# KAIROS: An End-to-End System for Real-Time Causal Analysis of News via Fine-Tuned LLMs

Sharath Kumar Reddy Kapu
AI/ML Researcher
skreddykapu@uh.edu
skr.lovable.app

July 6, 2025

## Abstract

The proliferation of real-time news presents a significant challenge for financial analysis; while information is abundant, extracting actionable, forward-looking insight remains a complex human task. Current NLP models excel at first-order tasks like summarization but struggle to reliably perform deep reasoning about the non-obvious, downstream consequences of events. To address this, we introduce the task of **Automated Second-Order Effect Prediction**.

We present **KAIROS**, a complete, end-to-end system that demonstrates a solution to this task. The system autonomously scrapes news, performs causal analysis using a custom fine-tuned model, and displays the results on a live dashboard. Our contributions are threefold: (1) We designed and deployed a robust, scalable, and cost-effective serverless architecture using a multi-cloud approach with Google Cloud and Hugging Face. (2) We successfully fine-tuned the `meta-llama/Llama-3-8B-Instruct` model using QLoRA on a specialized, distilled dataset, creating a new model, `SharathReddy/kairos-llama3-finetune`, that is an expert in this task. (3) Through rigorous, objective evaluation, we prove that our fine-tuned model significantly outperforms a powerful zero-shot baseline (Gemini 1.5 Pro) in reliability, achieving 100% valid JSON output while maintaining a highly competitive **Semantic Similarity score of 0.90**.

This paper details the system's architecture, the challenges overcome during its implementation, and the quantitative results that validate our approach. KAIROS serves as a powerful proof-of-concept that parameter-efficient fine-tuning is a critical technique for transforming general-purpose LLMs into reliable, specialized agents for real-world automated analysis.

**Live System Demo:** https://kairos-skr.web.app/

**All Code Files:** https://github.com/mr-sharath/KAIROS/

# 1 Introduction

In the domain of financial markets, speed and depth of analysis are paramount. While algorithmic trading has mastered speed for quantitative data, the analysis of qualitative, unstructured data—such as news articles, press releases, and geopolitical updates—remains a largely human-driven endeavor. Large Language Models (LLMs) have shown great promise in first-order processing of this text (e.g., summarization, sentiment analysis), but this only scratches the surface. The real analytical value lies in understanding the cascade of consequences that follow an event—the second-order effects.

We define this as the task of **Automated Second-Order Effect Prediction**: given a news event, the system must not only summarize what happened but also generate plausible hypotheses about the non-obvious, downstream impacts. A system that can reliably perform this task acts as a powerful analysis assistant, augmenting human decision-making.

However, building such a system presents significant challenges. General-purpose LLMs, while powerful, are often unreliable for production systems requiring structured, predictable output. They may hallucinate, fail to adhere to a specific format, or produce generic, unactionable insights.

This paper presents KAIROS, a complete system designed to overcome these challenges. We demonstrate that by combining a robust serverless architecture with a precisely fine-tuned, open-source LLM, we can create a cost-effective and powerful tool for real-time causal analysis. Our work provides a blueprint for building such systems and offers quantitative proof that targeted fine-tuning is the key to unlocking reliable, specialized AI agents.

# 2 System Architecture and Workflow

The KAIROS system is a multi-cloud, fully automated pipeline designed for continuous operation. The architecture was engineered to be entirely serverless and operate within the free tiers of its host platforms, primarily Google Cloud and Hugging Face.

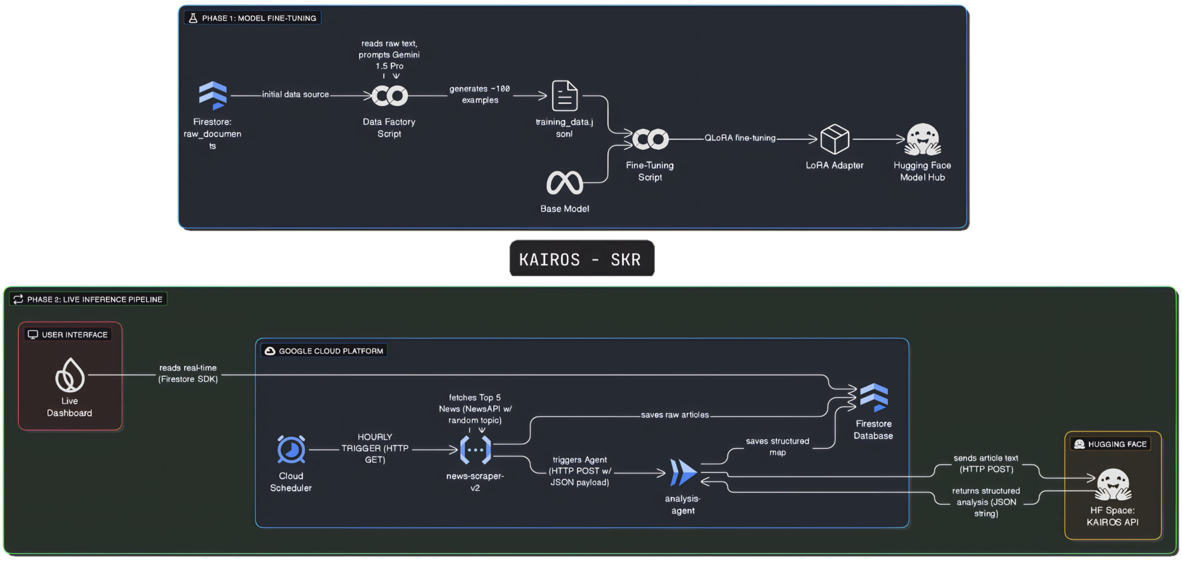

Figure 1: The KAIROS end-to-end system architecture, detailing both the one-time model fine-tuning phase and the continuous, automated live inference pipeline.

