

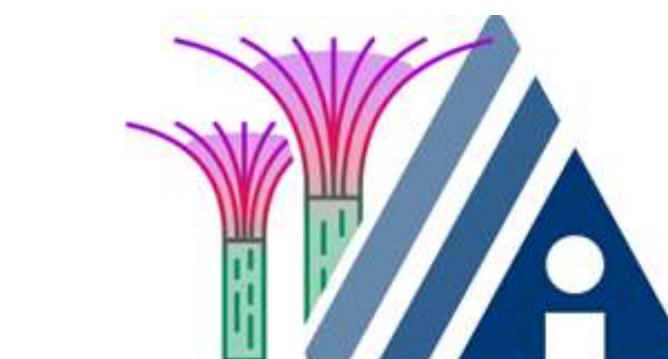
# GAICo: A Unified Framework for Multi-Modal GenAI Evaluation

Streamlining Deployed, Extensible, and Reproducible AI Assessment

Nitin Gupta, Pallav Koppisetty, Kausik Lakkaraju, Biplav Srivastava

The AI4Society Group

pip install gaico



ai4society.github.io/projects/  
GenAIResultsComparator

## Why GAICo?

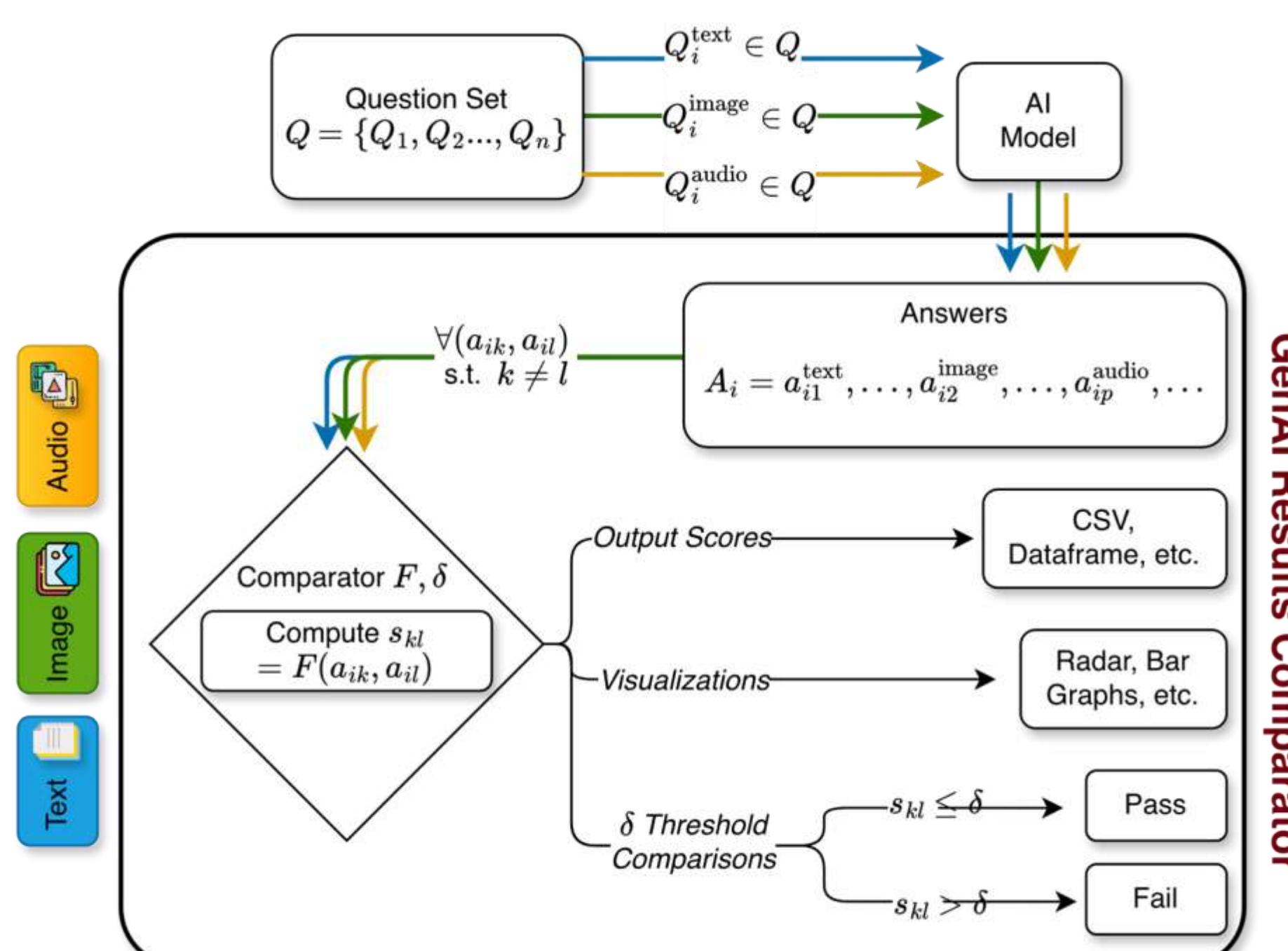
Fragmented GenAI Evaluation  
Composite AI systems (LLMs + image/audio generators) lack unified evaluation tools.

