

# Building Applications with LLMs

From traditional ML  
engineering to AI engineering

**Sinan Tang @Zalando**

Women+ in Data and AI Festival  
27.09.2024



# Agenda

- **Demystifying AI engineering**
- **Meet the Zalando Assistant**
- **AI engineering techniques:  
adapting models**
- **AI engineering techniques:  
evaluation**
- **Summary**
- **Q&A**



# 01

## **Introduction to AI Engineering**

The AI tech stack



# Demystifying AI Engineering

Welcome to this year's  
Women+ in Data and AI \_ 

Language Models

Large Language Models

Foundation Models



# Demystifying AI Engineering

...

Welcome to this year's  
Women+ in Data and AI Conference

Talk  
Event

...

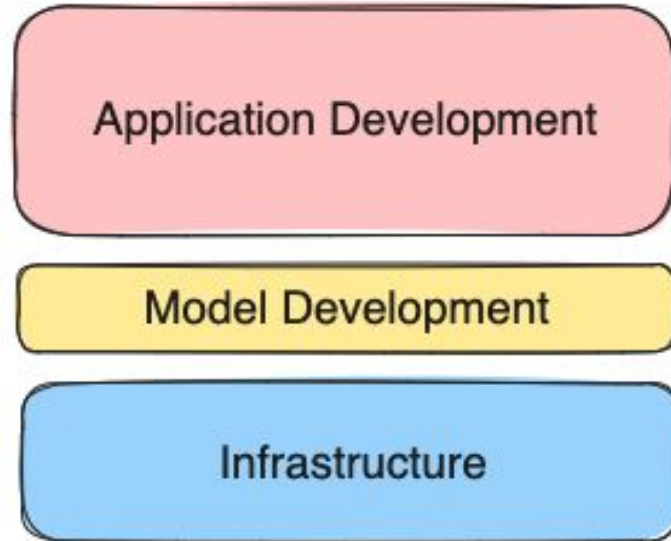
Language Models

Large Language Models

Foundation Models

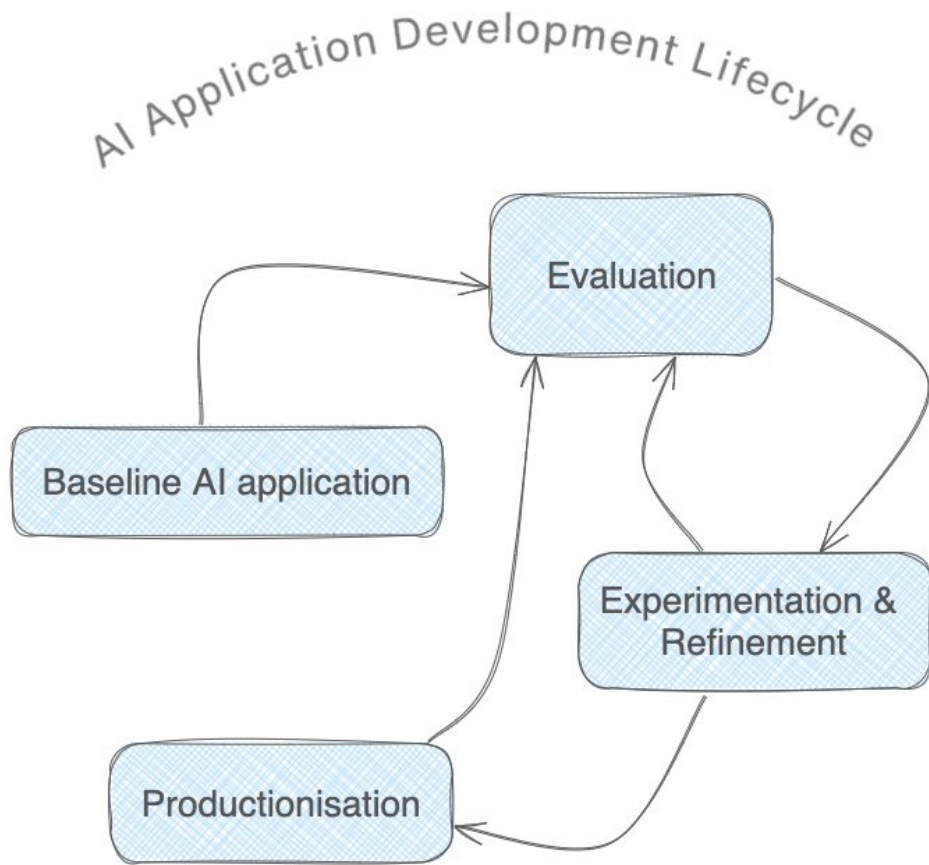


# AI Engineering Stack



# AI Engineering vs. ML Engineering

“it’s less about model development and more about adapting and evaluating foundation models”



