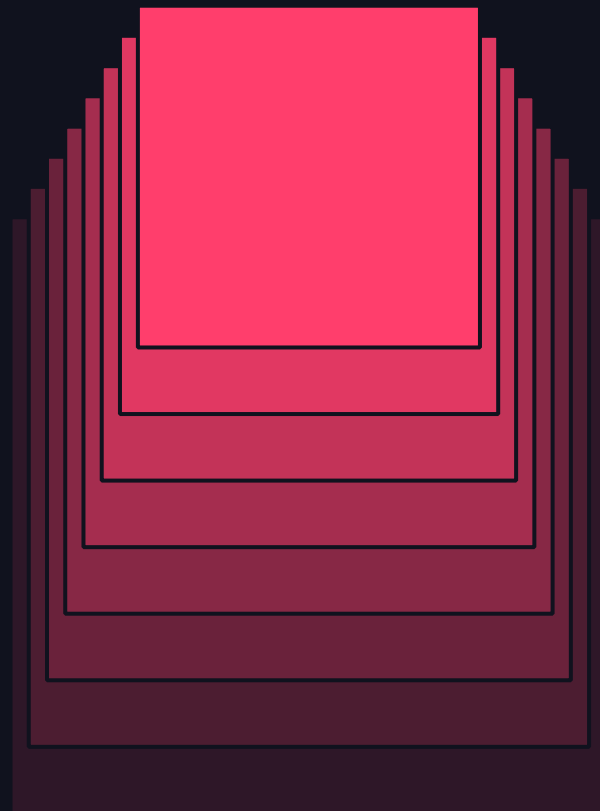


# LLM Evaluation: Auditing Fine-Tuned LLMs for Guaranteed Output Quality



**Loic Pauletto, Ph.D.**  
*Data Scientist, Miraki*



**Pierre Lourdelet**  
*Data Scientist, Miraki*

# Who are we ?



- Empowering clients since 2012
- World leading marketplace platform
- +450 B2B and B2C clients
- +200k shops
- 8.6B\$ GMV in 2023

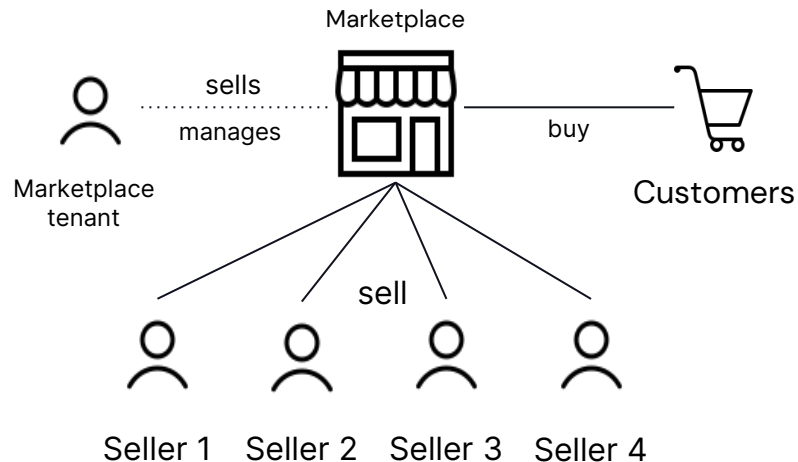


## The Data Science Team : Miradoge

- Develop and industrialize AI features
- Facilitation and improvement of Client Operations
- Platform Security and Fraud Prevention
- Revenue Growth Through Personalization Solutions

# What is a marketplace ?

A marketplace: a platform where multiple providers sell products and services



Some Mirakl-powered marketplaces



# Main Topics

1. Catalog Data
2. The onboarding struggle
3. How to auto-structure information ?
4. Challenges
5. Ensure Quality
6. Engineering integration
7. Key takeaways

