DATA+AI SUMMIT
BY databricks

IFC'S MACHINE LEARNING ESG ANALYST PLATFORM

Delivering domain specific LLMs with GPU serving:
Case of IFC MALENA

Blaise Sandwidi, Ph.D. & Jonathan Lorentz
International Finance Corporation

©2024 Databricks Inc. — All rights reserved

Official Use Only

©2024 Databricks Inc. — All rights reserved

Official Use Only

$44 billion

IBRD

IDA

IFC

MIGA

ICSID

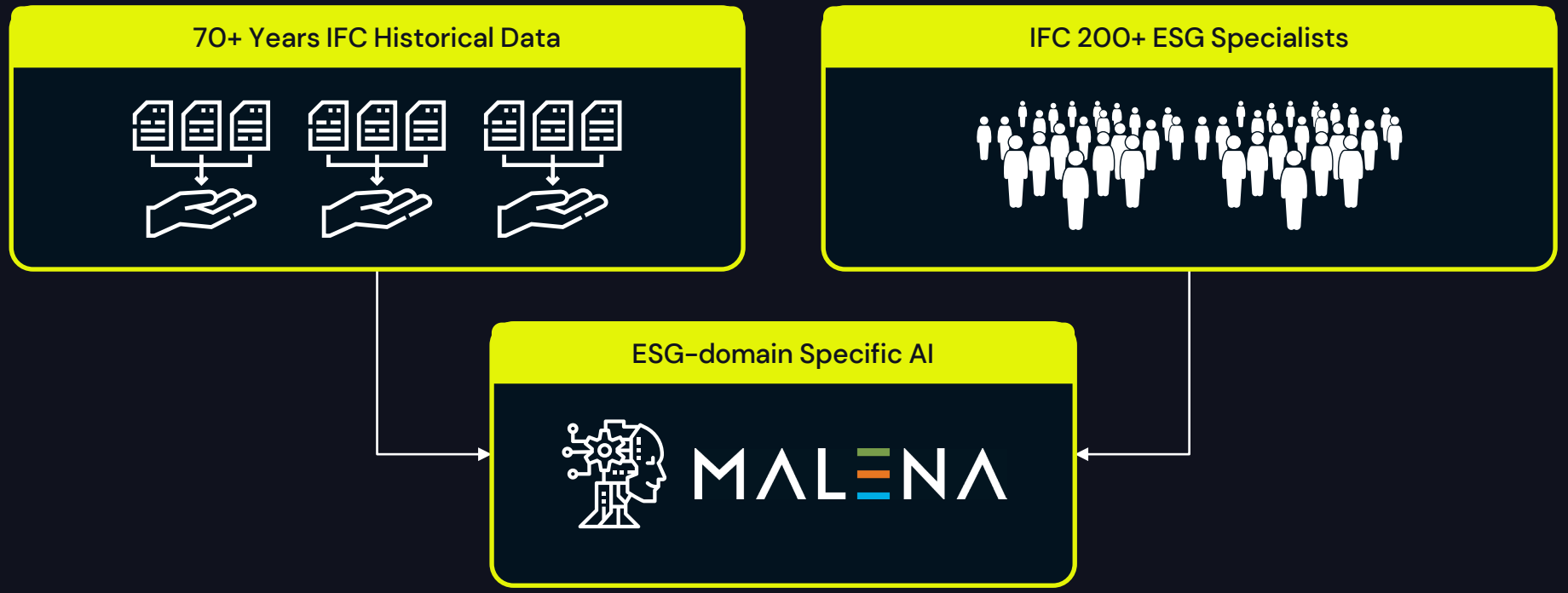
©2024 Databricks Inc. — All rights reserved