# 02

## Meet the Zalando Assistant

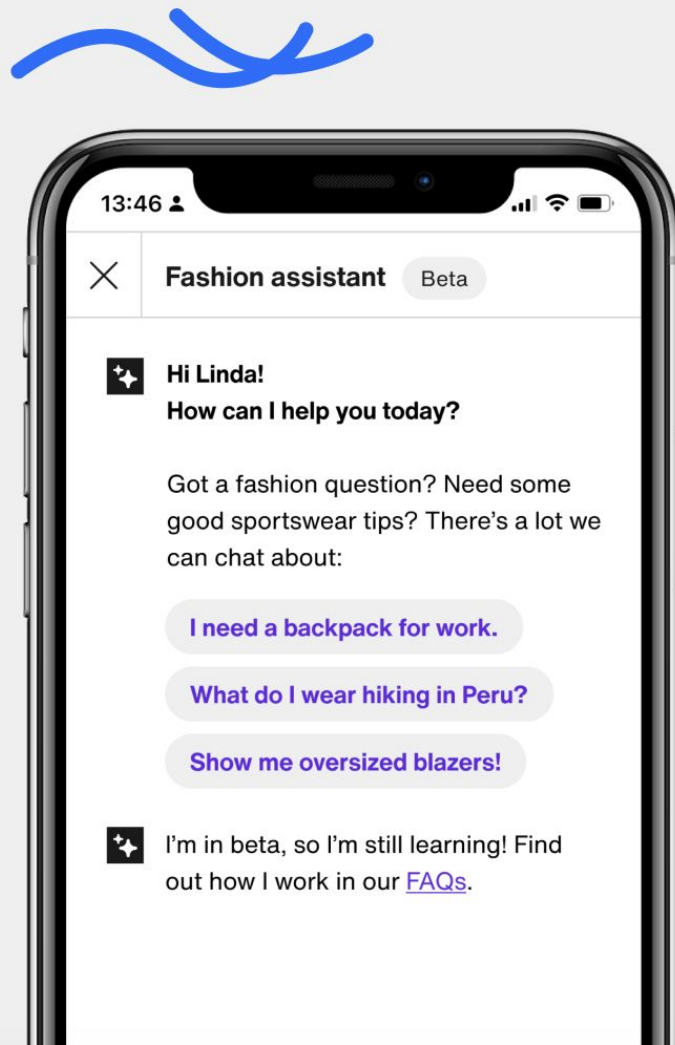


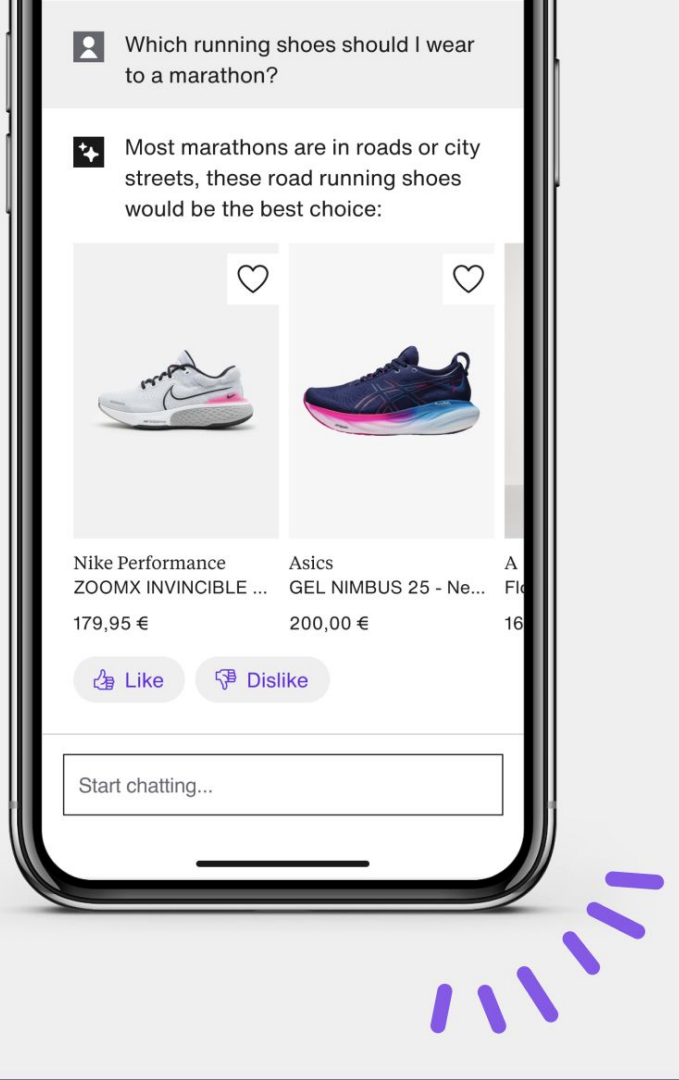


# Zalando Assistant (ZA)

ZA is an AI-powered experience allowing Zalando customers to discover fashion items, style tips and more by using their own language.