# *Context:* Catalog- based environment



# What is a catalog ?

## In a marketplace context



**Title:** iPhone 15 6,1" 5G  
128GB Black  
**Description:** This is an iPhone ...  
**Brand :** Apple  
**Screen:** 6.1" OLED  
**Color:** Black  
**Processor:** A16 Bionic Chip  
**Storage Capacity:** 128GB  
**DAS Regulations:** Compliant  
**Weight:** 171g



ワンプラス10プロ  
5G256GBグリーンRedmi  
プロセッサと120Hz流  
体ディスプレイで、強  
力なパフォーマンスと  
超滑らかな応答性を  
解き放ちます。ワ  
ープチャージ80T  
で、15分で1日の  
パワーを充電。

Brand	Size	Color	Condition
Apple	128GB	Black	New
Redmi	256GB	Blue	New

# Mirakl catalog in numbers :

**310 M  
products**

**+650 k  
categories**

**+450  
marketplaces**

**+200 k  
shops**



# The data

## An example



**Title:** iPhone 15 6,1" 5G 128 GB Black

**Description:** This is an iPhone ...

**Brand :** Apple

**Screen:** 6.1" OLED

**Color:** Black

**Processor:** A16 Bionic Chip

**Storage Capacity:** 128GB

**DAS Regulations:** Compliant

**Weight:** 171g



# The onboarding struggle



# Challenging environment

## Input



```
ID123,ProduktDetails,€Preis,Menge_Bestand,Info
101,Drahtlose Maus,29,99,150,Eine hochpräzise drahtlose Maus <b>mit ergonomischem Design</b> und langer
Batterielaufzeit.
102,Laufschuhe,49,99,75 pcs,Bequeme und langlebige Laufschuhe <i>für alle Geländearten</i>.
```

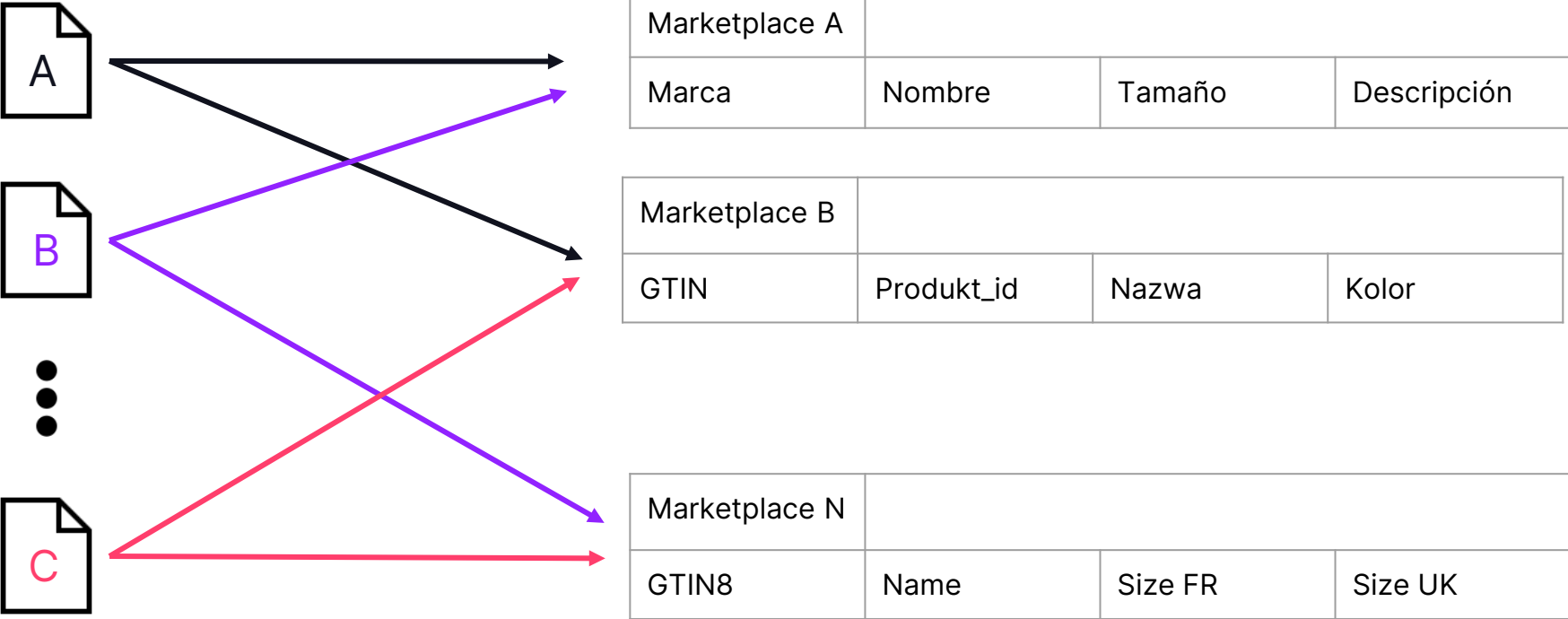


```
ID_#, $Item, €Cost, Stock#, Detail_Info
101,WrLsMse,29.99,,<p>A high-precision wireless mouse with <b>ergonomic design</b> and long battery life.</p>
102,RnShs,"49.99USD",75 pcs,<p>Comfortable and durable running shoes designed for <i>all terrains</i>.</p>
103,Steel H2O Btl,19,99,"200",<p>Insulated water bottle that keeps drinks <b>cold for 24 hours</b> or <b>hot for 12
hours</b>.</p>
A104,,None,300 pcs,<p>Soft and breathable cotton t-shirt available in multiple colors and sizes.</p>
```