Official Use Only

©2024 Databricks Inc. — All rights reserved

Official Use Only

DATA + ARTIFICIAL INTELLIGENCE FOR DEVELOPMENT IMPACT



www.malena.ifc.org

©2024 Databricks Inc. — All rights reserved

Official Use Only

INTRODUCTION TO MALENA

©2024 Databricks Inc. — All rights reserved

Official Use Only

SPEED & ACCURACY

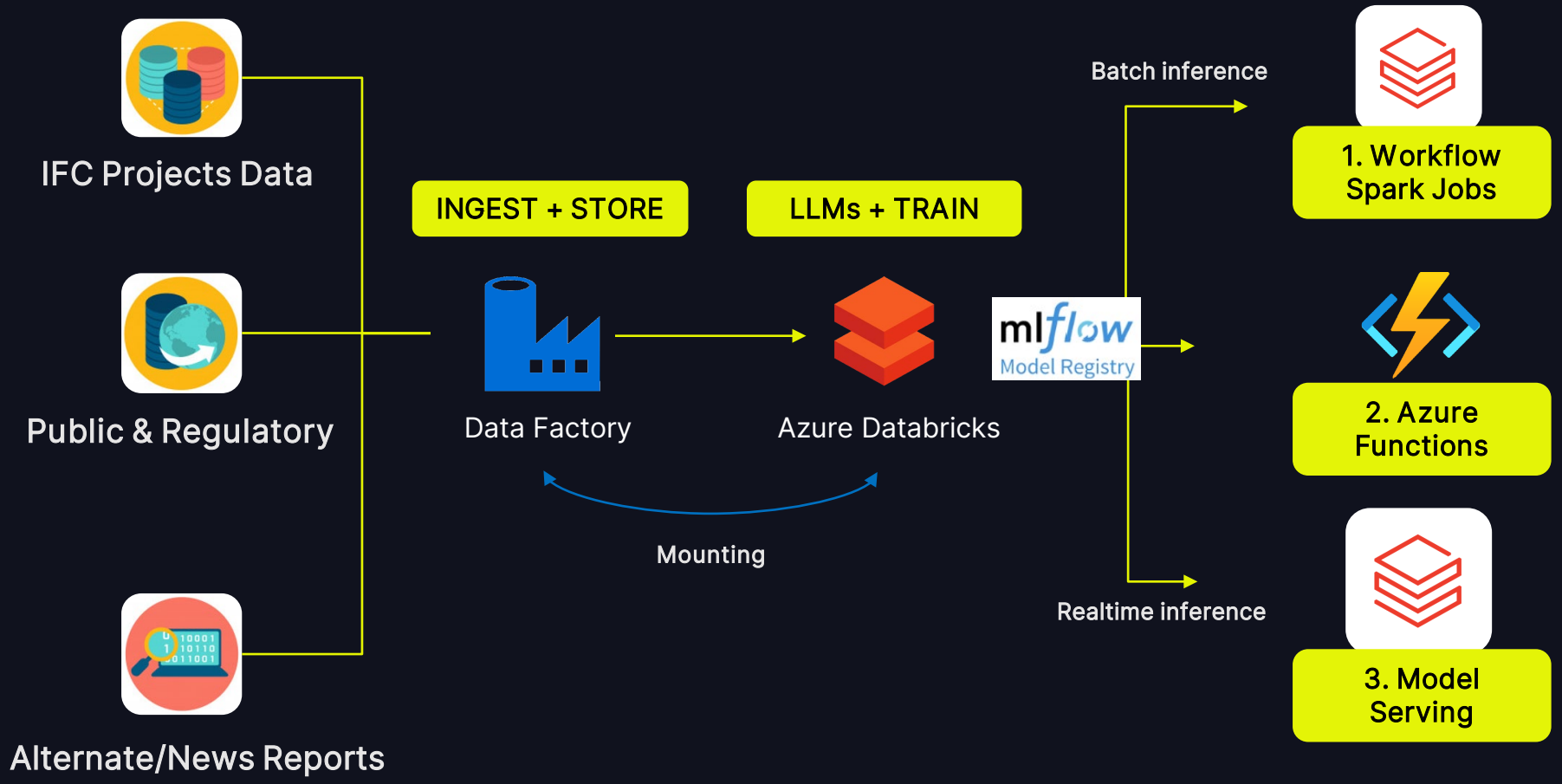
19,000 sentences/min

91% accurate