It's multilingual and able to interact with customers in any European language.





**The core ZA experience is completely dynamic:** A fluid conversation between the customer and the assistant (powered by ChatGPT and Zalando's own semantic search model).

*"I'm looking for a wardrobe refresh. Bright colours, fun patterns, unusual cuts."*

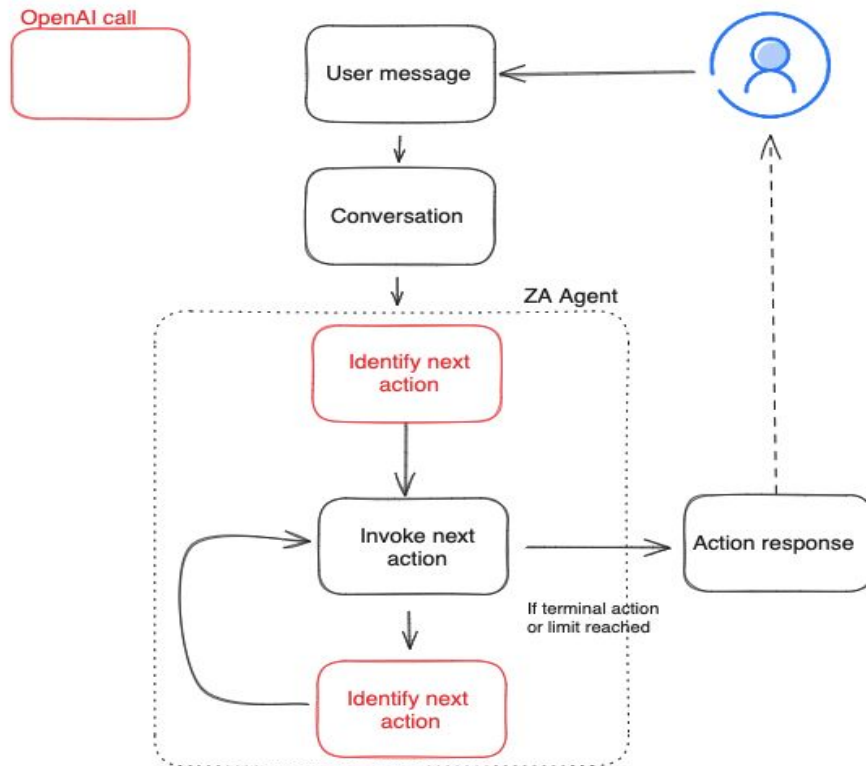
*"I need sport fashion ideas"*

*"Need help to buy my fiancé a birthday present!"*

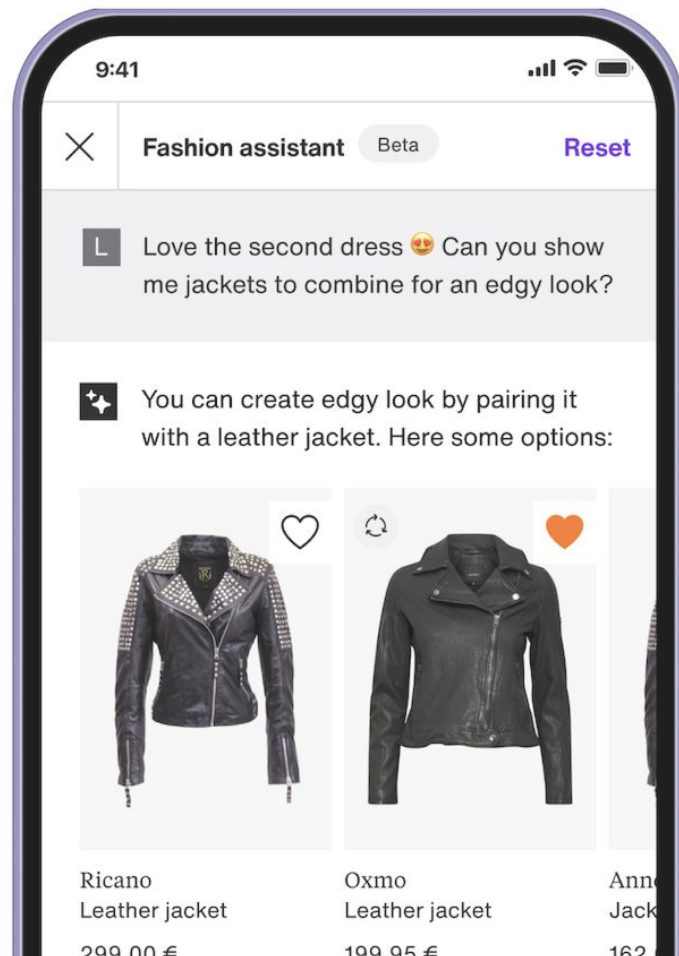
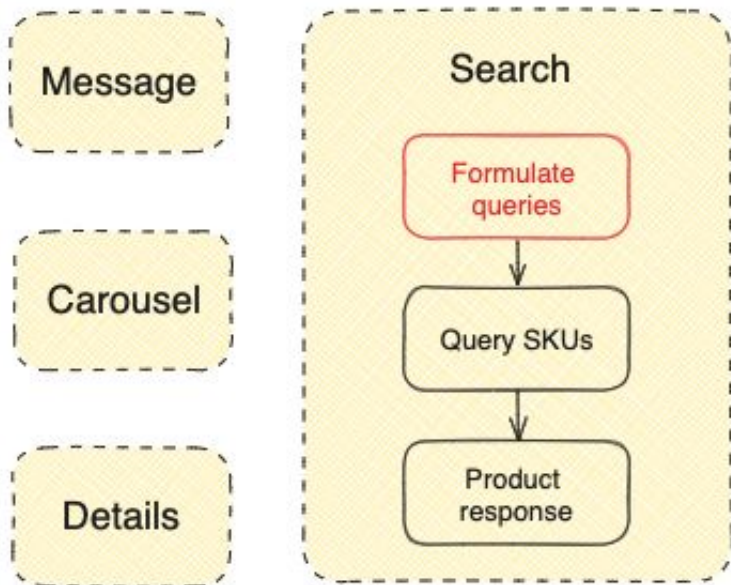
*"I want Stockholm style"*

