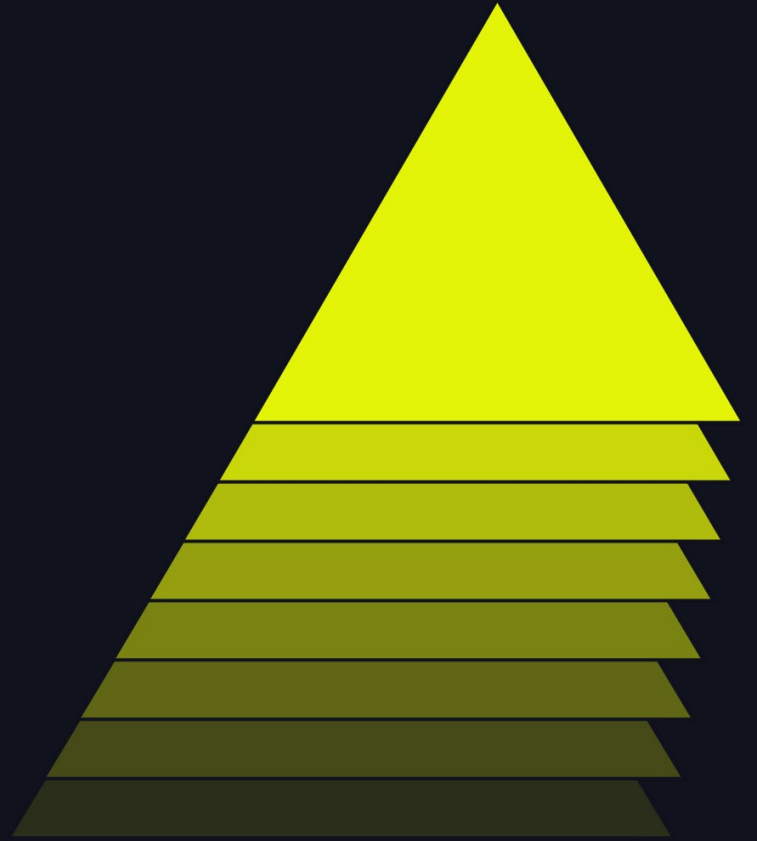


LEVERAGING LLMS FOR EMAIL PROCESSING IN CUSTOMER CENTRES

Joanna Lenczuk | CKDelta
06/12/2024



AGENDA

Leveraging LLMs for Email Processing in Customer Centres

- About CKDelta
- Overview of the Challenge
- Workflow Walkthrough
 - Email Classification
 - Email Sentiment
 - Email Summarisation
 - Email Responses
- Model Performance & Cost Comparison
- Challenges & Conclusions


ABOUT CKDELTA



ABOUT US

CKDelta builds intelligent applications designed to provide enhanced insight into business performance



 A member of CK Hutchison Holdings

Our goals

- Reduce costs by creating efficiency
- Increase revenue by driving innovation
- Enhance safety
- Improve sustainability

Industries



Utilities



Logistics



Transport



Retail



Financial services



ABOUT ME

My team and I delivered the first implementation of the Virtual Customer Agent

Data scientist at CKDelta

- Predictive modelling for utilities and logistics
- MLOps implementations
- LLMs for customer communication support



Joanna Lenczuk

OVERVIEW OF THE CHALLENGE



ABOUT THE CUSTOMER

First implementation was adopted for UK Power Networks but it's a common use-case and can be adopted for different industries and modes of contact

UK Power Networks

- The largest electricity distributor in the UK
- Maintains electricity cables and lines in London, the South-East and East of England
- Supplies energy to 19 million people

Customer Centre

- Email is the main channel of communication
- 20 agents working full time on handling requests and inquires
- B2B inbox where engineers can raise technical questions and requests