©2024 Databricks Inc. — All rights reserved

Official Use Only

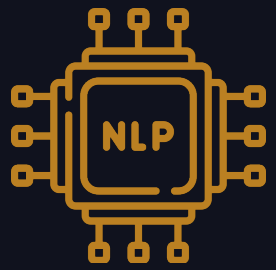
INFRASTRUCTURE SET UP



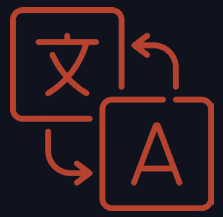
©2024 Databricks Inc. — All rights reserved

Official Use Only

ESG-DOMAIN AI FOR EMERGING MARKETS



ESG Sentiment Analysis (**91%** acc)



Translation

Named Entity Recognition

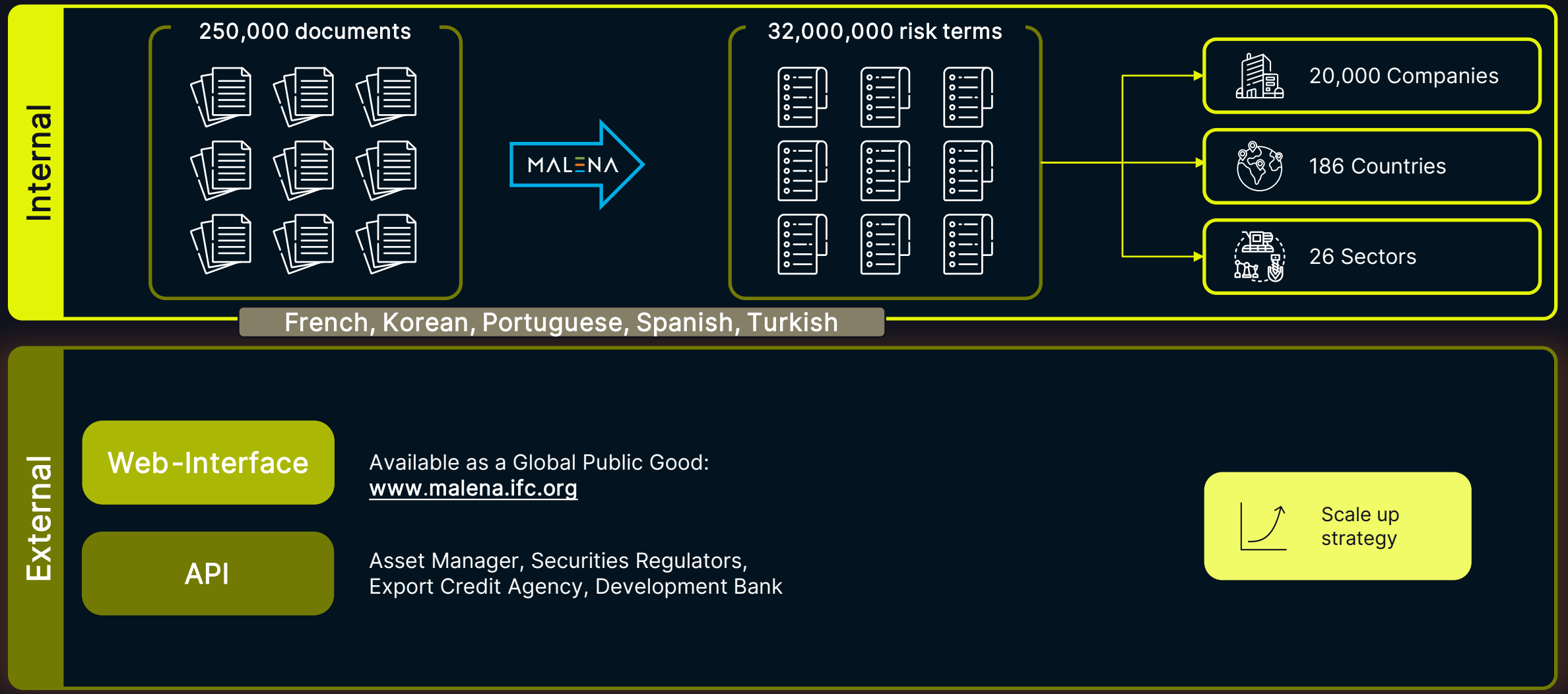
RAG + Question & Answering (**85%** acc)