# Integrating LLM into Zalando Assistant

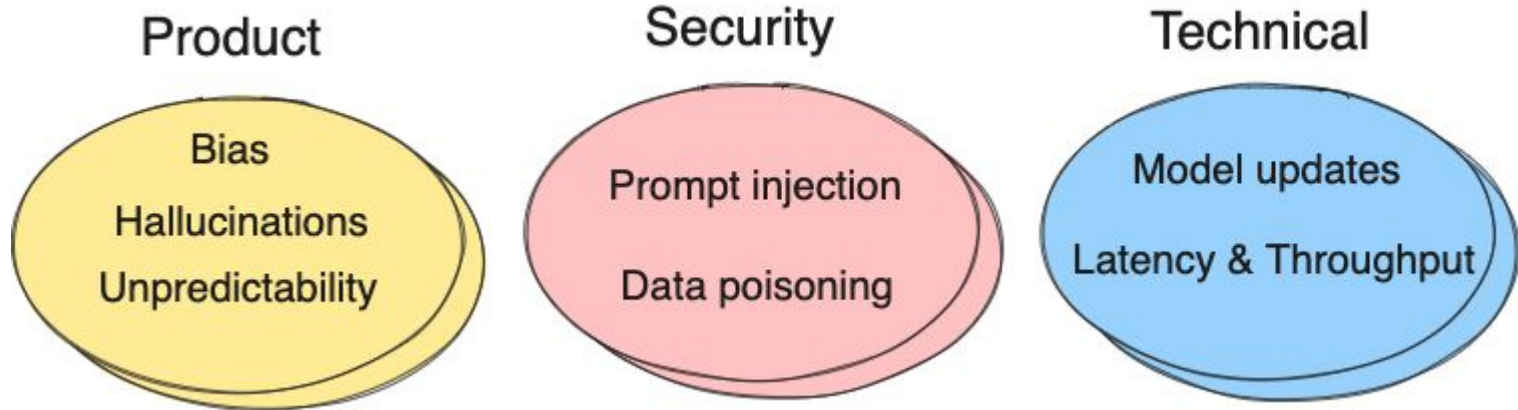
## Main Flow



# Integrating LLM into Zalando Assistant Actions



# Challenges building Zalando Assistant & working with LLM



# 03

## **AI Engineering Techniques** Adapting AI Models



# How to make AI work for you

## Prompting

What is a good prompt?

The **TELeR** framework <Turn, Expression, Level of details, Role>

Level 0	No directive, Just Data
Level 1	Simple one-sentence directive expressing the high-level goal
Level 2	Multi-sentence (paragraph-style) directive expressing the high-level goal and the sub-tasks that need to be performed to achieve the goal
Level 3	Complex (bulleted-list-style) directive expressing the high-level goal along with a detailed bulleted list of sub-tasks to be performed
Level 4	A complex directive that includes the following: 1) Description of High-level goal, 2) A detailed bulleted list of sub-tasks, 3) A guideline on how LLM output will be evaluated/ Few-Shot Examples.
Level 5	A complex directive that includes the following: 1) Description of High-level goal, 2) A detailed bulleted list of sub-tasks, 3) A guideline on how LLM output will be evaluated/ Few-Shot Examples, 4) Additional relevant information gathered via retrieval-based techniques.

