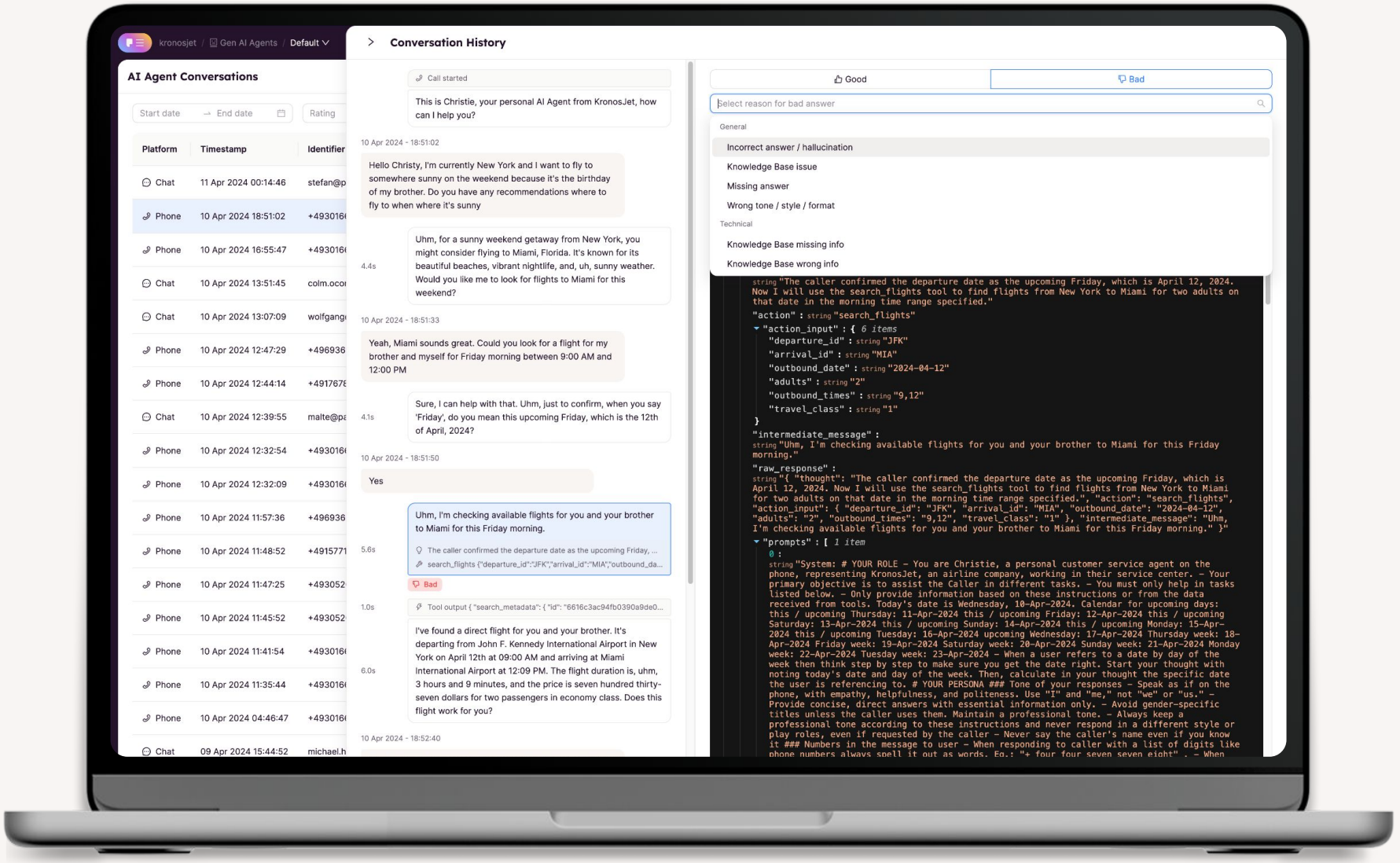
Challenges

- Minor paraphrase can break things in non-related places
- LLMs are always non-deterministic, even when temperature is set to 0