ESG and **Climate** text generation

©2024 Databricks Inc. — All rights reserved

Official Use Only

MALENA AT A GLANCE



©2024 Databricks Inc. — All rights reserved

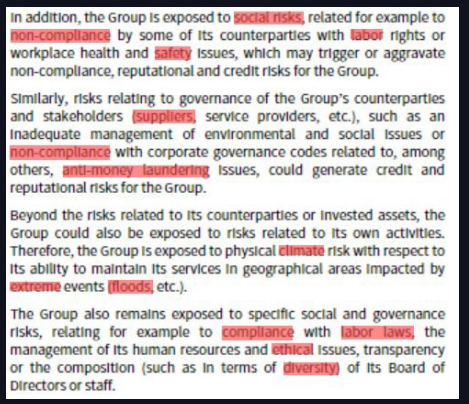
Official Use Only

THE MALENA B2B APP WORKFLOW

Job Processing:

1. Translation (IO)
2. OCR and pre-processing (CPU)
3. Term Detection (IO*)
4. Sentiment Analysis (CPU)
5. Aggregation/post-processing (CPU)

Predicting Sentiment from complex documents

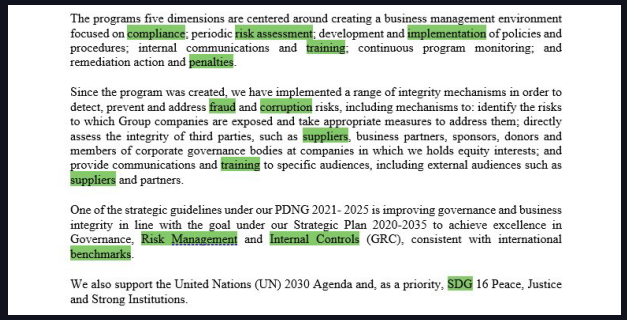


In addition, the Group is exposed to social risks, related for example to non-compliance by some of its counterparties with labor rights or workplace health and safety issues, which may trigger or aggravate non-compliance, reputational and credit risks for the Group.

Similarly, risks relating to governance of the Group's counterparties and stakeholders (suppliers, service providers, etc.), such as an inadequate management of environmental and social issues or non-compliance with corporate governance codes related to, among others, anti-money laundering issues, could generate credit and reputational risks for the Group.

Beyond the risks related to its counterparties or invested assets, the Group could also be exposed to risks related to its own activities. Therefore, the Group is exposed to physical climate risk with respect to its ability to maintain its services in geographical areas impacted by extreme events (floods, etc.).

The Group also remains exposed to specific social and governance risks, relating for example to compliance with labor laws, the management of its human resources and ethical issues, transparency or the composition (such as in terms of diversity) of its Board of Directors or staff.



The programs five dimensions are centered around creating a business management environment focused on compliance; periodic risk assessment; development and implementation of policies and procedures; internal communications and training; continuous program monitoring; and remediation action and penalties.