ID	PName	Price	StockQty	Desc
101	Wireless Mse	29.99	None	<p>High-precision mse <b>ergo design</b> long batt.</p>
102	RunShoes	49.99USD	75	<p>Comfy & durable shoes <i>4 all terrains</i>.</p>
103	Stl Wtr Btl	19,99	200 pcs	<p>Water btl keeps <b>cold 24hrs</b> <b>hot 12hrs</b>.</p>
A104		None	three hundred	Soft cotton t-shirt multiple clr sizes.



# Complex Extraction



# Missing information

## An example

### Specifications



Apple - iPhone 14 Plus (Unlocked)

★★★★★ 4.9 (158 Reviews)

#### General

Product Name	iPhone 14 Plus (Unlocked)
Brand	Apple
Model Number	
Series ⓘ	Unlocked 14/14 Plus
Year of Release	
Color	
Color Category	Black
Storage	



# The missing attributes hassle

# How to auto-structure information ?



# Data we have

Description
Galaxy S23 Ultra - 512GB, Phantom Black, Experience next-level photography with Galaxy S23 Ultra's advanced 108MP camera and 100X Space Zoom.
<pre>&lt;body&gt; &lt;div class="product-container"&gt; &lt;div class="product-title"&gt;Phone 14 Pro - 128 Go, Violet&lt;/div&gt; &lt;div class="product-description"&gt;Plongez dans une expérience utilisateur encore plus riche &lt;/div&gt; &lt;div class="product-condition"&gt;Etat : Neuf&lt;/div&gt; &lt;/div&gt; &lt;/body&gt; &lt;/html&gt;</pre>
Google Pixel 7 - 128 GB, Obsidian, Mantente a la vanguardia con la traducción en tiempo real del Google Pixel 7 y el chip Tensor G2 para operaciones de IA mejoradas. Su Adaptive Battery aprende el uso de tus apps y amplía la eficiencia energética. Nuevo
ワンプラス10プロ5G256GBグリーン。Snapdragon 8 Gen 1 プロセッサと120Hz 流体ディスプレイで、強力なパフォーマンスと超滑らかな応答性を解き放ちます。ワープチャージ 80Tで、15分で1日のパワーを充電。



# Our situation

Data we have

Data we need

Description	Brand	Size	Color	Condition
Galaxy S23 Ultra - 512GB, Phantom Black...				
<body> <div class="product-cont...				
Google Pixel 7 - 128 GB, Mantente...				
ワンピース10プロ5G256GBグリ...				





