

LLMops (MLops) Pipeline

LLM-Powered Real-Time Translation System

“A scalable multilingual translation system leveraging Kafka, FastAPI, and Hugging Face — achieving real-time, low-latency LLM deployment with full observability.”



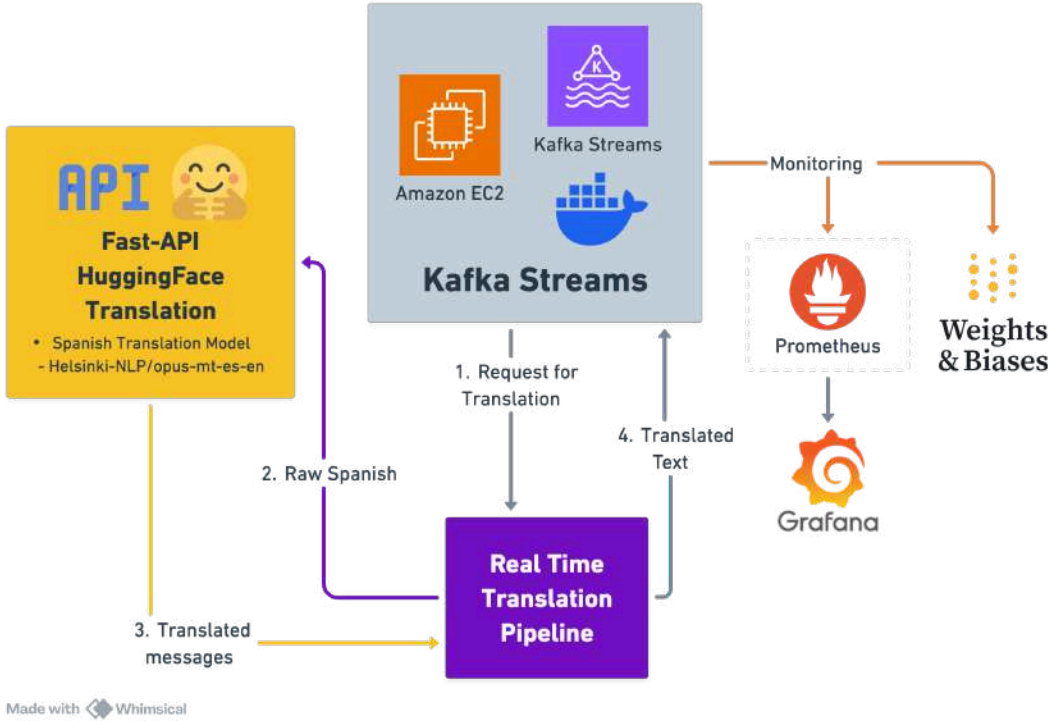
A Kafka-Driven Pipeline with MLOps Automation

Author: Wonha Shin | University of Rochester

Tools: Kafka • FastAPI • Hugging Face • Docker • Prometheus • Grafana • W&B • AWS EC2

This project presents a **real-time multilingual translation pipeline** operationalized with **LLMOps** principles. The system leverages **Kafka** for distributed streaming, **FastAPI** for real-time translation serving, and **Hugging Face’s Helsinki-NLP/opus-mt-es-en** model for Spanish → English translation. The entire pipeline was deployed on **AWS EC2**, with containerized monitoring through **Prometheus + Grafana**, and experimental integration with **Weights & Biases (W&B)** for logging latency and translation quality.

🧩 The goal: achieve low-latency, high-throughput translation while maintaining observability, scalability, and reliability.



Kafka for Data Streaming

Backbone for message ingestion and routing — handles millions of multilingual messages per day with three brokers and ZooKeeper for high availability.

Translation Model (Hugging Face)

Spanish → English translation with `Helsinki-NLP/opus-mt-es-en`, deployed as a FastAPI microservice for sub-second inference latency.

API Management (FastAPI)

REST endpoints for real-time translation, seamlessly integrated with Kafka topics for streaming ingestion and output publishing.

Monitoring & Observability

Prometheus + Grafana dashboards for Kafka metrics, latency, and resource utilization.

Early-stage W&B integration for BLEU and drift metrics.

Deliverables

Key Components of the Pipeline

1. Kafka for Data Streaming

- Serves as the **backbone** of message ingestion and routing.
- Three brokers + three ZooKeeper nodes ensure **fault tolerance and high availability**.
- Kafka topics are partitioned across brokers to manage:
 - `processed_news_spanish` → incoming RSS data
 - `translated_news_spanish_to_english` → model outputs
 - `final_translations` → enriched, ready-to-serve data

2. Translation Model

- Utilized **Hugging Face’s Helsinki-NLP/opus-mt-es-en** model for **Spanish → English** translation.
- Deployed via **FastAPI microservice** on an EC2 instance.
- Designed for **low latency (<1s)** and **context-preserving translations**.
- Future versions may include **multi-language** and **adaptive retraining** capabilities.

3. API Management (FastAPI)

- Handles translation requests and communicates with Kafka topics.
- **Producer–Consumer architecture:**
 - `producer.py` → fetches El País RSS feeds, sends to Kafka.
 - `RealTimeTranslationPipe.java` → processes summaries, sends to FastAPI.
 - `translation_service.py` → performs model inference and publishes results back to Kafka.
- Provides REST endpoints for on-demand translation queries.