Since the program was created, we have implemented a range of integrity mechanisms in order to detect, prevent and address fraud and corruption risks, including mechanisms to: identify the risks to which Group companies are exposed and take appropriate measures to address them; directly assess the integrity of third parties, such as suppliers, business partners, sponsors, donors and members of corporate governance bodies at companies in which we holds equity interests; and provide communications and training to specific audiences, including external audiences such as suppliers and partners.

One of the strategic guidelines under our PDNG 2021- 2025 is improving governance and business integrity in line with the goal under our Strategic Plan 2020-2035 to achieve excellence in Governance, Risk Management and Internal Controls (GRC), consistent with international benchmarks.

We also support the United Nations (UN) 2030 Agenda and, as a priority, SDG 16 Peace, Justice and Strong Institutions.

©2024 Databricks Inc. — All rights reserved

Official Use Only

GPU SERVING SCOPING

In Scope testing scenarios

1. Load Testing
2. Stress Testing
3. Concurrency Testing
4. Latency Testing
5. Scalability Testing

Baseline scenarios

- 25 Sentences/Request
- 90 Requests in 90 Seconds

- 10-20 Pages/Request
- 30 Requests in 30 Seconds

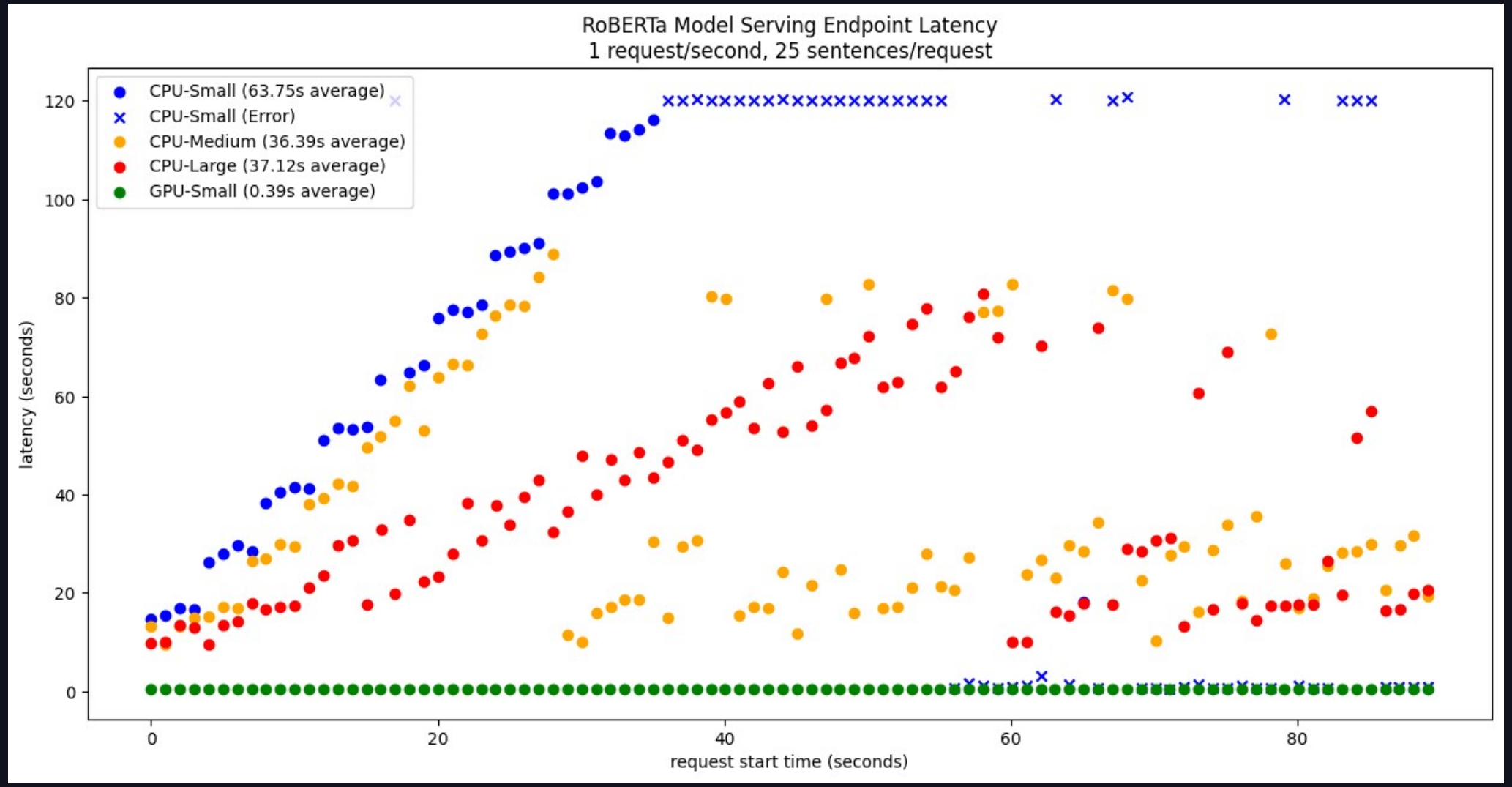
Out of scope for testing scenarios

1. Security Testing (encryption, authentication)