# The process

An LLM based Mirakl product



# Output

**Galaxy S23 Ultra - 512GB, Phantom Black**, Experience next-level photography with Galaxy S23 Ultra's advanced 108MP camera and 100X Space Zoom.



```
{  
  "Brand": "Samsung",  
  "Size" : "512GB",  
  "Color" : "Black"  
}
```

# Expected result

Seller Data

Extracted Data

Description	Brand	Size	Color	Condition
Galaxy S23 Ultra - 512GB, Phantom Black...	Samsung	512GB	Black	
<body> <div class="product-cont...	Apple	128GB	Purple	New
Google Pixel 7 - 128 GB, Mantente...	Google	128GB	Black	New
ワンピース10プロ5G256GBグ...	One Plus	256GB	Green	New



# Challenges



# Engineering constraints

## 1. Cost management

*High volume of traffic, need to find the most efficient solution.*

## 2. Fast response time

*Ensure quick response times for a seamless user experience, requiring the model to generate results rapidly and handle multiple requests efficiently.*

# Evaluation: A Major Challenge

We do not have any ground truth

# Ideas to acquire ground truth

1. Use data we already have



Need to be qualified, sparse

2. Manually create our dataset



Not humanly possible

3. Mix of human & GPT-4



How to control quality ?



# The process

An LLM based Mirakl product



Detect hallucinations as soon as possible  
to avoid snowball effect



# Challenges

What can happen ?

Inaccuracy and Hallucination



Consequences

Danderositv


Latency 1054ms · 138 tokens

Enter user message...

User



Add

Run  ↵

# Conversion

## An example of mistake

- Example of a pants product
- Input :

Elevate your wardrobe with our sleek and stylish Pants. These trousers offer a perfect blend of comfort and elegance, featuring a tailored fit, premium fabric, and versatile design suitable for any occasion. Experience the ideal combination of form and function with deep pockets and easy maintenance. Perfect for the modern man seeking both style and practicality. Available in size 44 (Italian).

IT	38	40	42	44	46	48
UK	6	8	10	12	14	16
US	0-2	4	6	8	10	12
FR	34	36	38	40	42	44



Size
40

# Ensure Quality



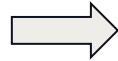
# Two missions

How do we “Ensure quality” ?

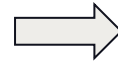
1. Create a dataset
2. Evaluate fine-tuned models

# First step: Prompt engineering

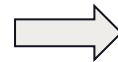
Ask to explain



Say "I do not know"



Give examples



```
23: Prompt Python    
```

```
.....  
...  
For each answer you provide, you must explain your choice.  
...  
If there is no information provided about an attribute, respond with  
"None".  
Never add information that is not present in the seller's  
description.  
...  
Example: If the product is a t-shirt and it is not mentioned whether  
it is a top or a bottom, you can infer that it is a top.  
...  
.....
```

# Detecting hallucinations

- Use LLM as a judge
- Set metrics

# LLM as a judge

USER

You are an experienced marketplace operator and you are checking the work of one of your trainee. He was supposed to extract information from this text :

Google Pixel 7 - 128 GB, Obsidian, Mantente a la vanguardia con la traducción en tiempo real del Google Pixel 7 y el chip Tensor G2 para operaciones de IA mejoradas. Su Adaptive Battery aprende el uso de tus apps y amplía la eficiencia energética.

He extracted this : {Brand : Google , Size:128GB,Color :Green, Condition : New}





# LLM as a judge

USER

You are an experienced marketplace operator and you are checking the work of one of your trainee. He was supposed to extract information from this text :

Google Pixel 7 - 128 GB, Obsidian, Mantente a la vanguardia con la traducción en tiempo real del Google Pixel 7 y el chip Tensor G2 para operaciones de IA mejoradas. Su Adaptive Battery aprende el uso de tus apps y amplía la eficiencia energética.