- ❑ **Fragmented tools:** BLEU (text), SSIM (images), librosa (audio) → each siloed
- ❑ **No unified interface:** Incompatible APIs across modalities
- ❑ **Poor reproducibility:** Ad-hoc scripts prevent cross-team comparisons
- ❑ **Hard to debug:** Difficult to isolate orchestrator vs. specialist model failures

The GAICo Solution  
GAICo processes multi-modal AI outputs (text, image, audio, structured data), computes similarity scores against references, and generates:

1. Raw data reports (CSV, DataFrames)
2. Visualizations (radar/bar charts)
3. Pass/fail assessments (threshold  $\delta$ )

```
from gaico.metrics import BaseMetric
class MyCustomMetric(BaseMetric):
    def calculate(self, generated, reference):
        # Metric logic here
        return similarity
```



- ✓ Unified API: Single interface for text, images, audio, and structured data
- ✓ 15+ Metrics: N-gram to semantic to specialized (PlanningLCS, SSIM, AudioSNR)
- ✓ Extensible: Add custom metrics via BaseMetric class
- ✓ Reproducible: Standardized CSV reports and visualizations

## Case Study: Debugging AI Travel Assistants

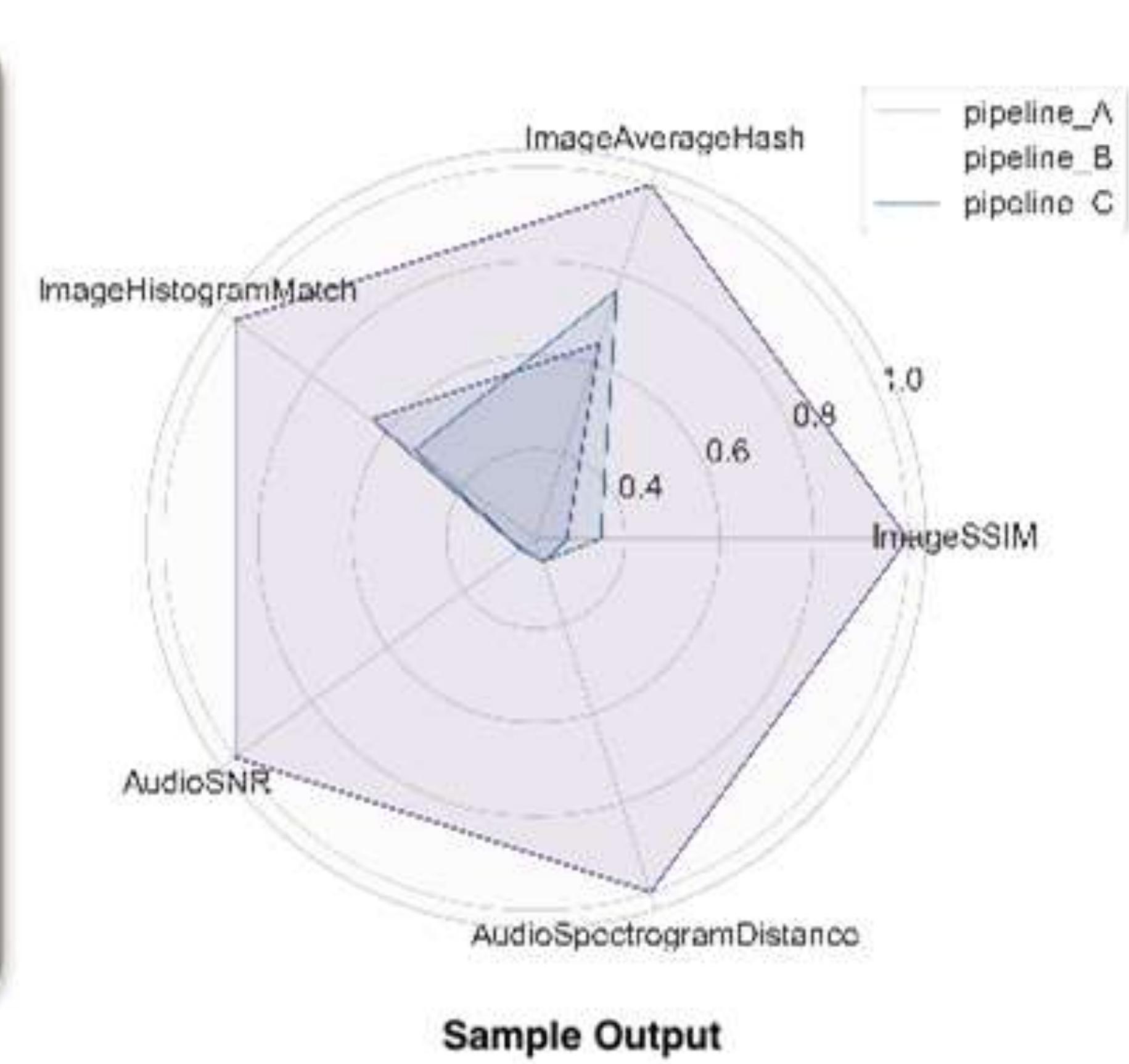
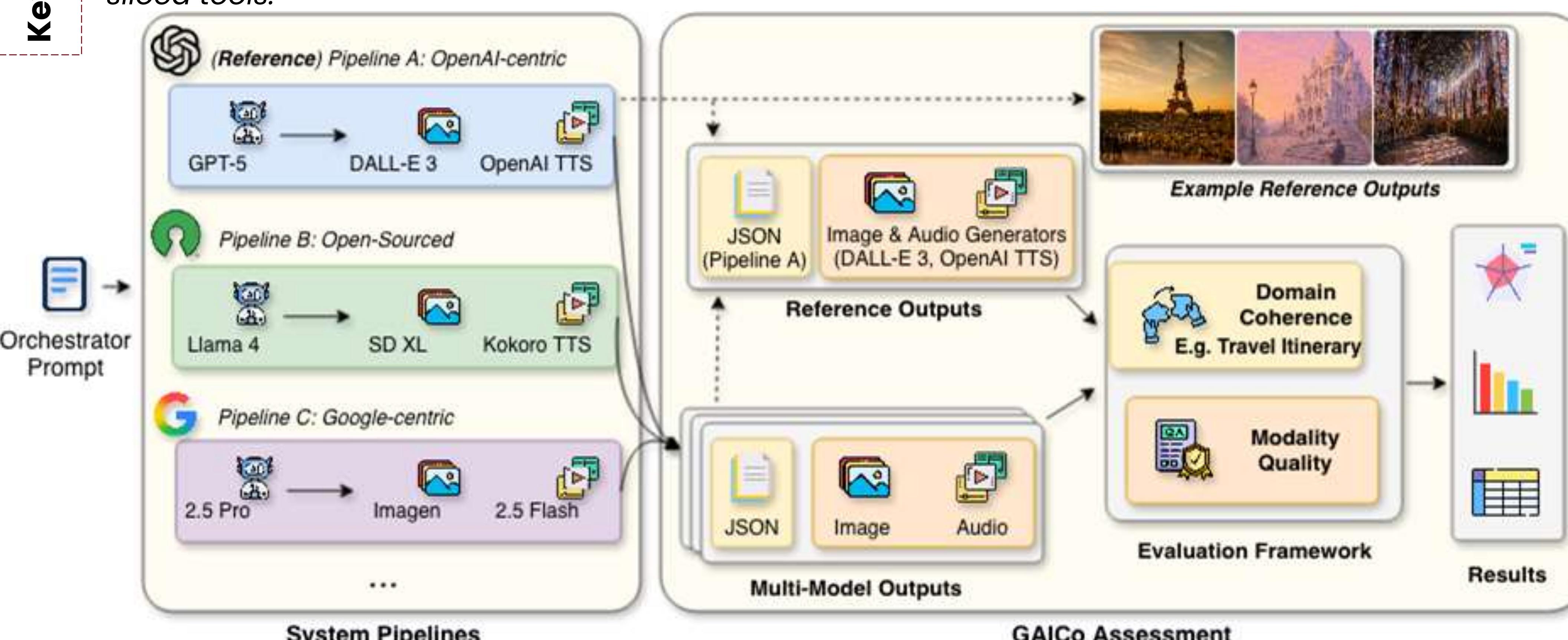