4. Monitoring and Observability

- **Prometheus:** collected system-level and Kafka metrics (CPU, memory, latency, replication).
- **Grafana:** real-time dashboards visualized:
 - Kafka cluster health, topic throughput, and JVM memory usage.
 - Node resource utilization and network performance.
- **Weights & Biases (W&B):**
 - Attempted integration for **latency**, **BLEU score**, and **system drift** tracking.
 - Identified challenges in real-time metric alignment → outlined solutions for schema consistency and custom panels.

System Insights & Lessons Learned

- **Scalability:** The Kafka-based architecture effectively handled high-throughput workloads, highlighting its robustness for real-time applications.
- **Integration Challenges & Future Fixes**
 - **Scalability:** Kafka-based architecture efficiently handled high-throughput multilingual workloads.
 - **Integration Challenges:** W&B metric schema mismatch revealed areas for improved real-time alignment.

▼ Final Outputs

```
da pasara a los servicios sociales de la junta.html", \ "published": \ "Thu, 12 Dec 2024 13:43:07
GMT\"},"timestamp":"2024-12-13T16:20:11.873203454Z"}
{"text":{"\ "title\": \ "La Junta incumple su plan para proteger el acu\u00edfero de Do\u00f1ana i
niciado hace 10 a\u00f1os\","summary\": \ "La organizaci\u00f3n conservacionista WWF denuncia q
ue tras una d\u00e9cada, el Ejecutivo andaluz solo ha completado el 23% de las medidas previstas\
", \ "link\": \ "https://elpais.com/clima-y-medio-ambiente/2024-12-12/la-junta-incumple-su-plan-para
-proteger-el-acuifero-de-donana-iniciado-hace-10-anos.html\","published\": \ "Thu, 12 Dec 2024 14
:42:32 GMT\"},"timestamp":"2024-12-13T16:20:21.878298692Z"}
{"text":{"\ "title\": \ "Tu d\u00e9cimo esconde una X: el reto matem\u00e1tico de la Loter\u00eda
de Navidad\","summary\": \ "Puedes enviar tu respuesta hasta el 19 de diciembre. El s\u00e1bado
21 publicaremos la soluci\u00f3n\","link\": \ "https://elpais.com/loteria-de-navidad/2024-12-12
/tu-decimo-esconde-una-x-el-reto-matematico-de-la-loteria-de-navidad.html\","published\": \ "Thu,
12 Dec 2024 11:36:42 GMT\"},"timestamp":"2024-12-13T16:20:31.899738025Z"}
{"text":{"\ "title\": \ "Una ni\u00f1a de 11 a\u00f1os de Sierra Leona, \u00fanica superviviente
de un naufragio en el Mediterr\u00e1neo tras pasar 12 horas en alta mar\","summary\": \ "Los res
catistas de la organizaci\u00f3n Compass Collective creen que en la precaria embarcaci\u00f3n vi
ajaban otras 45 personas, que permanecen desaparecidas\","link\": \ "https://elpais.com/internaci
onal/2024-12-12/una-nina-de-11-anos-de-sierra-leona-unica-superviviente-de-un-naufragio-en-el-medi
terraneo-tras-pasar-12-horas-en-alta-mar.html\","published\": \ "Thu, 12 Dec 2024 13:28:50 GMT\"}
","timestamp":"2024-12-13T16:20:41.906531412Z"}
{"text":{"\ "title\": \ "Estas son las tres normas para acceder a los pases de cine nudista por prim
era vez en Espa\u00f1a\","summary\": \ "La pel\u00edcula \u2018T\u00fa no eres yo\u2019 organiza pases para p\u00fablico sin ropa en salas de cine comerciales de varias ciudades\","link\
": \ "https://elpais.com/espana/catalunya/barcelona-se-sale/2024-12-12/estas-son-las-tres-normas-par
a-acceder-a-los-pases-de-cine-nudista-por-primera-vez-en-espana.html\","published\": \ "Thu, 12 D
ec 2024 15:26:45 GMT\"},"timestamp":"2024-12-13T16:20:51.910500105Z"}
{"text":{"\ "title\": \ "La fiesta m\u00e1s rara del mundo\","summary\": \ "Abr\u00ed el libro co
mo si pasar la tarde leyendo en silencio con dos docenas de desconocidos fuera lo m\u00e1s normal
del mundo\","link\": \ "https://elpais.com/opinion/2024-12-12/la-fiesta-mas-rara-del-mundo.html\
", \ "published\": \ "Thu, 12 Dec 2024 04:00:00 GMT\"},"timestamp":"2024-12-13T16:21:01.926969762Z"
}
```

```
{"text":{"\ "title\": \ "The importance of space, the \u2018third teacher\u2019 of education
\u00f3n\","summary\": \ "In a world where everything changes at great speed, experts claim
the value of the environment as a priority factor in the learning of students\","link\":
\ "https://elpais.com/economy/training/2024-12-13/the-importance-of-space-the-third-master-of
-education.html\","published\": \ "Fri, 13 Dec 2024 08:58:57 GMT\"},"timestamp":"2024-12-1
4T01:31:24.787963647Z"}
{"text":{"\ "title\": \ "Chanel already has design\u00f1ador: the Belgian Matthieu Blazy\"," \
summary\": \ "The rumors are confirmed: the Belgian, who has turned Bottega Veneta into one
of the most acclaimed signatures of the panorama, to boast\u00e9l from pr\u00f3ximo a\u00f1o
to the important charge m\u00e1s to which a fashion designer can aspire\","link\": \ "http
s://elpais.com/moda/moda/2024-12-12/chanel-ya-tene-disenador-el-belga-matthieu-blazy.html\
", \ "published\": \ "Thu, 12 Dec 2024 17:30:33 GMT\","timestamp":"2024-12-14T01:31:34.4478020
95Z"}
{"text":{"\ "title\": \ "Europe must unite in space to compete on the world stage\","summary
\": \ "The Director-General of the European Space Agency refers to the need for Europe to inv
est in astron\u00f3mic research to strengthen its economy\u00e9a and increase its influenc
e on the world\","link\": \ "https://elpais.com/science/2024-12-13/europa-must-unite-in-spa
ce-to-compete-in-the-scene-global.html\","published\": \ "Fri, 13 Dec 2024 12:50:35 GMT\"}
,"timestamp":"2024-12-14T01:31:42.094106304Z"}
```