Level 6

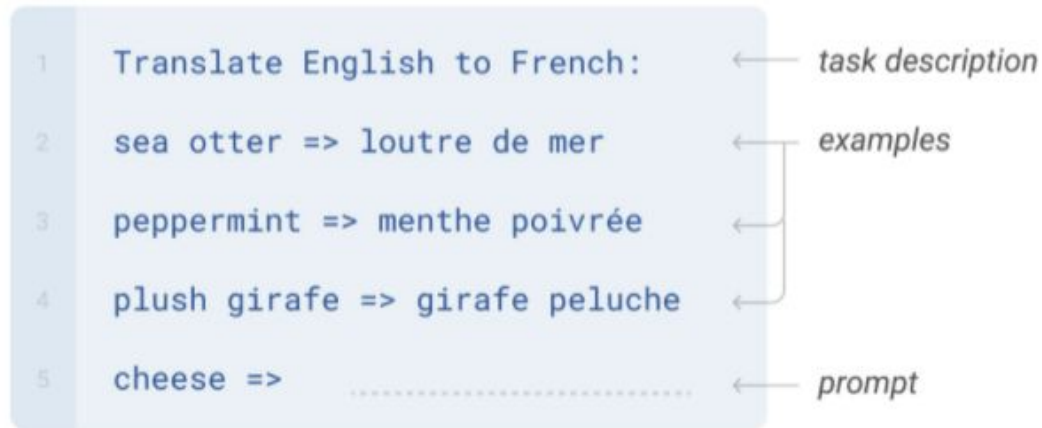
A complex directive that includes the following: 1) Description of High-level goal, 2) A detailed bulleted list of sub-tasks, 3) A guideline on how LLM output will be evaluated/ Few-Shot Examples, 4) Additional relevant information gathered via retrieval-based techniques, 5) An explicit statement asking LLM to explain its own output.