Evaluated 3 AI travel assistant pipelines generating 3-day Paris itineraries (text plans, images, audio):

Two-Part Evaluation Strategy  
Challenge: Failure attribution in composite AI systems

GAICo's Approach:

1. **Plan Coherence:** Compare JSON outputs against baseline (Pipeline A)
2. **Modality Quality:** Compare images/audio against per-pipeline references; isolates specialist model quality from orchestrator prompts

Key Insight  
GAICo pinpointed that performance issues stemmed from orchestrator LLMs, not specialist models, difficult to detect with siloed tools.



## Resources - 17 Ready-to-Run Jupyter Notebooks

### QUICKSTART (5 notebooks)

- ❑ Basic workflow & Experiment module
- ❑ Single/multi-metric evaluation
- ❑ Audio, image, structured data, & case-study evaluation

### ADVANCED (5 notebooks)

- ❑ Custom thresholding techniques
- ❑ LLM FAQ Analysis
- ❑ Custom visualizations

### DOMAIN-SPECIFIC (7 notebooks)

- ❑ Finance evaluation
- ❑ Election analysis
- ❑ Recipe generation
- ❑ Planning sequences & time-series

### COMMUNITY IMPACT

- ❑ ~16,000 downloads (Jun-Dec 2025)
- ❑ 17 Ready-to-run notebooks
- ❑ 100% Open-source & actively maintained
- ★ Growing Community contributions & adoption

pallav@email.sc.edu  
niting@email.sc.edu