▼ Example 1

`Producer.py` log output — shows successful ingestion of Spanish RSS articles into Kafka topics (`processed_news_spanish`).

Input (Spanish)

```
{
  "title": "Josep Borrell ser\u00e1 el nuevo presidente del CIDOB",
  "summary": "La elecci\u00f3n busca \u201ereforzar el posicionamiento internacional\u201c de este \u201ethink tank\u2019 barcelo\n\u00e9s",
  "link": "https://elpais.com/espana/catalunya/2024-12-13/josep-borrell-sera-el-nuevo-presidente-del-
cidob.html",
  "published": "Fri, 13 Dec 2024 12:46:11 GMT"
}
```

Output (English)

```
{
  "title": "Albares asks the Polish Presidency to reactivate the negotiation for the official Catalan, Basque and Galician in the EU",
  "summary": "The head of diplomacy Española writes to his homólogo Sikorski in 'esperas of the European semester of Warsaw"
}
```

▼ Example 2

FastAPI translation response — demonstrates real-time Spanish → English translation using Helsinki-NLP/opus-mt-es-en with sub-second latency.

Input (Spanish)

```
{
  "title": "Chanel ya tiene diseñador: el belga Matthieu Blazy",
  "summary": "Se confirman los rumores: el belga, que ha convertido a Bottega Veneta en una de las firmas más aclamadas del panorama, pasa a presumir de uno de los cargos más importantes a los que puede aspirar un diseñador de moda.",
  "published": "Thu, 12 Dec 2024 17:30:33 GMT"
}
```

Output (English)

```
{
  "title": "Chanel already has a designer: the Belgian Matthieu Blazy",
  "summary": "The rumors are confirmed: the Belgian, who has turned Bottega Veneta into one of the most acclaimed signatures of the panorama, to boast one of the important charges to which a fashion designer can aspire."
}
```

▼ Example 3

`RealTimeTranslationPipe.java` stream log — confirms end-to-end message flow from Kafka ingestion to translated output publishing.

Input (Spanish)

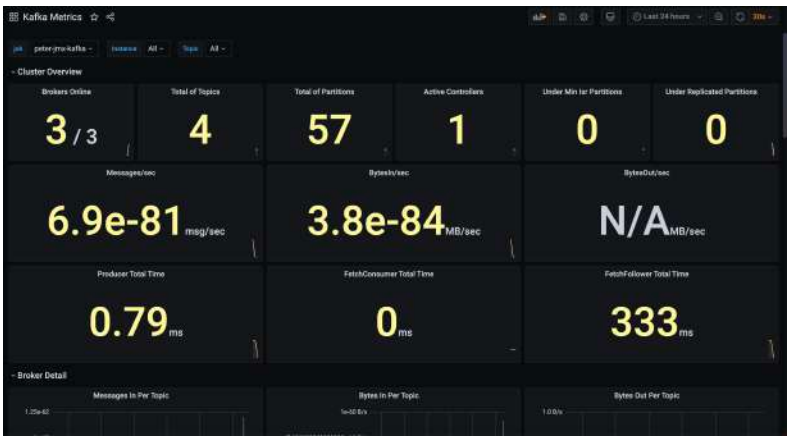
```
{
  "title": "La IA es una quimera real: puede hacer el mundo radicalmente mejor",
  "summary": "Expertos debaten sobre los posibles impactos de la inteligencia artificial en el futuro del trabajo y la educación."
}
```

Output (English)