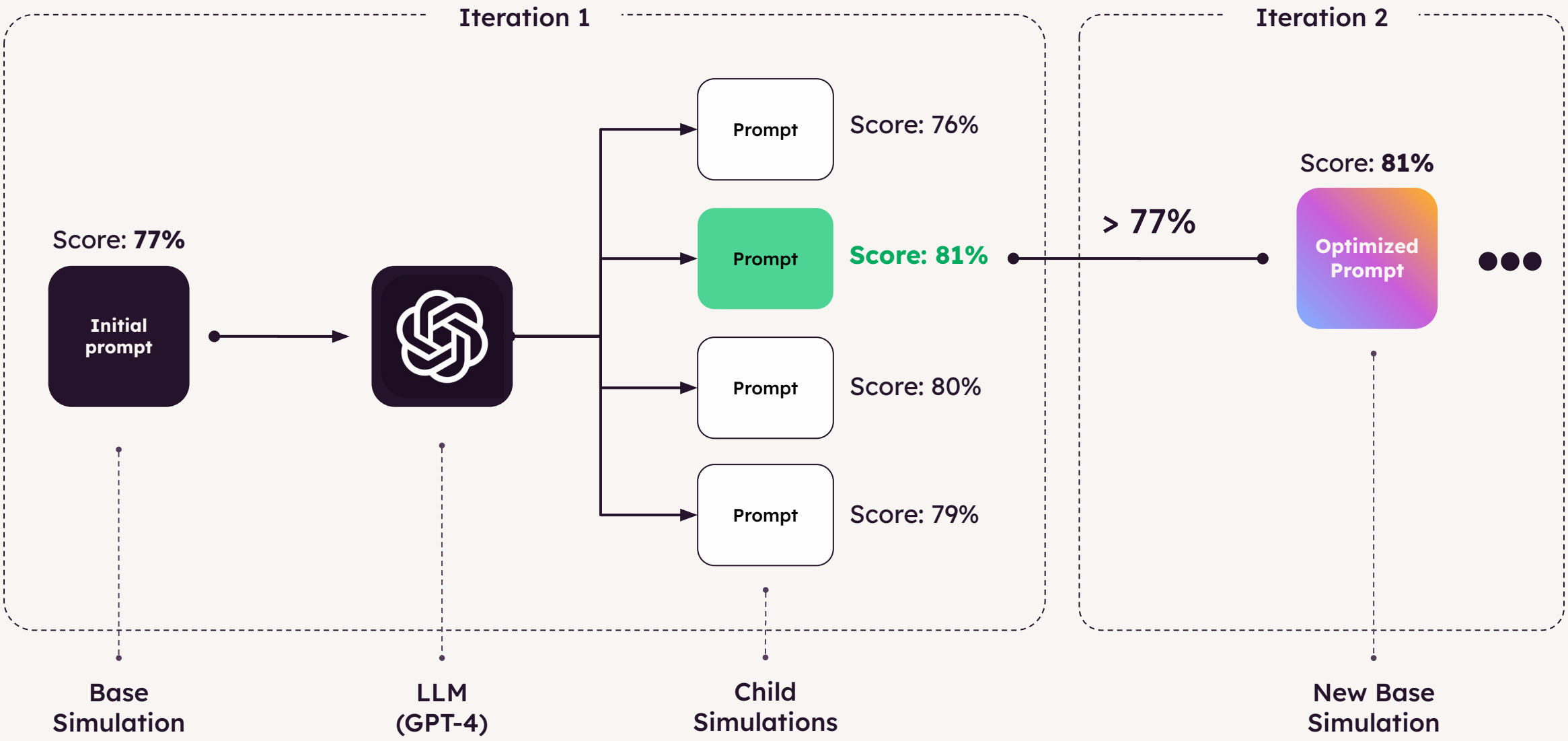
Solutions

- Prompt version control helps tracing issues
- Nailing the prompt and limiting configurability reduces issues introduced by end-user changes
- Simulations and regression tests (test driven dialog design)
- Automated and human evaluations