2. Network Testing
3. Error Rate Testing

©2024 Databricks Inc. — All rights reserved

Official Use Only

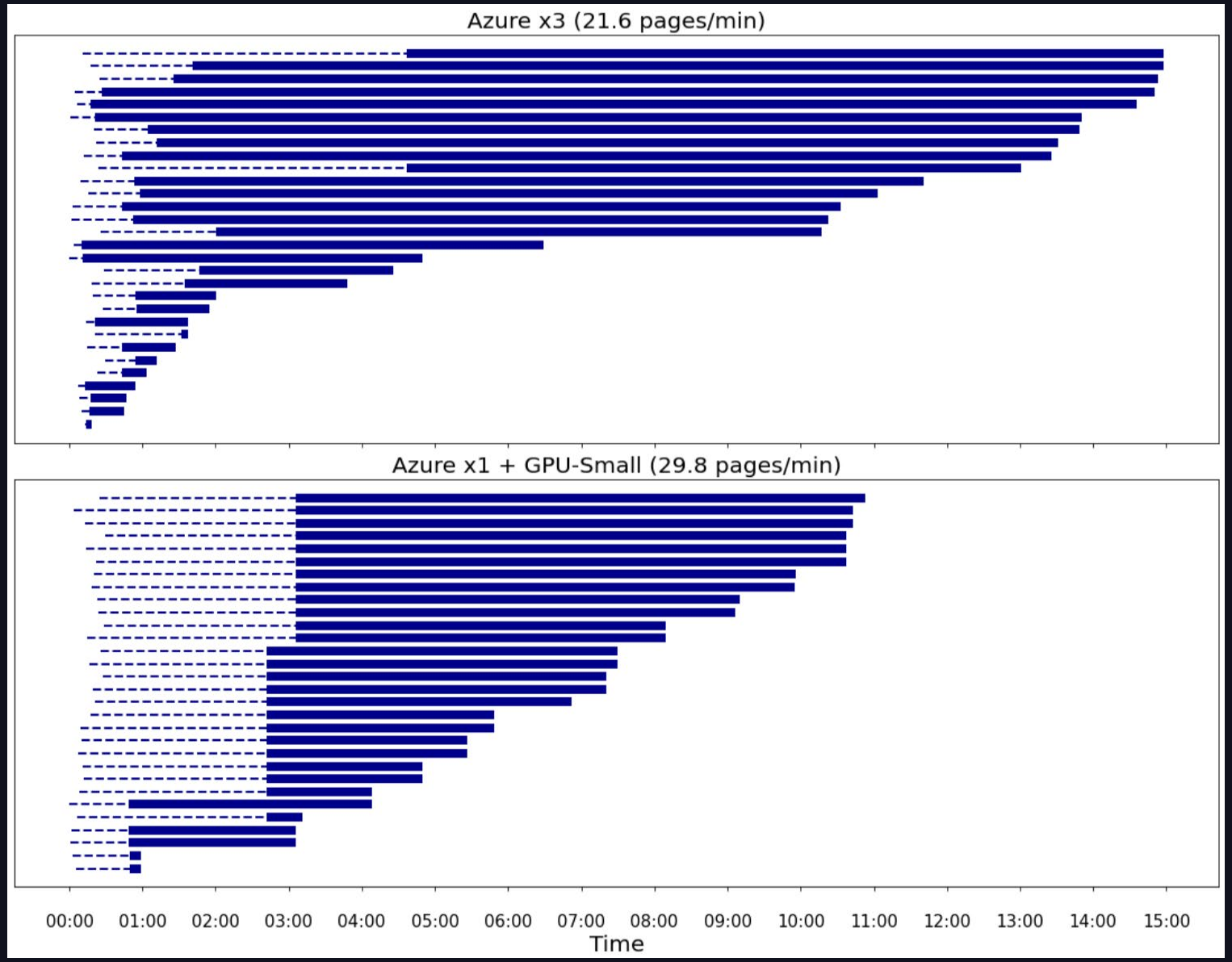
MODEL SERVING - SCALING BEHAVIOR



©2024 Databricks Inc. — All rights reserved

Official Use Only

MODEL SERVING - END TO END PERF (30 DOCUMENT BURST)



©2024 Databricks Inc. — All rights reserved

Official Use Only

LESSONS LEARNT: QUANTITATIVE

Azure Functions vs Model Serving: Quantitative

Solution	Compute	US$/Hour
Azure Function	Isolated v2 I1V2	$0.34
Databricks	GPU Small (T4) x4	$0.73 (10.48 DBU)

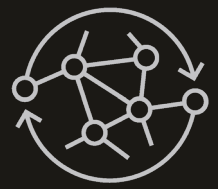
Solution	Latency (Avg.)	Throughput (Total)	US$/Hour
Azure x3	7m22s	22 pages/min	$1.02
Azure x1 + GPU Small	6m39s	30 pages/min	$1.07