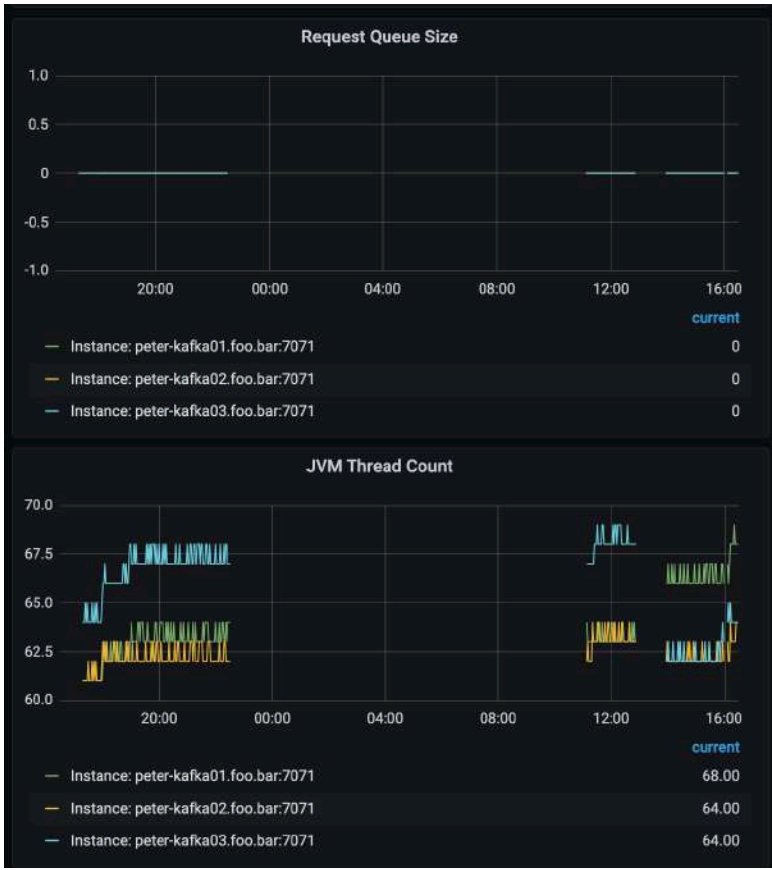
```
{
  "title": "AI is a real chimera: it can make the world radically better",
  "summary": "Experts discuss the potential impacts of artificial intelligence on the future of work and education."
}
```


▼ Monitoring Systems Outputs

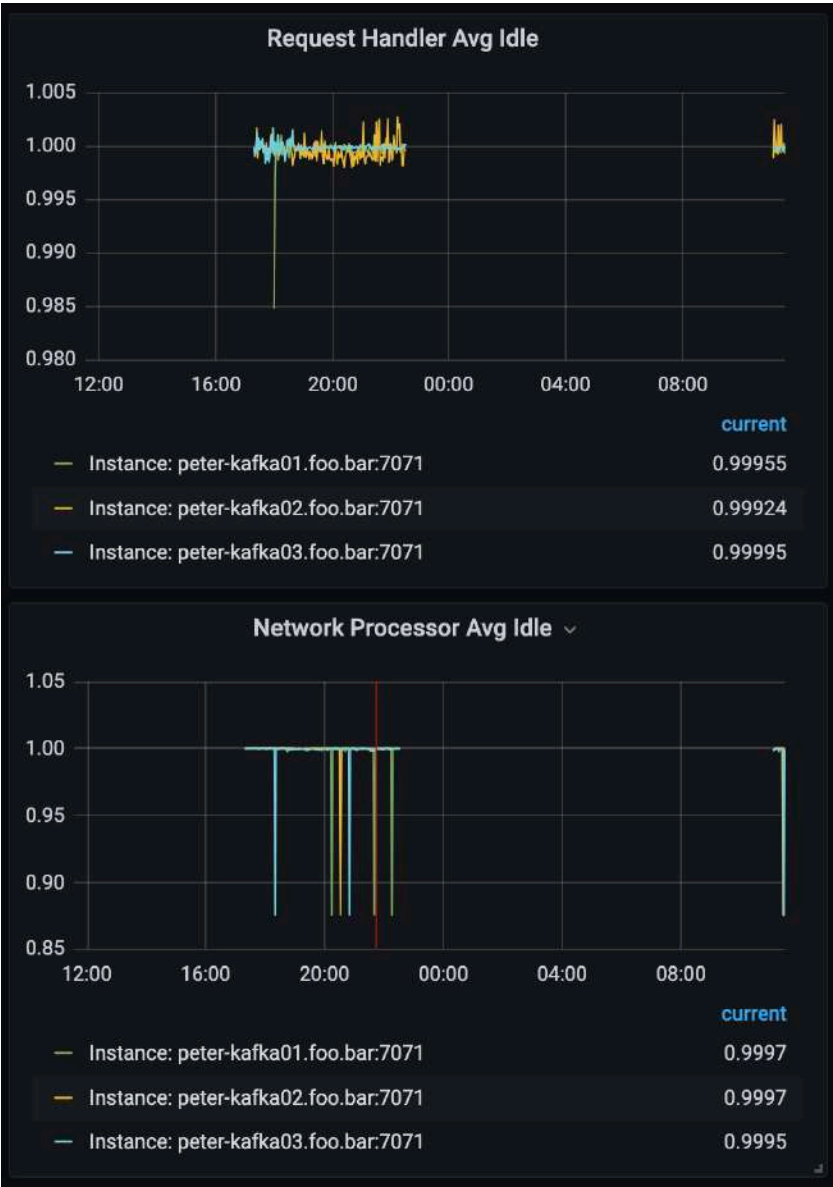
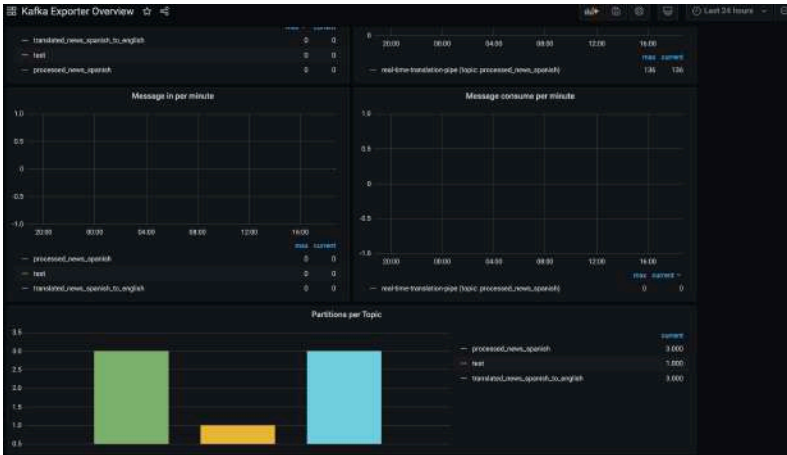
Grafana dashboard visualizing Kafka throughput, node CPU/memory usage, and topic replication health during live translation