PROBLEM TO SOLVE

Customer emails review process is long and prone to errors

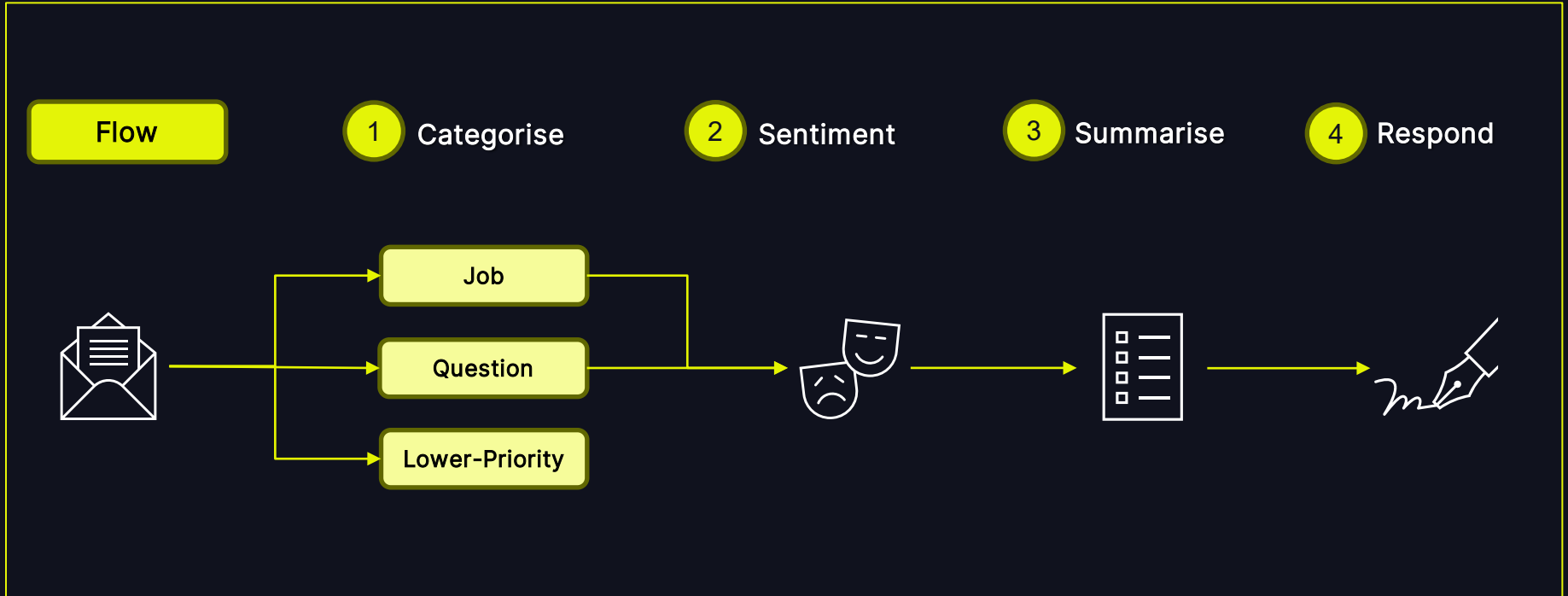


Slow response time

- Manual process of handling and distributing emails
- Low customer satisfaction
- Risk of regulatory fines
- Increasing number of emails
- Long and complex inquires

SOLUTION OVERVIEW

Improving and accelerating the process of reviewing customer emails



BENEFITS

Tangible benefits for both UK Power Networks and their customers

2.5h

saved per day

The team leader spent ~3h every day manually categorising emails. Automating email classification saves 30% of their time.

An agent spent ~3h every day reviewing emails. Reading summaries takes on avg. 17% of that time, freeing 2.5h each day.

+19%

improvement in identifying the most urgent emails

The Customer Centre team used to manually identify 79% of the most urgent emails. Our model identifies 98% of them, significantly reducing the risk of regulatory fines.