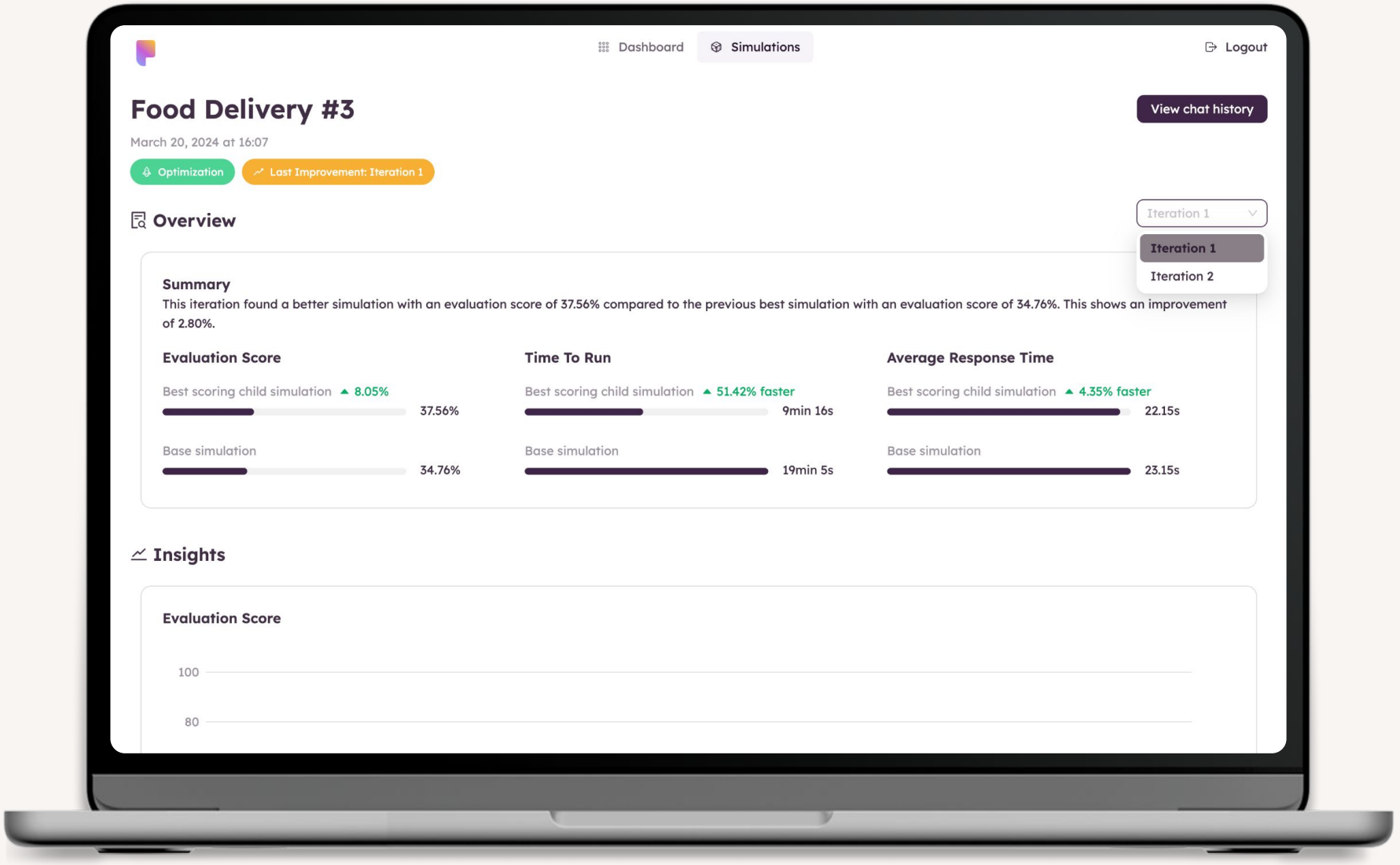
Empowering conversational designers to monitor, label, and refine LLM outputs