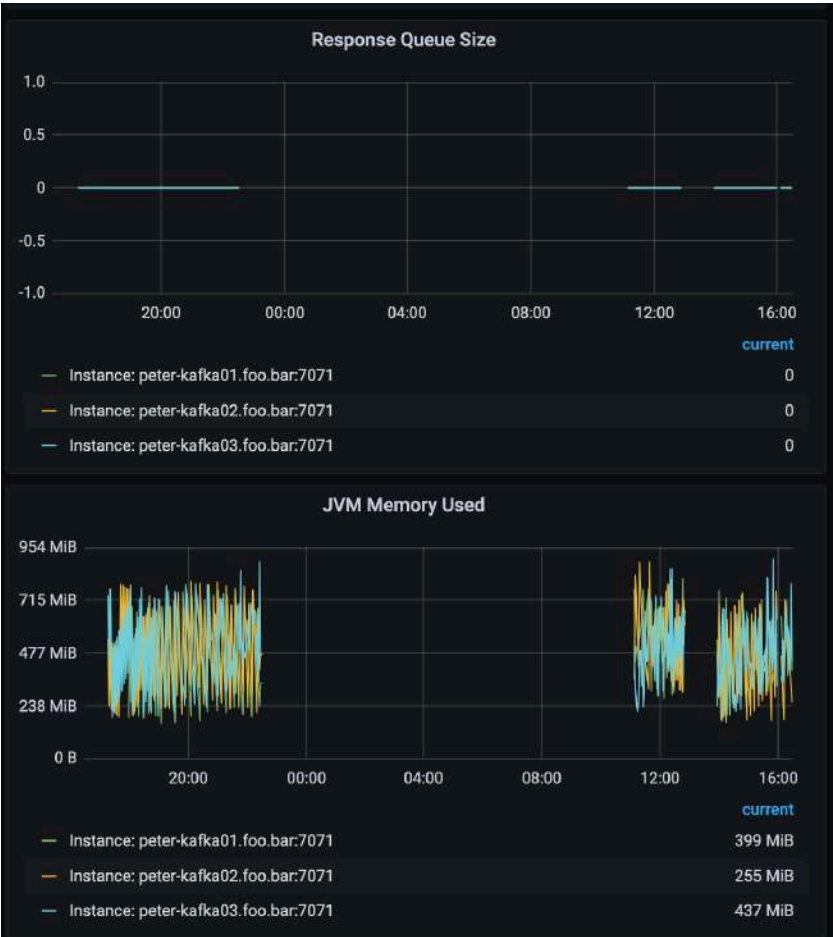
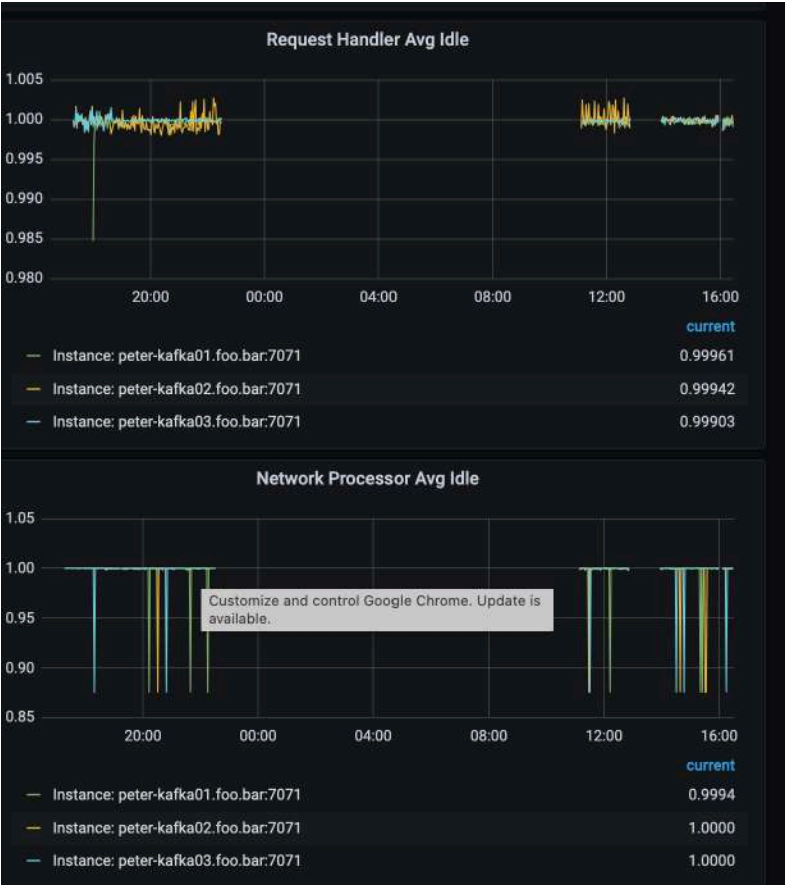