5s

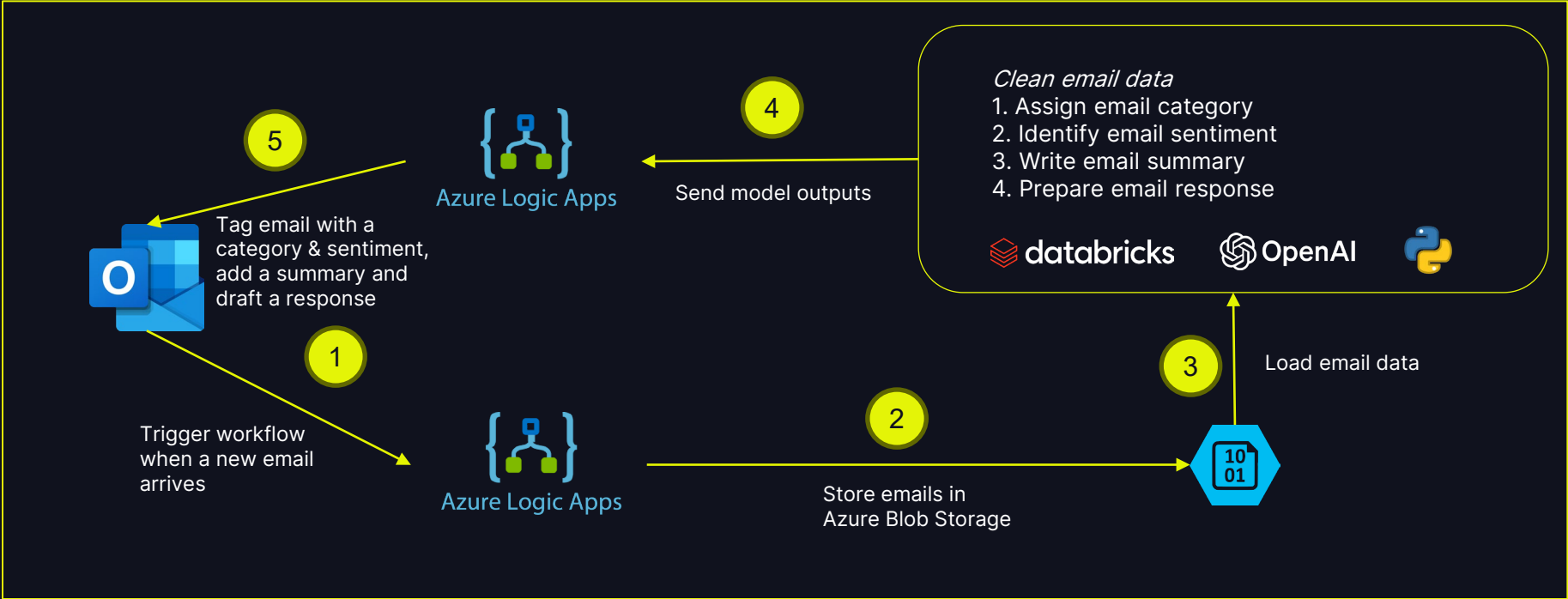
time to process an email

Providing email categories, sentiment, summaries and draft responses takes 5-10 secs, which translates to higher customer satisfaction. The manual process used to take 1.5 days.