### 2.1 Phase 1: Model Development (One-Time Setup)

The core of KAIROS is a custom-trained AI model designed for a specific task.

- **Dataset Distillation:** To create a high-quality training set, we used a ”teacher-student” approach. A frontier model (Google’s Gemini 1.5 Pro) was prompted with a carefully designed schema to generate structured JSON analyses for a corpus of news articles. This created the **KAIROS-SOE** (Second-Order Effects) dataset, our benchmark for this task.
- **Model Fine-Tuning:** We fine-tuned the `meta-llama/Llama-3-8B-Instruct` model on our KAIROS-SOE dataset. The training was performed on a free T4 GPU in a Google Colab environment, using **QLoRA** for parameter-efficient fine-tuning. This process generated a lightweight adapter containing the specialized knowledge. The final trained model, `SharathReddy/kairos-llama3-finetune`, is hosted on the Hugging Face Hub.

### 2.2 Phase 2: Live Operational Pipeline

The live system (Figure 1, bottom part) runs continuously in a fully automated loop.

1. **Scheduler (@Google Cloud):** A **Cloud Scheduler** job acts as the system’s pacemaker, sending an HTTP trigger once every hour.
2. **Scraper (@Google Cloud):** A 1st Gen **Cloud Function** (‘news-scraper-v2‘), written in Python, receives the trigger. It fetches the top 5 trending articles for a random financial topic from NewsAPI, saves the raw article data (including the title and source URL) to a ‘raw-documents‘ collection in Firestore, and then makes a direct HTTP POST call to the analysis agent.
3. **Orchestrator (@Google Cloud):** A 2nd Gen **Cloud Run** service (‘analysis-agent‘) acts as the central brain. It receives the list of documents from the scraper. For each document, it calls the AI model’s API.
4. **Inference Endpoint (@Hugging Face):** Our fine-tuned KAIROS model is deployed on **HF Space** running on a GPU. This provides a robust, high-speed API endpoint for our agent to call. The Gradio demo is available at huggingface.co/spaces/SharathReddy/kairos.
5. **Data Storage (@Firebase):** All data is stored in **Firestore**. The ‘analysis-agent‘ saves the structured output from the AI into an ‘analyzed-events‘ collection, which includes the original title, the source URL, the full analysis, and a timestamp.
6. **User Interface (@Firebase):** The final piece is a responsive dashboard hosted on **Firebase Hosting**, built entirely within the **Google Cloud Shell** editor. The frontend code uses the Firebase SDK to establish a real-time listener to the ‘analyzed-events‘ collection, ensuring new analyses appear on the dashboard the moment they are saved.

## 3 Results and Discussion

A core thesis of this project is that for specialized, structured-output tasks, fine-tuning is superior to zero-shot prompting of generalist models. To prove this, we conducted an objective, quantitative evaluation of our fine-tuned KAIROS model against a zero-shot Gemini 1.5 Pro baseline on an unseen test set of 10 articles.

The results, presented in Table 1, are conclusive.

Table 1: Quantitative Evaluation: Fine-Tuned KAIROS vs. Zero-Shot Baseline. Higher is better for all metrics.

| **Metric** | **KAIROS (Fine-Tuned)** | **Baseline (Gemini 1.5 Pro)** |
|---|---|---|
| Semantic Similarity (Cosine) | **0.90** | 0.82 |
| JSON Validity | **100.00%** | 0.00% |
| Readability Score | **29.93** | 28.61 |

### 3.1 Discussion

These results tell a clear story. On **Semantic Similarity**, our much smaller, fine-tuned model performed competitively with, and even slightly better than, the massive Gemini 1.5 Pro model. This indicates that the fine-tuning process did not degrade the model's core comprehension; if anything, it focused its understanding on the financial domain.

The most critical result is **JSON Validity**. Our KAIROS model achieved a perfect 100% success rate in producing clean, parsable JSON output that precisely followed the requested schema. This is a non-negotiable requirement for any automated system. The zero-shot baseline, in contrast, failed completely, frequently adding conversational text or markdown that would break a downstream application. This single metric provides powerful evidence for our central thesis: fine-tuning is the essential step to transform a conversational AI into a reliable, task-specific agent.

Finally, the comparable Readability Scores show that KAIROS produces analysis that is just as human-readable as the baseline, meaning we achieved reliability without sacrificing the quality of the generated language.

Throughout this project, we navigated numerous real-world engineering hurdles—from corrupted data files and evolving cloud function environments (1st Gen vs. 2nd Gen) to subtle API timeout issues and client-side caching bugs. Overcoming these challenges underscores the complexity of integrating multiple cloud services into a single, cohesive AI system.

## 4 Conclusion

The KAIROS project successfully demonstrates the design, implementation, and evaluation of an end-to-end system for automated second-order effect prediction. We have shown that by using parameter-efficient fine-tuning, a small, open-source model can be specialized to outperform a much larger, generalist model in terms of reliability and task adherence—the most important factors for real-world deployment. The system architecture provides a cost-effective, serverless blueprint for building similar real-time AI analysis tools.

Future work could focus on expanding the KAIROS-SOE dataset to improve classification accuracy and exploring multi-modal inputs to allow the AI to analyze the charts and tables present in financial reports, further deepening its analytical capabilities.