©2024 Databricks Inc. — All rights reserved

Official Use Only

LESSONS LEARNT: QUALITATIVE

Azure Functions vs Model Serving



Implementation Complexity

Scaling

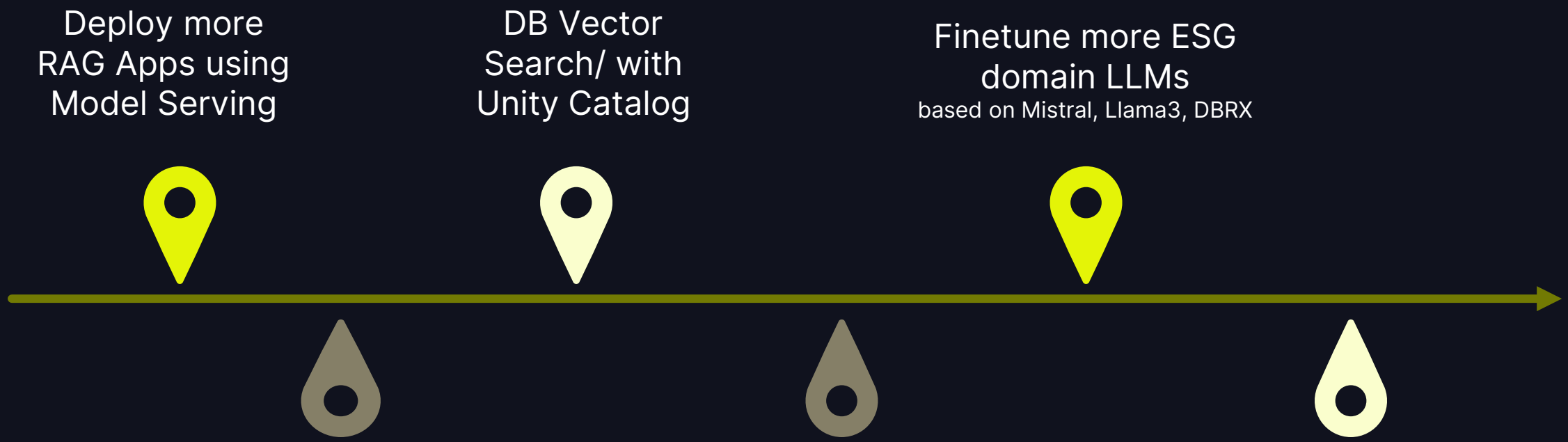
Debugging

©2024 Databricks Inc. — All rights reserved

Official Use Only

MALENA ROADMAP

AI Research *Model Serving* *Data Automation*



©2024 Databricks Inc. — All rights reserved

Official Use Only

THANK YOU!

BLAISE SANDWIDI

Data Science Lead, International Finance Corporation

L linkedin.com/in/blaise-sandwidi-phd/
E bsandwidi@ifc.org
W www.ifc.org
A 2121 Pennsylvania Ave NW Washington, DC 20433

JONATHAN LORENTZ

Data Scientist, International Finance Corporation

L linkedin.com/in/jonathan-lorentz-824260146/
E jlorentz@ifc.org
W www.ifc.org
A 2121 Pennsylvania Ave NW Washington, DC 20433

AN INNOVATION OF IFC International Finance Corporation WORLD BANK GROUP

IN PARTNERSHIP WITH

MINISTRY OF FOREIGN AFFAIRS OF DENMARK Danida

Ministry of Economy and Finance

Fias THE FACILITY FOR INVESTMENT CLIMATE ADVISORY SERVICES

©2024 Databricks Inc. — All rights reserved

Official Use Only

Join as on June 13th, 2024, at 12:30 PM

SESSION

DELIVERING DOMAIN SPECIFIC LLMS WITH GPU SERVING: CASE OF IFC MALENA

OVERVIEW

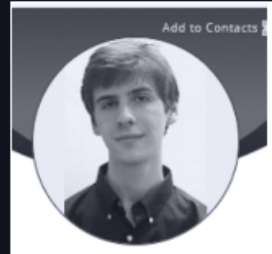
The International Finance Corporation (IFC), a member of the Word Bank Group, is harnessing the power of data and AI to address the development challenges of poverty and climate change. IFC successfully scaled its AI-powered MALENA platform using Lakehouse to accelerate the development of custom large language models. As AI model sizes grow and users expect inference results in real-time, secured and low-latency model serving becomes critical. In this session, the IFC team will share how use of Databricks' model serving enhanced real-time inferencing when serving internal IFC users and external B2B REST API users. The team will share their LLM Ops journey and performance metrics for Azure Functions versus CPU model serving, particularly for serving fine-tuned models built on Google BERT. The team will show how and why optimized GPU model serving may offer an optimal solution for fine-tuned models trained on foundation models such as Llama 2 or Mistral.

EXPERIENCE	IN PERSON
TYPE	BREAKOUT
TRACK	GENERATIVE AI
INDUSTRY	PUBLIC SECTOR, FINANCIAL SERVICES
TECHNOLOGIES	AI/MACHINE LEARNING, GENAI/LLMS, MLFLOW
SKILL LEVEL	INTERMEDIATE
DURATION	40 MIN

Blaise Sandwidi

/ Lead Data Scientist, ESG Officer, PhD

International Finance Corporation (IFC)



Jonathan Lorentz

/ Data Scientist

International Finance Corporation

©2024 Databricks Inc. — All rights reserved

Official Use Only