- JVM Metrics

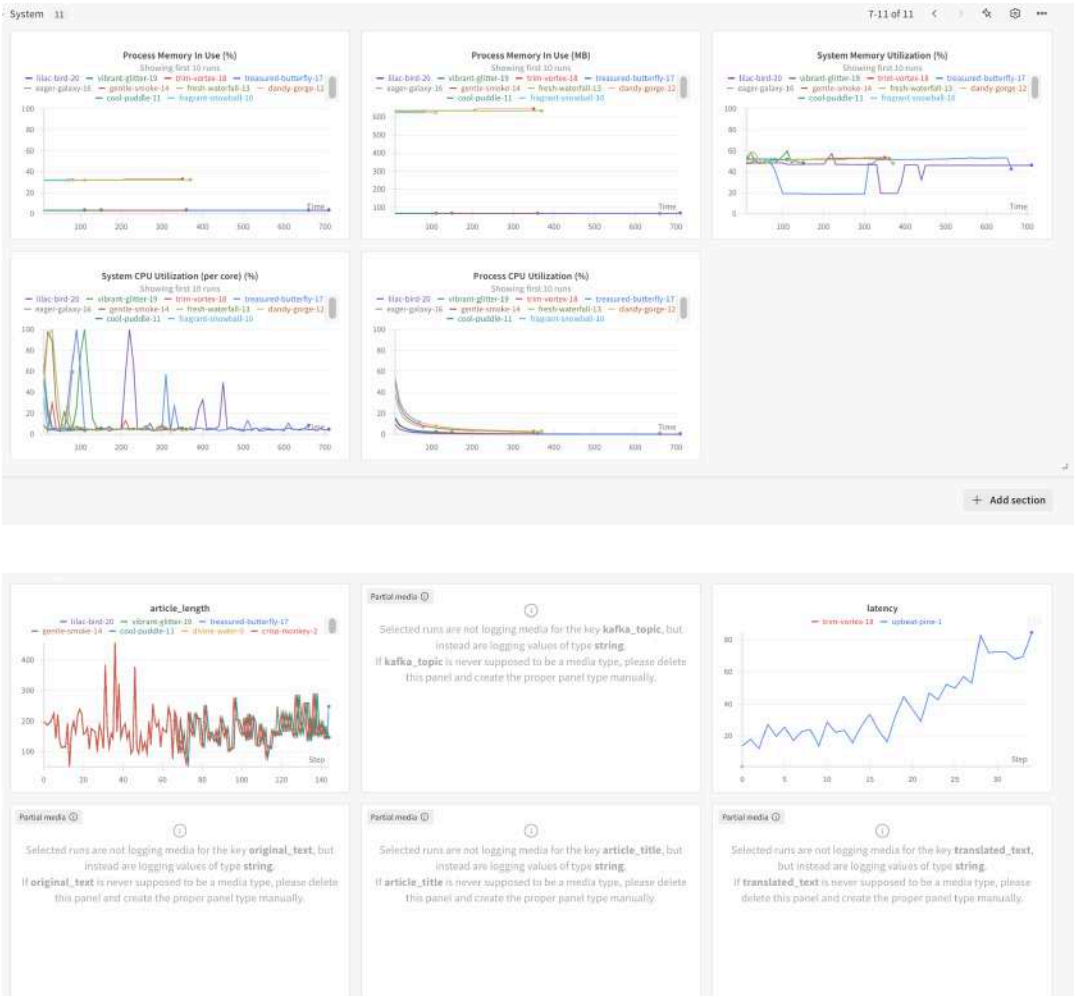


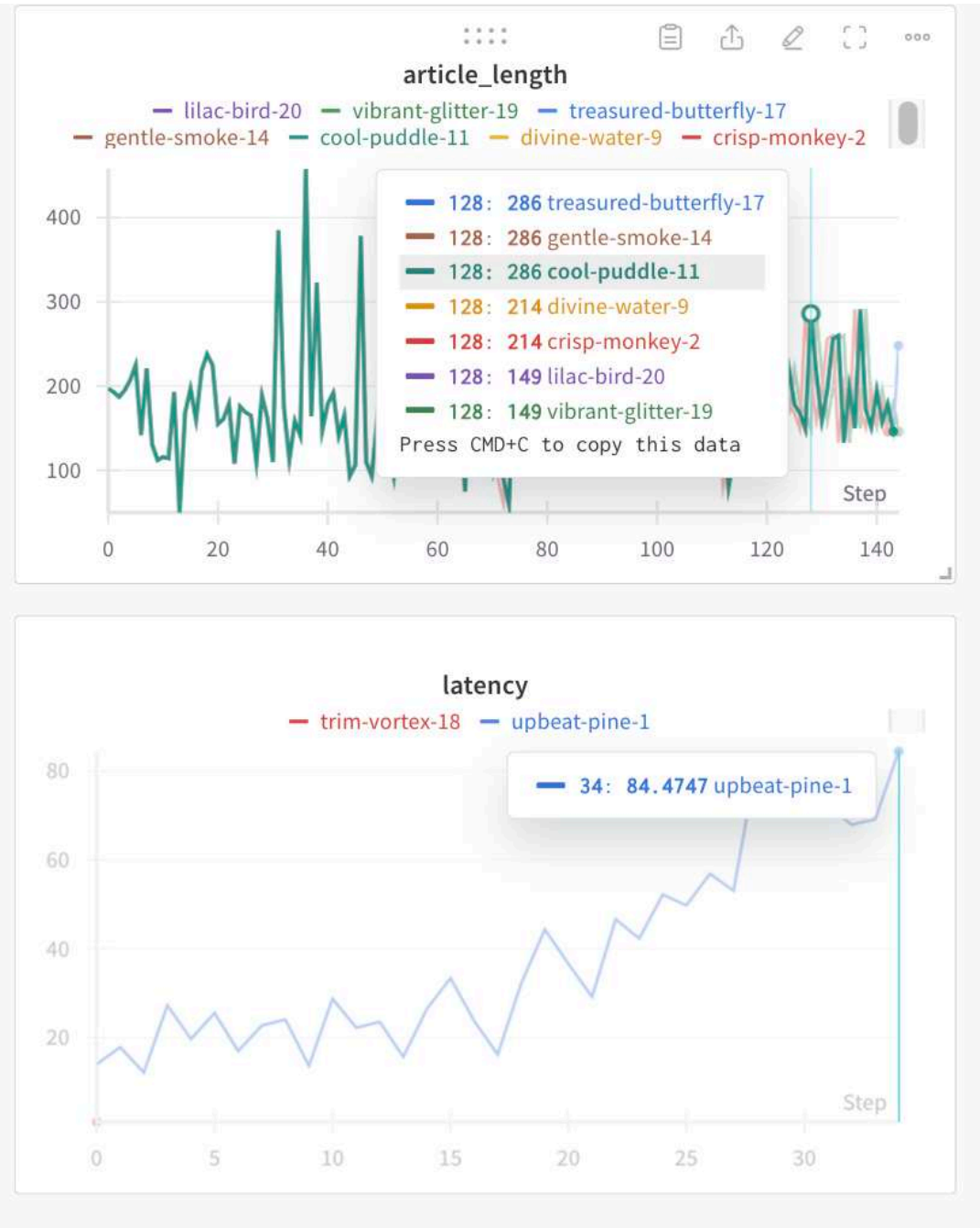