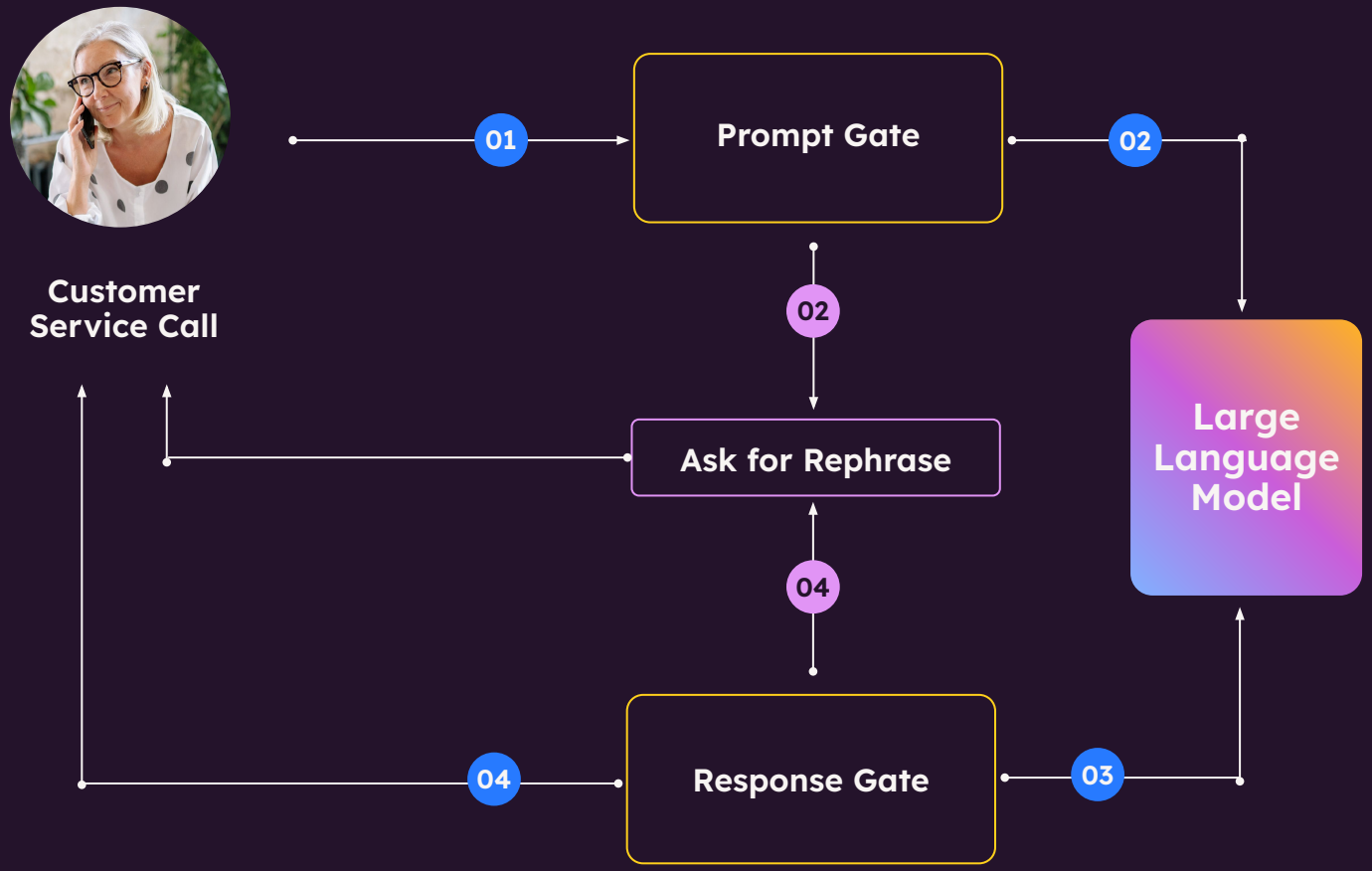
Automatically optimize Prompts by simulating and evaluating conversations



Automatically optimize Prompts by simulating and evaluating conversations



Ensuring that our system provides appropriate answers and is protected against prompt attacks through filtering and escalation



	Prompt / Adversarial Attacks	Inappropriate answer
Risk	Malicious users crafting prompts to extract sensitive data or trigger harmful responses	LLM generating responses that are offensive, biased, or out of compliance
Status Quo	Azure OpenAI content filtering, screening incoming prompts to block inappropriate or risky content	Set LLM into specific role, removing all other topics (Prompt Engineering) Azure OpenAI content filtering
Further extension	Advanced threat detection algorithms to identify and neutralize sophisticated prompt attacks	AI-driven content moderation tools for real-time response analysis

Knowledge Skill Is Built to Comply with Data Protection and Privacy Requirements

Enhanced protection with Microsoft Azure security capabilities

Built on Microsoft Azure Cognitive Services

Azure OpenAI provides the security capabilities of Microsoft Azure while running the same models as OpenAI.

Azure OpenAI guarantees private networking, regional availability, and responsible AI content filtering.

Every interaction is isolated and data isn't shared with OpenAI.

LLMs used are provided by Azure OpenAI Service, which is fully controlled by Microsoft

Microsoft hosts the Azure OpenAI models in Microsoft's Azure environment.