He extracted this : {Brand : Google , Size:128GB,Color :Green, Condition : New}

# LLM as a judge

**USER**

You are an experienced marketplace operator and you are checking the work of one of your trainee. He was supposed to extract information from this text :

Google Pixel 7 - 128 GB, Obsidian, Mantente a la vanguardia con la traducción en tiempo real del Google Pixel 7 y el chip Tensor G2 para operaciones de IA mejoradas. Su Adaptive Battery aprende el uso de tus apps y amplía la eficiencia energética.

He extracted this : {Brand : Google , Size:128GB,Color :Green, Condition : New}

**ASSISTANT**

Of course, here is the correction :








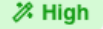


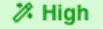


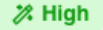


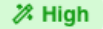


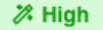


```
{
  "Color":
  {
    "Wrong value": "Green",
    "New value": "Black"
  },
  "Condition":
  {
    "Wrong value": "New",
    "New value": "Not mentioned"
  }
}
```

# ChainPoll

## A Galileo Product

- LLM as a judge
- Correctness for open-domain errors
- Context Adherence for closed-domain errors
- CoT explanation

# Galileo

Input	Output	 RAG Quality	 Output Quality	 Rating Summary	 Input Quality
		Context Adherence 	Correctness 	Like/Dislike	Prompt Perplexity 
You are given a list of colum...	```json {"Battery included": { ...	 High	 Low		5.666
You are given a list of colum...	{ "Brand": {"value":"Skechers...	 High	 High		8.212
You are given a list of colum...	```json {"Additional Features...	 High	 High		7.598
You are given a list of colum...	```json {"Accessories (option...	 High	 High		6.403
You are given a list of colum...	{ "Accroche permanente": {"v...	 High	 High		6.745

# Galileo

## Guardrail Metrics

- **Toxicity**  
*Measures the presence of offensive and harmful languages*
- **Personal identifiable information**  
*Tracks email, phone number, credit card number ...*
- **Tone**  
*Classifies the sentiment in the response*
- **Sexism**  
*Measures how 'sexist' an output can be perceived*
- **Custom metrics**  
*Metrics we can create*

# Dataset Creation

Here is the CARTER jacket from MAX MARA, perfect for the Fall Winter 2023/24 season. Made of 100 percent cashmere, this beige jacket offers comfort and luxury without fur. Belonging to the outerwear category, it is ideal for a stylish and modern woman. **Size (IT) 42**, this garment features a solid-color pattern for refined style.



```
{  
  "Size (FR)" : "42"  
}
```

Low  
Correctness

# Dataset Creation

Here are the new **black Air Max**, perfect for a modern and urban style. Featuring the iconic visible Air unit, these shoes offer exceptional comfort and cushioning. Designed for durability, they combine high-quality materials with a sleek and versatile design. Available in **US size 10**, they are ideal for completing any outfit with a touch of sporty sophistication.



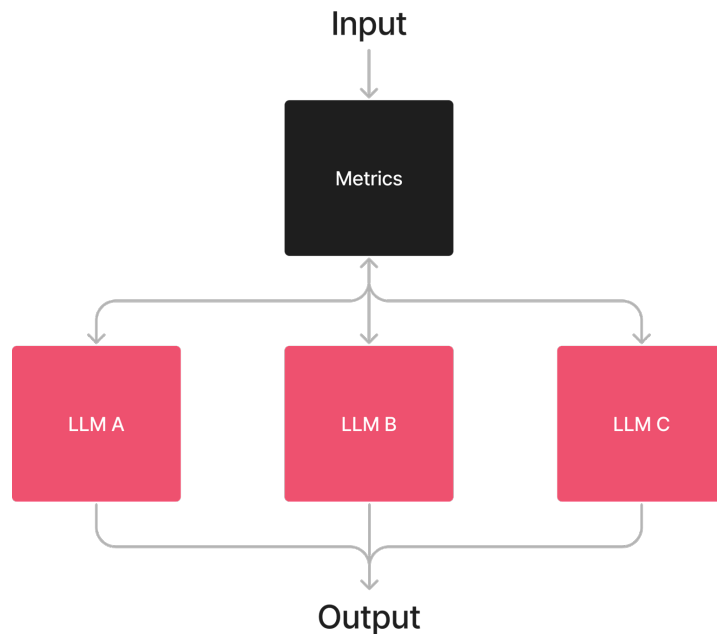
```
{
  "Brand" : "Nike",
  "Size" : "10",
  "Color" : "Black"
}
```