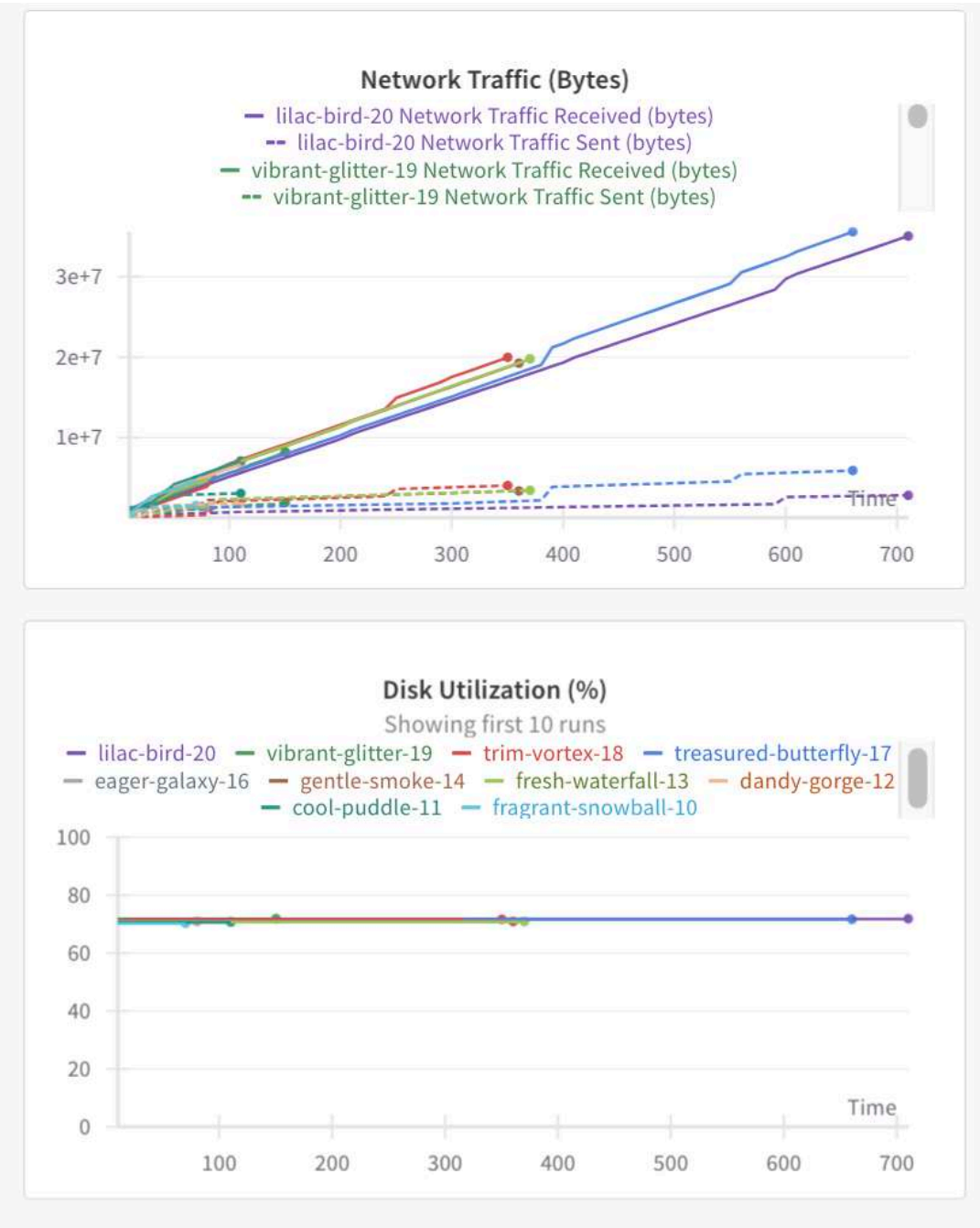
- Kafka Exporter





