From Semantic Extraction to Autonomous Agents: Architecting the Next Generation of Enterprise AI

Grigorios Tsoumakas
Professor @ Aristotle University of Thessaloniki
Cofounder & Chief Scientific Officer @ Medoid AI

ARISTOTLE UNIVERSITY OF THESSALONIKI

Medoid AI

Agenda

Work at Aristotle University

- Enterprise ontology

Work at Medoid AI

- Agent memory infrastructure
- The future of AI data infrastructure







Work at Aristotle University of Thessaloniki

Relation extraction

Semantic indexing

Summarization

November 4-5, 2026

Relation Extraction

Sample sentence from a post-mortem report: "*During the outage, the edge router failed to resolve traffic through the primary DNS server*"



Capabilities

- Construction of knowledge graphs, enabling GraphRAG
- Reduction of inference compute costs

Enterprise Applications

- Automated root cause analysis by mapping dependencies in incidence logs
- Compliance verification by linking obligations to internal policies

Christou, D., & Tsoumakas, G. (2026). Sub-Billion, Super-Frontier: Small Language Models Rival Zero-Shot Frontier LLMs on General and Literary Relation Extraction. *ArXiv, abs/2606.22606*.

Does Size Matter in Relation Extraction?

LLMs for relation extraction

The performance of frontier models is strong

Yet their high computational demands raise challenges in

- Accessibility
- Sustainability
- Privacy

Can small LMs close the gap?

Task: Relation classification

Datasets: 7 General, 2 Literature

Models: RoBERTA, Small (<1B, 3B) LMs fine-tuned with QLoRA (SmolLM, Qwen, Llama), Frontier LMs (OpenAI, Anthropic)

Prompt: 0-shot, 2-shot, 10 different paraphrased templates

Christou, D., & Tsoumakas, G. (2026). Sub-Billion, Super-Frontier: Small Language Models Rival Zero-Shot Frontier LLMs on General and Literary Relation Extraction. *ArXiv, abs/2606.22606*.

Average Results on the 7 General Datasets

Task-specific adaptation beats raw scale: SLMs surpass 0-zhot frontier LMs

2-shot prompting helps, especially the sub-billion models

In the niche literature domain, frontier models struggle more

Model	Params	Shots	μF1
Llama 3.2	3B	2	0.844
Qwen 2.5	0.5B	2	0.828
RoBERTa	125M	-	0.826
Llama 3.2	3B	0	0.821
SmolLM 2	360M	2	0.735
Qwen 2.5	0.5B	0	0.714
GPT-5.4	~ 2.2T	0	0.693
Sonnet 4.6	~ 1.5T	0	0.662
SmolLM 2	360M	0	0.527

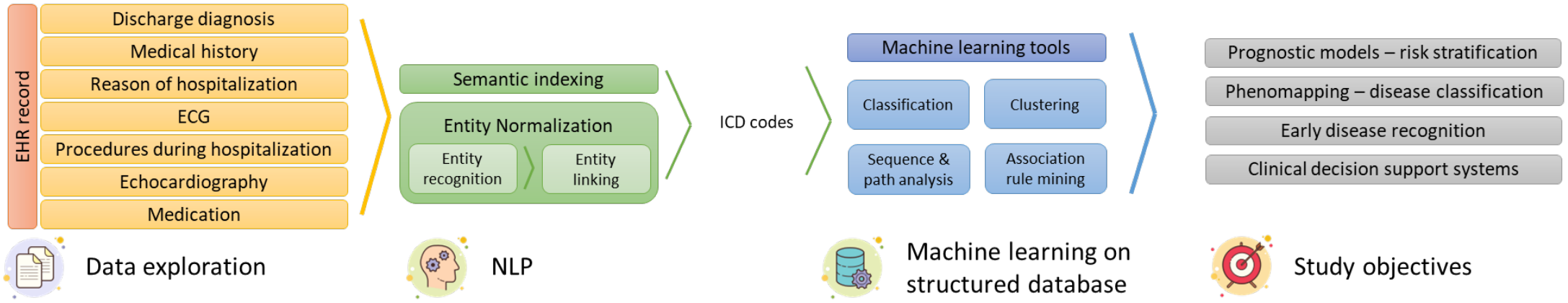
The CardioMining Study



NLP methods for clinical coding of unstructured textual information

Explainable ML methods for diagnostic and prognostic models





Dimitriadis, D., Patsiou, V., ..., Giannakoulas, G., & Tsoumakas, G. (2025). Overview of ElCardioCC Task on Clinical Coding in Cardiology at BioASQ 2025. Conference and Labs of the Evaluation Forum.