High Context  
Adherence

# Next step : Layering

## Future Opportunities

- Metrics will enable precise selection of the most suitable LLM
- The best-fitting LLM will be chosen based on specific input



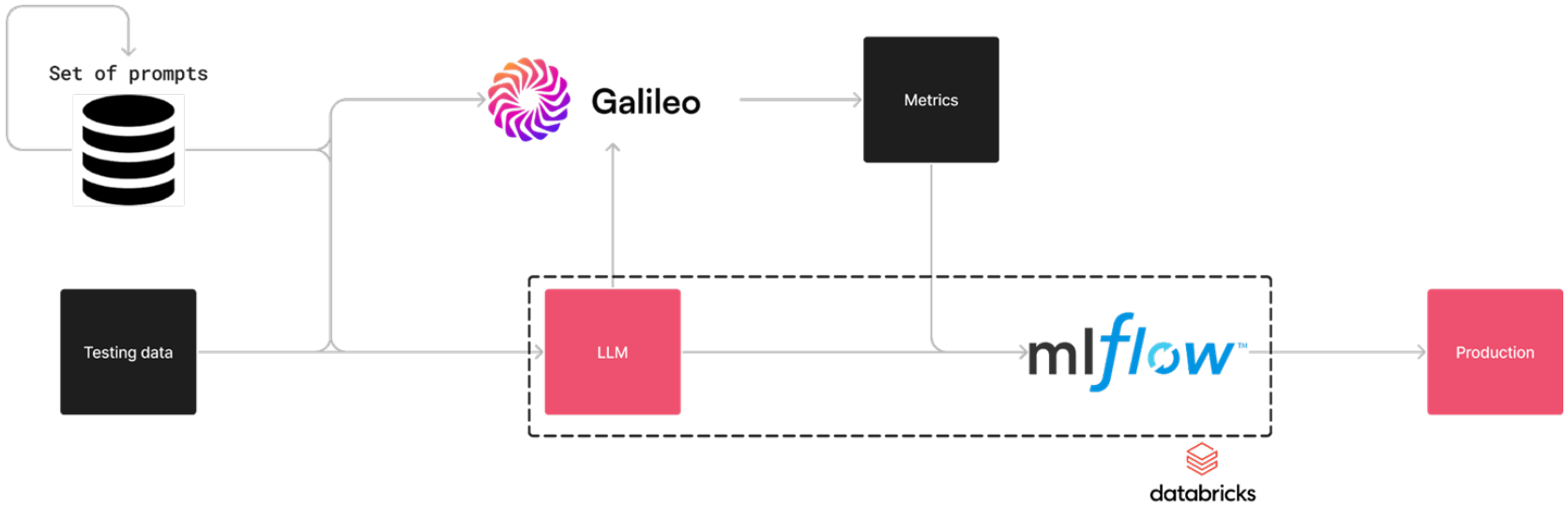


# Engineering integration



# Engineering side

- The evaluation pipeline



# Engineering side

- MLFlow interface



# Key Takeaways



# Key Takeaways

What we hope you will remember from this talk

- LLMs can be game-changers in crucial use cases
- LLMs are inherently prone to hallucinations
- We need to invest a lot of time and resources to ensure quality

**THANK  
YOU**