# How to make AI work for you

## Prompting

Few-shot learning

Prompt



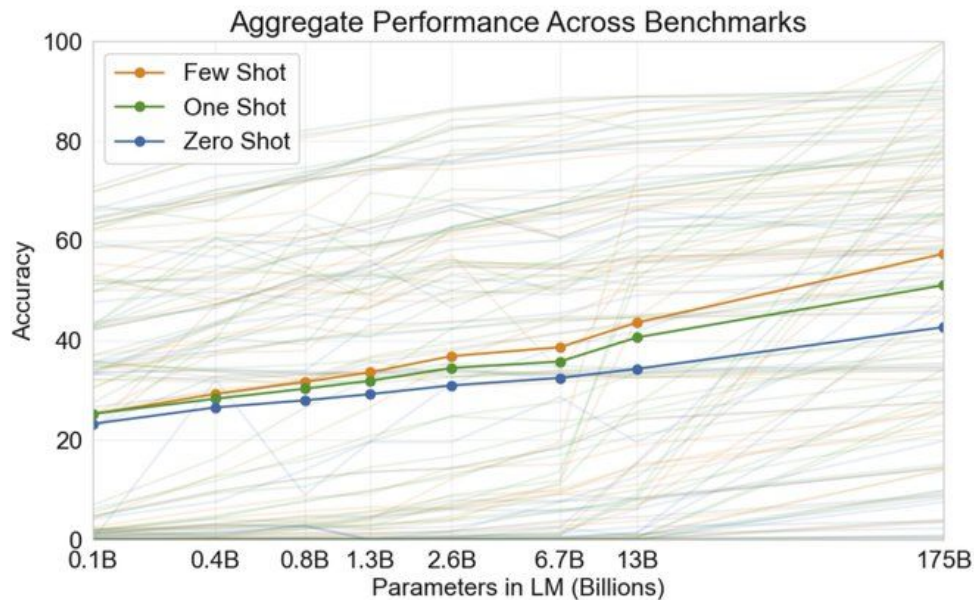


# How to make AI work for you

## Prompting

Few-shot learning

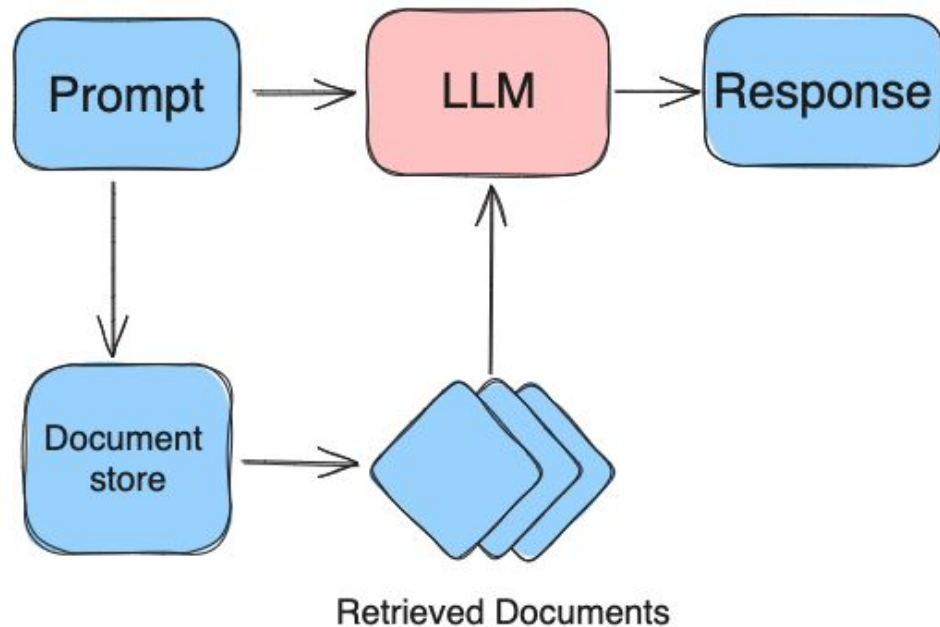
Performance benchmark




# How to make AI work for you

RAG

## Retrieval-Augmented Generation





# How to make AI work for you

## Limitations


### Limitations of prompt engineering & RAG

- ❖ Reliance on prompt quality
- ❖ Complexity and iteration
- ❖ Domain specificity
- ❖ Potential bias
- ❖ Limited control on the output
- ❖ Limited context window

# 04

## AI Engineering Techniques Evaluation





# How do you know it's good (enough)

## The problem of evaluation

“Lack of evaluations has  
been a key challenge for  
deploying to production.”

— OpenAI Dev Day 2023

# How do you know it's good (enough)

## Challenges

Design an itinerary for a  
weekend trip to Berlin



Day 1: Arrival and Exploring Central Berlin

Morning:

- Arrival in Berlin.
- Visit the Brandenburg Gate.

Lunch:

- Lunch at Augustiner am Gendarmenmarkt.

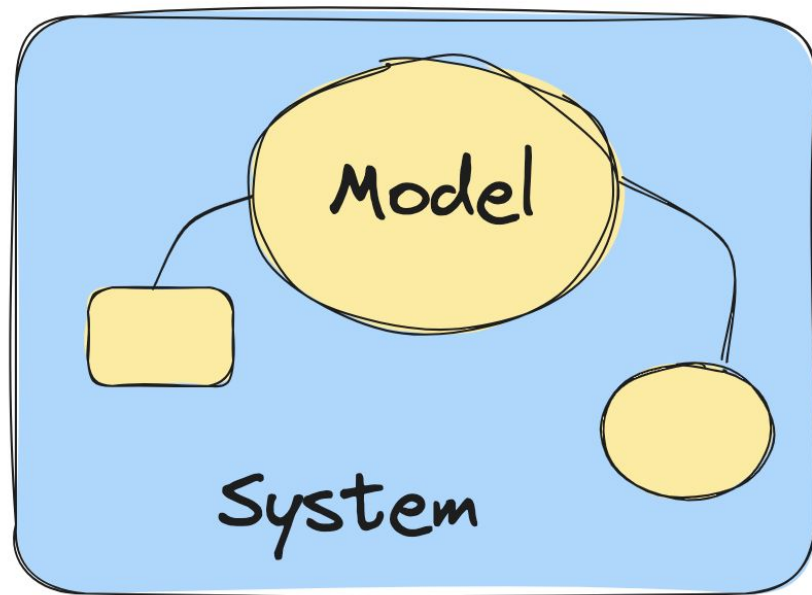
Afternoon:

- Explore Museum Island.
- Walk to Alexanderplatz.

...