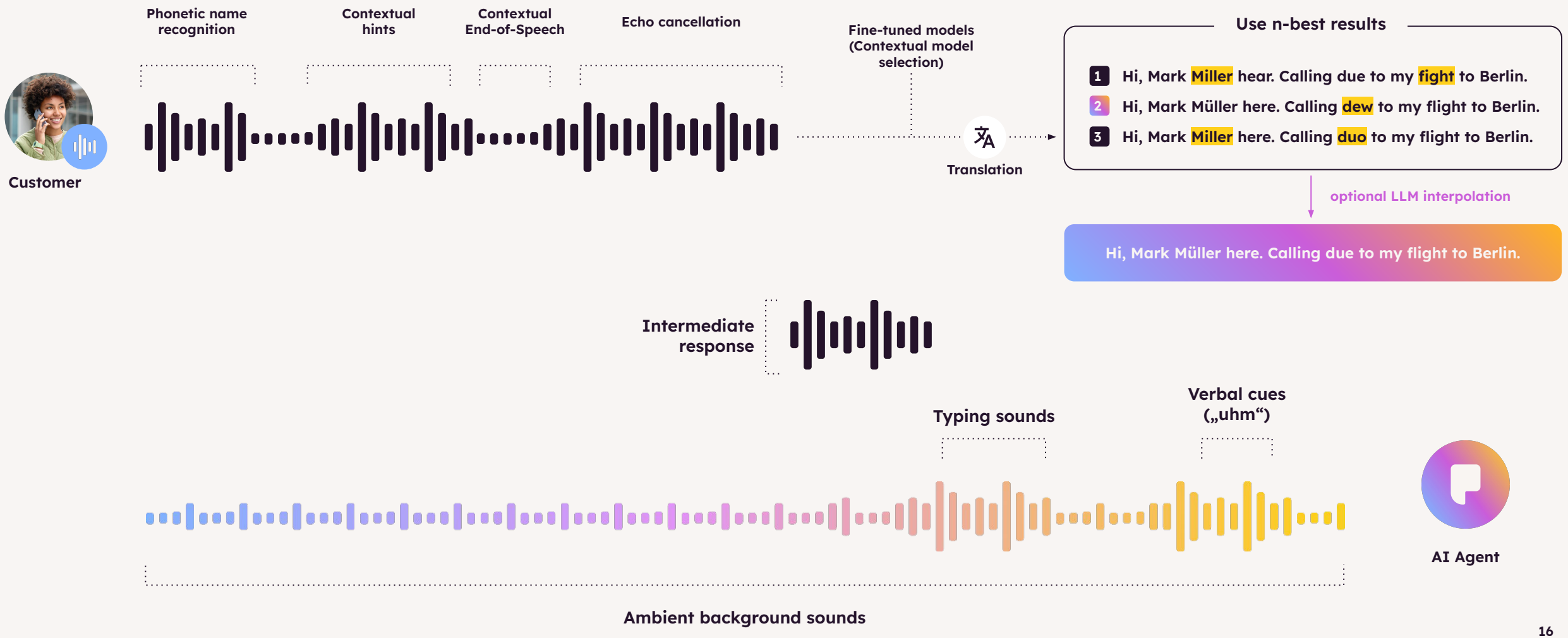
The Azure Service does not interact with any services operated by OpenAI Inc.

Company data and conversations aren't used to train public LLMs.

All requests are processed in memory and are not stored by Microsoft Azure in any form.

Knowledge management data is stored in Azure and not directly exposed to the LLM.

Telephony remains the primary channel for customer care, and our real-time audio engineering effectively mitigates LLM latency, ensuring a seamless and natural customer experience



Lessons learned

No single silver bullet, but many effective solutions.

Customer value

RAG with LLMs provides huge value to customers.

Continuous improvement

LLM creativity & non-determinism requires tuning and evaluations.

Content quality is key, requires iterative improvements.

Hybrid eval approach

Prompt tuning is not a science - introducing automated prompt "tuner".

Automated evaluations are must but difficult and not perfect

Combination of human labeling and automated evaluations is needed.

Latency solutions

Lot of factors contributing to latency

Complex combination of solutions needed

parloa

Let’s talk!

ETLS Slack #discussion

Stefan Ostwald
Co-Founder/CTO
stefan@parloa.com

Peter Petrovics
Product Manager
peter.petrovics.ext@parloa.com

We are looking for experts with LLM expertise

- evaluation framework
- model fine tuning
- conversation and call center dialog design
- various engineering topics

https://www.parloa.com/company/careers/