ELCardioCC Challenge 2025 - today

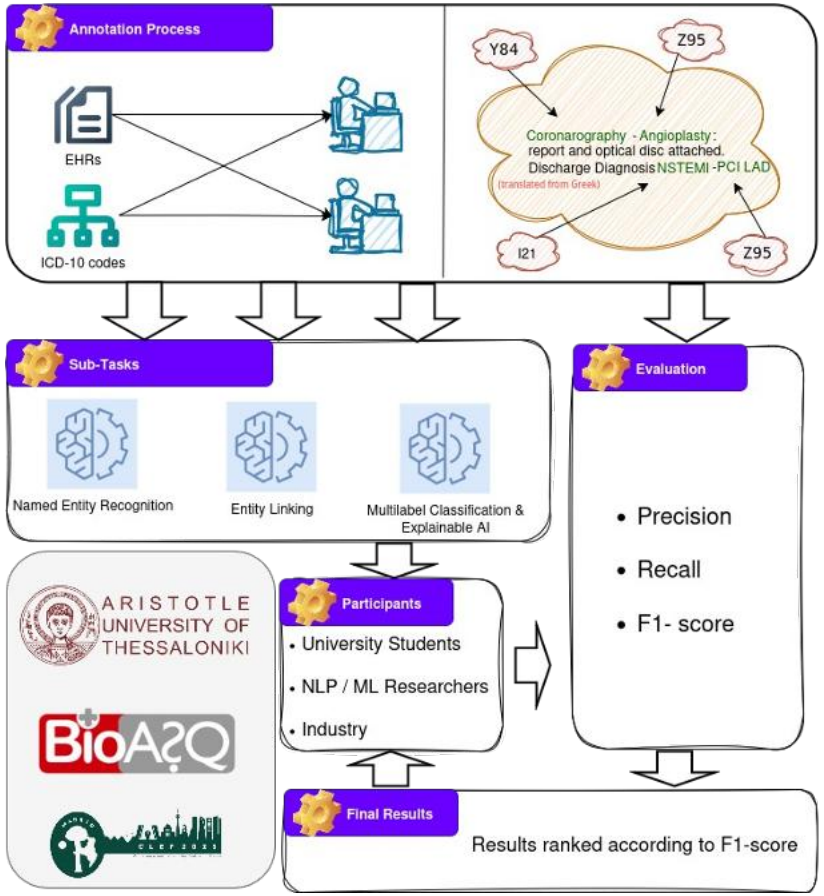
Clinical coding of Greek (low resource lang.) cardiology discharge letters with ICD-10 codes

- a) named-entity recognition + entity linking
- b) multi-label classification

Data: 2,500 train & 500 test discharge letters annotated by at least 2 experts at document & mention level (subset) with ICD-10 codes

Evaluation Metrics: Precision, Recall, F1

- 2025: NER+EL 0.85 F1
- 2026: MLC Transformer Encoder Ensemble 0.87 F1



E2ECardioAI: End-to-End AI-Powered Data Management and Analytics Platform for Enhanced Cardiovascular Care

Project details

12/2026 – 11/2029

1.4 million Euros

“Research – Innovate” Competitiveness Progr., NSRF 2021–2027

WP4: NLP Techniques for the Automated Structuring of Unstructured Clinical Data

Semantic indexing

Entity normalization

Generalization to simpler ontologies, like Diagnosis-Related Group (DRG)