WORKFLOW WALKTHROUGH

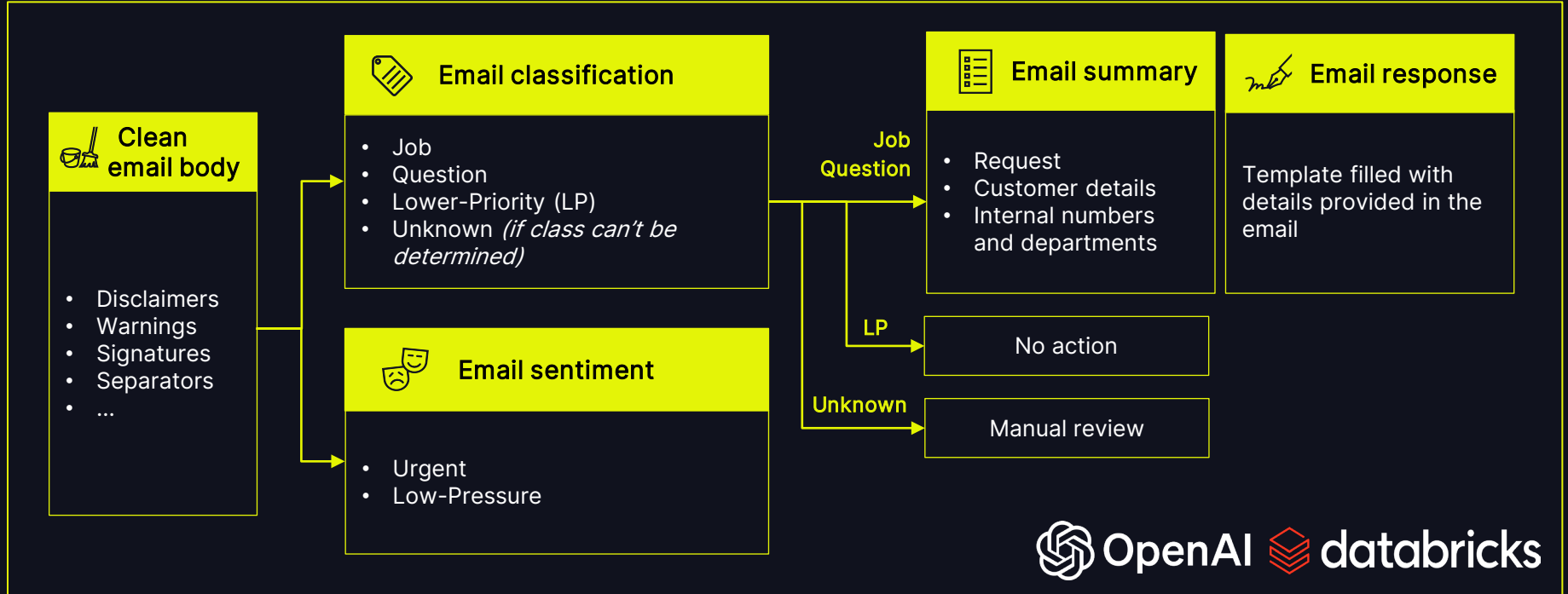
WORKFLOW OVERVIEW

Building an end-to-end solution in partnership with Databricks and Microsoft



TECHNICAL SOLUTION

All email-processing steps were implemented using Databricks OpenAI



EMAIL CLASSIFICATION

EMAIL CATEGORIES

Out of three email categories “Job” is the most important one

JOB

Hi,
Please find the G99 form attached.
Best,
Mark
Mark Jones
Engineer
mark.jones@mycompany.com
5 High Street, High Town, A29 931

QUESTION

I work for the Company Name. Because of our constructions work, we need to modify the power supplies in the area.

However, we have concerns around one supply which is already at full capacity. We could potentially need additional supplies, apart from the existing one. Please find attached the problematic location.

Could we keep the original supply and add another 4kva to the existing transformer?

Mark Doe
Specialist
Company Name

LOWER-PRIORITY (LP)

Please note that I am out of office with limited access to my emails, returning 23/01/2021.

In my absence, please forward any urgent queries to Mary Smith.

Kind regards,
Mark Jones

EVALUATION DATASET

Evaluating on 600 emails with an equal representation of each category

- 600 randomly selected emails – 200 from each category
- Complexity in retrieving labels
- Each email with a clean email body and a subject line free from post-processing alterations

Email category	Number of emails
Job	200
Question	200
LP	200

Table 1. Email categories representation

CLASSIFICATION RESULTS

LLMs improved identifying job-related emails by 19% p.

- Model based on OpenAI GPT 3.5 Turbo
- Attained overall accuracy of **75%**
- Identified **98%** of Jobs
- **~1.5%** of emails fell under the 'Unknown' category and required manual review

Human Benchmark

overall accuracy: 68%
jobs identified: 79%

	Predicted category			
	Job	Question	LP	Unknown
Job	195	3	0	1
Question	75	125	1	0
LP	32	29	131	8
Unknown	0	0	0	0

Table 1. Confusion matrix

	Precision	Recall	F1-score	Support
Job	0.65	0.98	0.78	199
Question	0.80	0.62	0.70	201
LP	0.99	0.66	0.79	200
Unknown	0.00	0.00	0.00	0.00
Accuracy	0.75	0.75	0.75	0.75
Weighted avg.	0.81	0.75	0.76	600

Table 2. Classification error metrics

EMAIL SENTIMENT



EMAIL SENTIMENT CLASSIFICATION

The priority is to identify and address urgent emails

URGENT

negative sentiment

- Customer clearly states the request is urgent
- Customer seems to be impatient
- Emails have been back and forth without a clear resolution

LOW-PRESSURE

neutral sentiment

- Requests with formal tones
- BAU approach
- An email chain can be long, but there's an agent assigned to resolve the query

EMAIL SENTIMENT CLASSIFICATION RESULTS

Using LLMs enabled identifying 80% of urgent emails

- Labels assigned manually after consulting the SMEs
- Model based on OpenAI GPT 3.5 Turbo
- Attained an overall accuracy of **93%**
- Identified **80%** of all urgent emails

		Metrics		
		Precision	Recall	F1-score
Sentiment class	Urgent	1.0	0.8	0.89
	Low-Pressure	0.9	1.0	0.95
	Accuracy	0.93	0.93	0.93

Table 1. Error metrics for sentiment classification

EMAIL SUMMARISATION

SUMMARISATION RESULTS

Summaries reduced the time needed to review emails by up to 90% p.