# How do you know it's good (enough)

Evaluating  
open-ended  
models

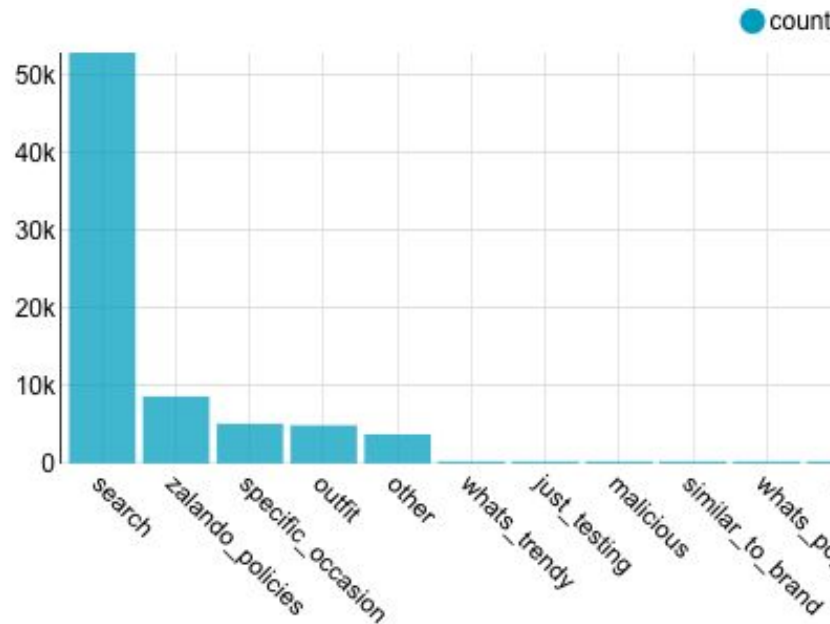


# ZA: Scalable Evaluation

## Evaluation Strategy

- Offline evaluation
  - Objective metrics (punchcard evaluation)
  - AI-as-a-Judge (LLM-based evaluation)
  - Human & machine annotations
- Comprehensive testing suite
  - Regression tests
  - Unit tests
  - Smoke tests
  - Conversation replay

Most common conversation types





# ZA: Scalable Evaluation



## Monitor

We can monitor if a change or a bugfix has the desired effect on a granular level.

Dashboard Conversation ZFA Conversation Tool Live

Sales Channel: All Annotation: GPT Customer Reaction: All Submit

Total: 8918

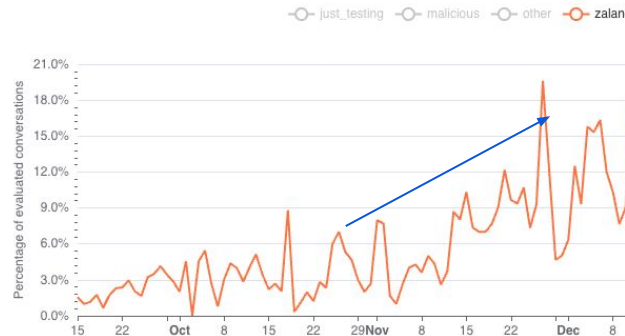
Date	Sales Channel	Language	Quality	Intent Tag	Issue Tag	Annotated By
2024-01-01 00:16	Germany	de	not_at_all	search		lim_eval
2024-01-01 00:15	Germany	de	very	zalando_policies		lim_eval
2024-01-01 00:12	Germany	de	neutral	search		lim_eval
2024-01-01 00:10	Germany	de	somewhat	search		lim_eval
2024-01-01 00:08	Germany	de	not_at_all	zalando_policies	FASHION_ONLY	lim_eval
2024-01-01 00:07	Germany	de	somewhat	search		lim_eval
2023-12-31 23:54	Germany	de	somewhat	search		lim_eval
2023-12-31 23:51	United Kingdom	en	very	search		lim_eval
2023-12-31 23:50	Germany	de	neutral	search	WRONG_SIZE	lim_eval
2023-12-31 23:49	Germany	de	neutral	search	WRONG_BRAND	lim_eval

< 1 2 3 4 5 ... 892 >

## Find

We can more easily find examples of conversations with specific issues.