- Weights & Bias






▼ Final Outputs

Summary of translated article payloads emitted to the `translated_news_spanish_to_english` topic for downstream analytics.

```
JSONDecodeError: Expecting ',' delimiter: line 1 column 481 (char 480), skipping message.
Title: Europe must unite in space to compete on the world stage
Summary: The Director-General of the European Space Agency refers to the need for Europe to invest in astron3mic research to strengthen its economyia and increase its influence on the world
Link: https://elpais.com/science/2024-12-13/europa-must-unite-in-space-to-compete-in-the-scene-global.html
Published: Fri, 13 Dec 2024 12:50:35 GMT
Timestamp: 2024-12-14T01:31:42.094106304Z
```

```
INFO:root:Received translation request: {"title": "Cinco a\u00f1os despu\u00e9s, nadie sabe cu\u00e1ntas personas tienen covid persistente: \u201cTe ven como una vaga que no quiere trabajar\u201d", "summary": "La dificultad del diagn\u00f3stico, el desconocimiento de muchos profesionales y la falta de marcadores biol\u00f3gicos complican el reconocimiento y atenci\u00f3n a la enfermedad", "link": "https://elpais.com/sociedad/2024-12-13/cinco-anos-despues-nadie-sabe-cuantas-personas-tienen-covid-persistente-te-ven-como-una-vaga-que-no-quiere-trabajar.html", "published": "Fri, 13 Dec 2024 04:30:00 GMT"}
INFO:root:Received translation request: {"title": "El mejor libro de cada a\u00f1o de la \u00faltima d\u00e9cada en \u2018Babelia\u2019", "summary": "El suplemento cultural de EL PA\u00cdS publica este s\u00fabado su tradicional n\u00famero especial con lo m\u00e1s destacado del a\u00f1o. Recordamos los t\u00edtulos vencedores en las 10 ediciones anteriores", "link": "https://elpais.com/babelia/2024-12-13/el-mejor-libro-de-cada-ano-de-la-ultima-decada-en-babelia.html", "published": "Fri, 13 Dec 2024 04:30:00 GMT"}
INFO:root:Translation successful: {"title": "\u00bfCu\u00e1l es for you the word of the a\u00f1o?", "summary": "THE PA\u00cdS wants to know cu\u00e1l es, for its readers, the word of this 2024. D\u00e9janos your candidate in the poll to find\u00e9is in this news", "link": "https://elpais.com/culture/2024-12-13/cual-es-para-ti-la-word-del-ano.html", "published": "Fri, 13 Dec 2024 04:40:00 GMT"}
INFO:root:Received translation request: {"title": "Josep Borrell ser\u00e1 el nuevo presidente del CIDOB", "summary": "La elecci\u00f3n busca \u201creforzar el posicionamiento internacional\u201d de este \u2018think tank\u2019 barcelon\u00e9s", "link": "https://elpais.com/espana/catalunya/2024-12-13/josep-borrell-sera-el-nuevo-presidente-del-cidob.html", "published": "Fri, 13 Dec 2024 12:46:13 GMT"}
INFO:root:Translation successful: {"title": "Albares asks the Polish Presidency to reactivate the negotiation for the official Catalan\u00e9n, Basque and Galician in the EU", "summary": "The head of diplomacy Espa\u00f1ola writes to his hom\u00f3logo Sikorski in \u201c\u00edsperas of the European semester of Warsaw \u201d, "link": "https://elpais.com/espana/2024-12-13/albares-pide-a-la-presidency-polaca-reactivate-the-negotiation-for-the-official-de-catalan-euskera-y-gallego-en-la-ue.html", "published": "Fri, 13 Dec 2024 16:38:11 GMT"}
INFO:root:Received translation request: {"title": "\u2018Astro Bot\u2019 gana el premio a mejor videojuego de 2024 en The Game Awards, los galardones m\u00e1s importantes del sector", "summary": "Junto al GOTY, condecoran tambi\u00e9n a \u2018Metaphor: ReFantazio\u2019 y \u2018Balatro\u2019 y anuncian los lanzamientos de \u2018The Witcher IV\u2019, protagonizado por Ciri, y lo nuevo de los creadores de 'The Last of Us\u2019", "link": "https://elpais.com/cultura/2024-12-13/astro-bot-gana-el-premio-a-mejor-videojuego-de-2024-en-the-game-awards-los-galardones-mas-importantes-del-sector.html", "published": "Fri, 13 Dec 2024 10:42:28 GMT"}
^C
```

 **Future Work:** Integrate adaptive retraining, expand to multilingual (EN–FR–KR), and implement automated W&B drift triggers for full LLMOps automation.