- Model based on OpenAI GPT 3.5 Turbo
- Attained a semantic textual overlap (similarity of the meaning, regardless of the phrasing) of **83%**
- The reading time of long email chains reduced by **82-90% p.**
- SMEs confirmed the reliability and completeness of summaries after manual review

Metric name	Score
Semantic Textual Overlap	0.83
Precision (Information Retrieval Metric)	0.68
Key-phrase Overlap	0.56

Table 1. Error metrics for summarisation

	Word count	Reading time	Word count of summary	Reading time of summary
Job	300	90s	52	15.6s
Question	587	176.1s	54	16.2s

Table 2. Reading-time metrics for summarisation

EMAIL RESPONSES



TEMPLATE-BASED EMAIL RESPONSES

Determining if customer's query matches a scenario covered by a template

Scenario Matching

1. Inquiring about the **fuse sizes** of an existing connection
2. Inquiring about the **available capacity of a connection**
3. Inquiring about the **available capacity of a network**



Yes

Response Generation

- Using a **provided template** to generate a response
- Modifying the template based on **details in customer's email**
- **Addressing customer's inquiry** and guiding them through the steps

EMAIL RESPONSES RESULTS

The initial results highlight the complexity of each template scenario

Learnings

- ✓ More labelled examples are needed
- ✓ The subjectivity of generated responses is a significant challenge
- ✓ Focusing on a narrow use-case is the first step to generating reliable responses

100% of emails matching the template scenario were identified and responded to.

25% of emails not related to any template received an unnecessary response.

MODEL PERFORMANCE & COST COMPARISON

DISTILBERT RESULTS

DistilBert shows promising results and can reduce costs in the future

	DistilBert	OpenAI
Job (Recall)	0.80	0.98
Question (F1)	0.71	0.70
LP (Precision)	0.9	0.99
Overall (Accuracy)	0.79	0.75

Table 1. Error metrics for DistilBert vs. OpenAI

- Model evaluated on the same 600 emails used for the original task
- Finetuning DistilBert on a batch of different 600, balanced-class emails
- The model is much smaller and can be run on a small GPU cluster or with CPU
- Promising results for optimizing the costs in the future

META LLAMA V2 RESULTS

Meta LLaMa V2 performs significantly worse than OpenAI GPT 3.5 Turbo

	13B Meta LLaMa V2	70B Meta LLaMa V2	OpenAI
Job (Recall)	0.98	0.81	0.98
Question (F1)	0.19	0.46	0.70
LP (Precision)	0.95	0.99	0.99
Overall (Accuracy)	0.55	0.67	0.75

Table 1. Error metrics for Meta Llama V2 vs. OpenAI

- Model evaluated on the same 600 emails used for the original task
- Two versions of Llama V2 model tested: 13B and 70B
- Both versions have significantly worse results than Open AI GPT 3.5 Turbo, especially regarding queries

ASSUMPTIONS FOR COSTS COMPARISON

The costs comparison assumes 3 million tokens are used every hour

Costs based on the average usage seen on sample runs:

3 million tokens per hour

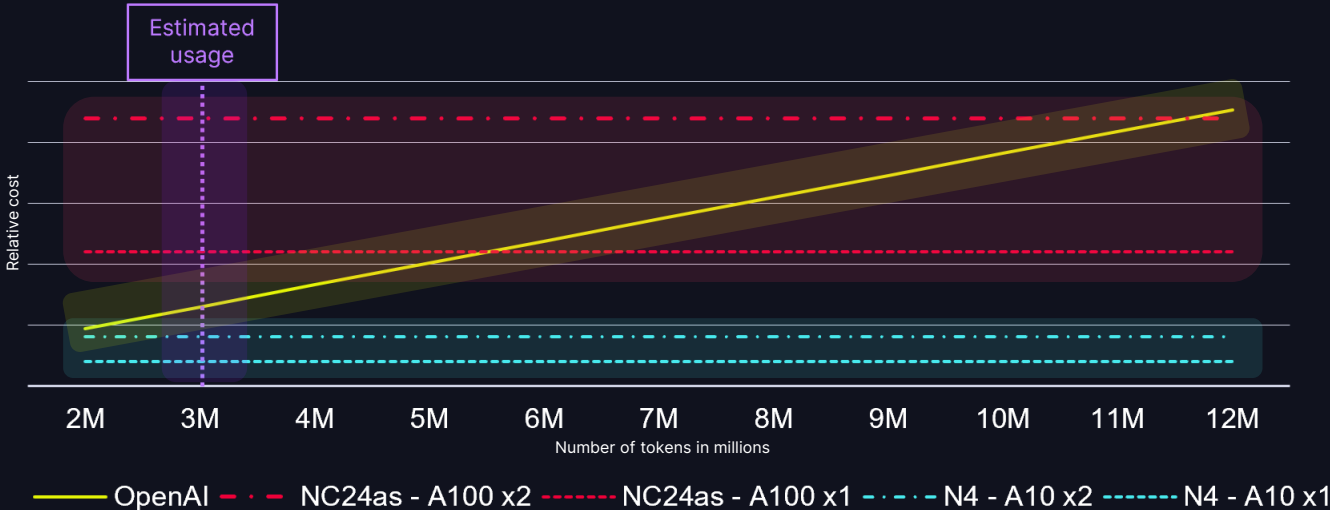


Assumption	Value
Hours online	10
Days working	6
Number of weeks	4

Table 1. Assumptions for costs comparison

COSTS COMPARISON

There is potential for reducing operating costs of using LLMs in the long run



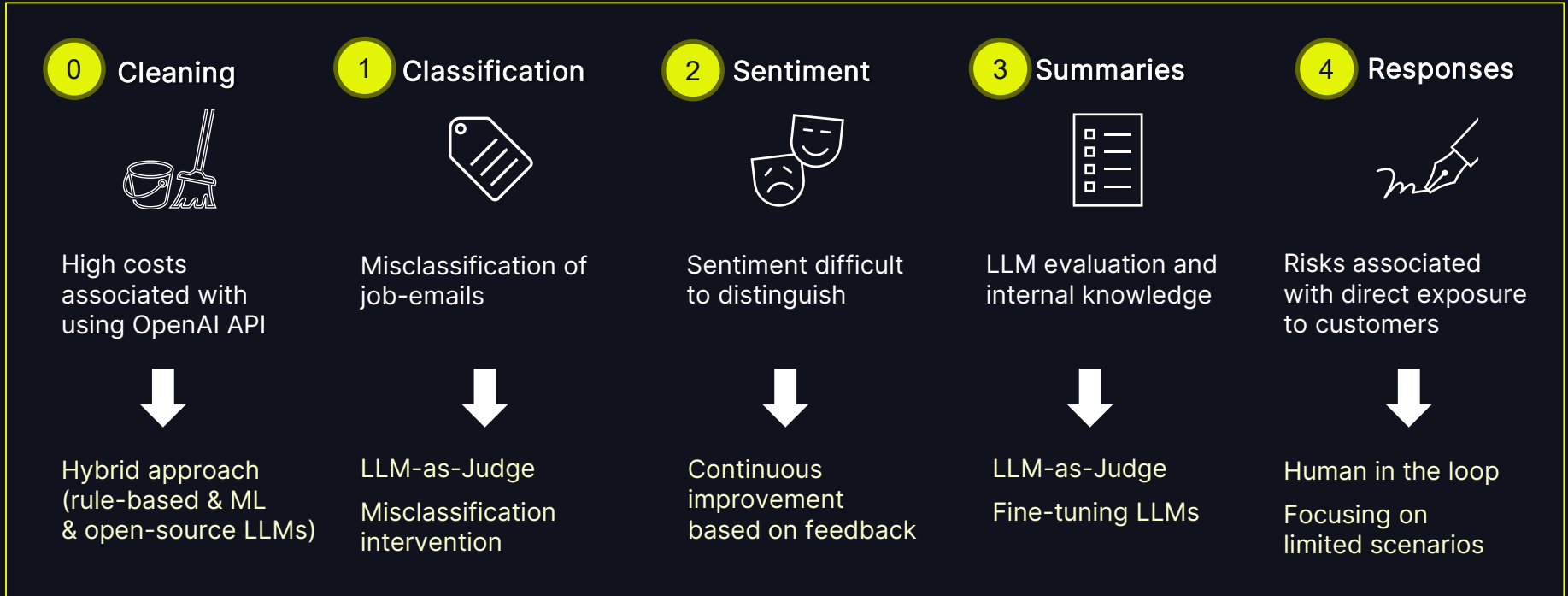
- ✓ small open-source models
- ✓ serverless endpoints
- ✓ using LLMs for the most complex tasks



CONCLUSIONS & CHALLENGES

CHALLENGES

The biggest challenge is reducing costs while maintaining high performance



CONCLUSIONS

The solution can be easily implemented for different industries and channels of customer communication

- Using LLMs improves the identification of the most urgent emails by 19% p.
- It saves 2.5 hours of the team's time every day, allowing them to focus on the most complex queries and personalised support.
- Automated email processing allows to reply to the customer in 5 seconds instead of 1.5 days.
- Using OpenAI and tuning prompts enables fast iterations, crucial at the early stage of development.
- There's potential for reducing operating costs by using open-source models on small machines and limiting use-cases handled by LLMs.

DATA+AI SUMMIT