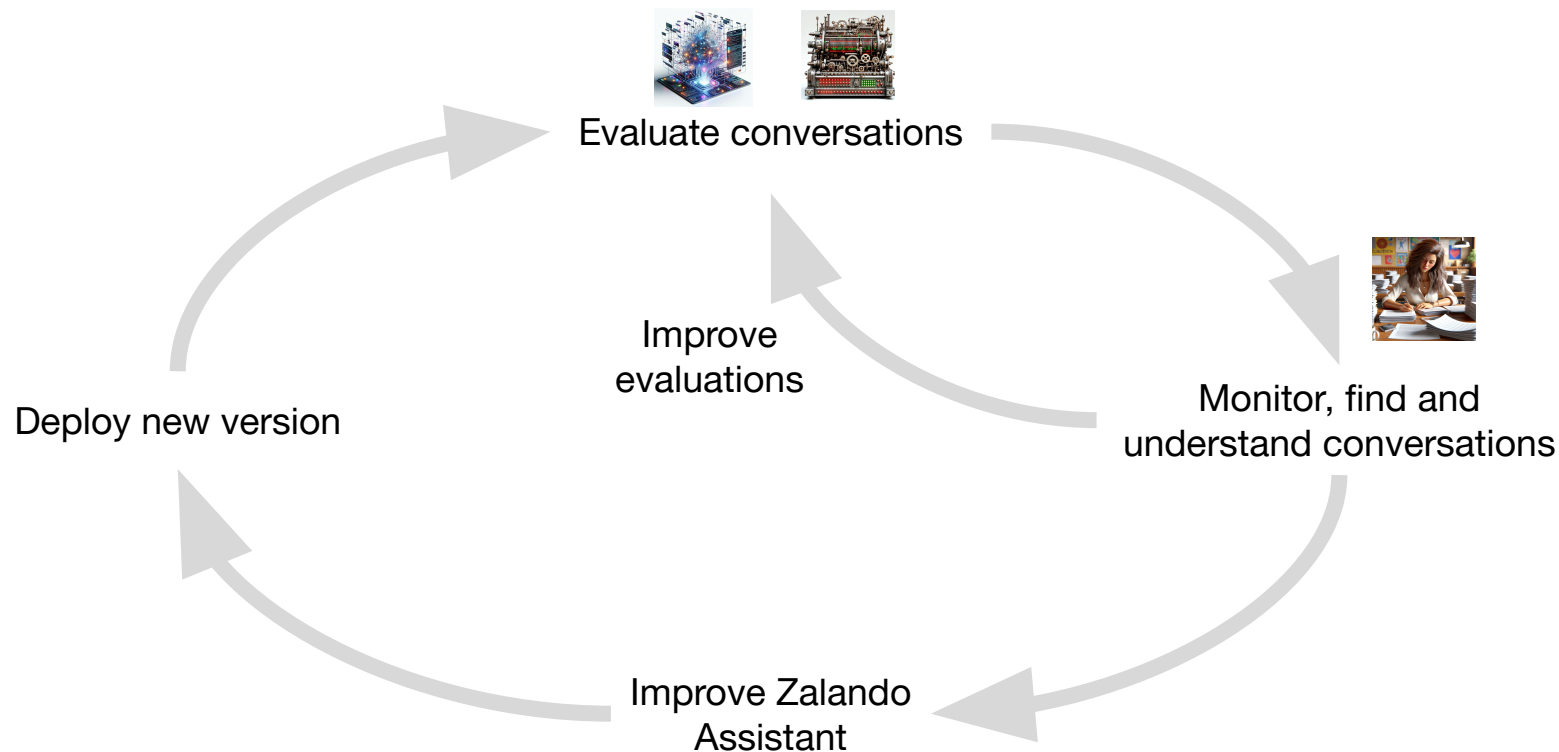
## Conversation type - Unsupported



## Understand & improve

We can better understand how customers are using the agent on average and improve the experience to meet their expectations.

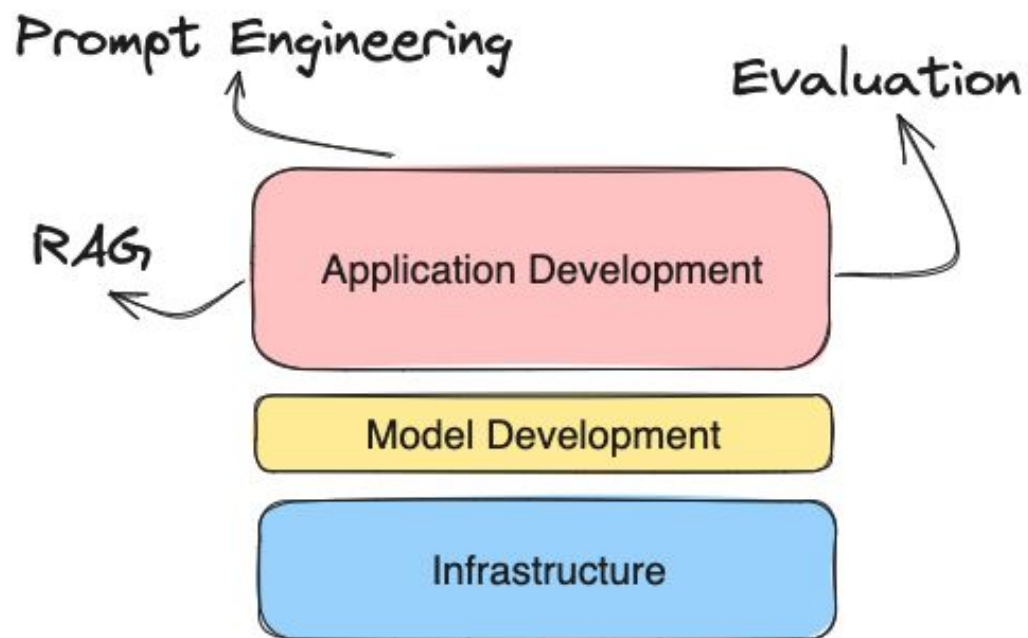
# ZA: Scalable Evaluation



# 05

## Summary





**Thank** you