Multi-Label Classification

Bin Liu, Assoc. Prof.
Chongqing University of Posts & Telecommunications

PhD graduate of AUTH 2017-21



Batch Selection in Deep Multi-Label Classification

- Zhou, A., Liu, B., Peng, Z., Wang, J., Tsoumakas, G. (2024). Multi-label Adaptive Batch Selection by Highlighting Hard and Imbalanced Samples. In Proc. **ECML PKDD 2024**
- Zhou, A., Liu, B., Wang, J., Tsoumakas, G. (2025). Batch selection for multi-label classification guided by uncertainty and dynamic label correlations. In Proc. **AAAI 2025**
- Liu, B., Peng, H., Wei, Z., Zhang, J. & Tsoumakas, G. (2026) D2ACE: Multi-Label Batch Selection Guided by Dual Dynamics and Adaptive Correlation Enhancement. Proceedings of **IJCAI 2026**

Takeaways

- Focus on samples where the model is uncertain in early epochs
- Focus on high loss, yet non-noisy, samples at late epochs
- Dynamically weight the labels based on their importance (higher mean uncertainty & hardness)
- Diffuse global label correlation patterns in uncertainty and hardness of samples at the local level

HARNESS: Human-Centered Neural Abstractive Summarization

Project details

10/2025 – 09/2028

200k Euros budget

Hellenic Foundation for Research & Innovation

Objectives

Mitigate the cost of obtaining in-domain data for neural abstractive summarization models

Personalize the output of neural abstractive summarization models to different user profiles

Increase the trustworthiness of users to neural abstractive summarization models

Advance lay summarization of biomedical literature

Paraskevopoulou, M.N., Passali, T., & Tsoumakas, G. (2026). Less is More: Quality-Aware Training Data Selection for Scientific Summarization.

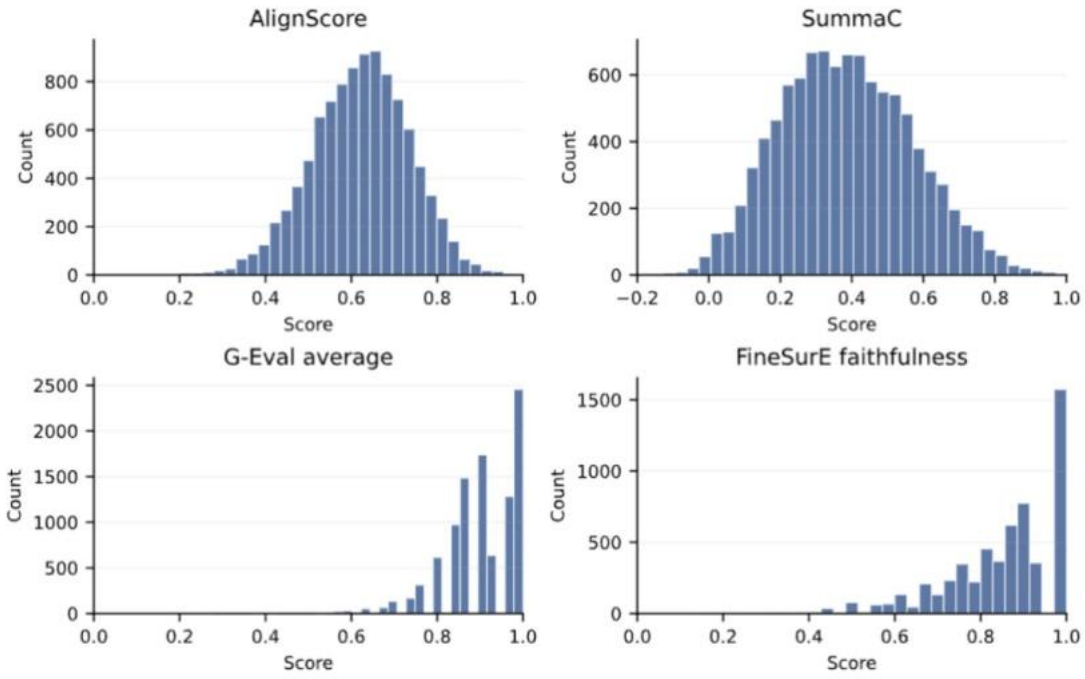
Scientific Article Summarization

https://huggingface.co/datasets/nefelipar/pmc-large

A new scientific summarization corpus

Dataset aspect	PubMed	PMC-large
Source	PMC	PMC-OA
Target	Author abstract	Author abstract
Pairing	article-abstract	article-abstract
Size	133K	1.88M
Avg. doc. length	3,016	3,918
Avg. sum. length	203	218
Sections	Level-1	Nested
Format	Plain text	Markdown
Section filtering	Up to conclusion	Content-aware
Figures/tables	Removed	Removed
Inline refs	Not specified	Placeholders
Math/Citations	Special token	Placeholders
Code	Not specified	Placeholders
Length filtering	Length-based	Percentile-based
Language filter	Not reported	English
License filter	Not reported	CC BY/CC0
Target leakage filter	Not reported	Applied

Are abstracts reliable summaries?



Paraskevopoulou, M.N., Passali, T., & Tsoumakas, G. (2026). Less is More: Quality-Aware Training Data Selection for Scientific Summarization.

Quality-Aware Training Data Selection

Can quality-focused fine-tuning improve summarization quality?

Model	Rank-based Overall ↑	Align ↑	SummaC ↑	G-Eval ↑	FineSurE ↑ / Succ.
Zero-shot	0.444	0.735	**0.462**	0.822	0.847 / 76.6
One-shot	0.306	0.712	0.409	0.523	0.616 / 45.0
Two-shot	0.028	0.652	0.378	0.419	0.526 / 33.6
Random 1K	0.361	0.812	0.377	0.831	0.921 / 93.6
All 5K	0.583	0.825	0.382	0.845	0.921 / 95.8
AlignScore 1K	0.778	**0.850**	0.388	0.839	**0.932** / 95.2
FineSurE 1K	0.667	0.834	0.395	0.834	0.928 / 94.0
SummaC 1K	0.472	0.819	0.403	0.819	0.908 / 94.2
G-Eval 1K	0.556	0.819	0.382	**0.848**	0.921 / 93.6
Aggregate 1K	**0.806**	0.844	0.398	0.845	0.931 / 94.8

Qwen 2.5 3B summarization quality on 500 held out articles.
Quality drives selection from a pool of 5k articles.

Paraskevopoulou, M.N., Passali, T., & Tsoumakas, G. (2026). Less is More: Quality-Aware Training Data Selection for Scientific Summarization.

Scaling to 100k articles

Model	Rank-based Overall ↑	Align ↑	SummaC ↑	G-Eval ↑	FineSurE ↑ / Succ.
AlignScore selection from 100K					
Random 1K	0.036	0.810	0.388	0.838	0.906 / 94.8
AlignScore 1K	0.625	**0.887**	**0.405**	0.824	0.949 / 95.8
Random 2K	0.125	0.816	0.384	0.846	0.916 / 95.0
AlignScore 2K	0.571	0.879	0.395	0.841	**0.957** / 95.0
Random 5K	0.179	0.825	0.392	0.843	0.913 / 93.8
AlignScore 5K	**0.679**	0.883	0.398	0.849	**0.957** / 95.4
Random 10K	0.250	0.833	0.390	0.848	0.917 / 94.2
AlignScore 10K	0.589	0.866	0.397	0.868	0.947 / 94.4
All 100K	0.429	0.828	0.391	**0.874**	0.935 / 95.2
Aggregate selection inside the top 10K AlignScore pool					
Agg-All 1K	0.536	0.867	0.400	0.852	0.947 / 94.2
Agg-NoSummaC 1K	0.429	0.864	0.392	0.856	0.945 / 94.4
Agg-All 2K	0.679	0.877	0.401	0.855	0.954 / 94.6
Agg-NoSummaC 2K	0.786	**0.879**	0.401	0.864	0.957 / 96.2
Agg-All 5K	**0.839**	0.876	**0.406**	0.861	**0.961** / 93.6
Agg-NoSummaC 5K	0.750	0.876	0.399	**0.867**	0.958 / 95.2

Work at Medoid AI

Real-world RAG deployments

Continual learning (CEREBRA-AI)

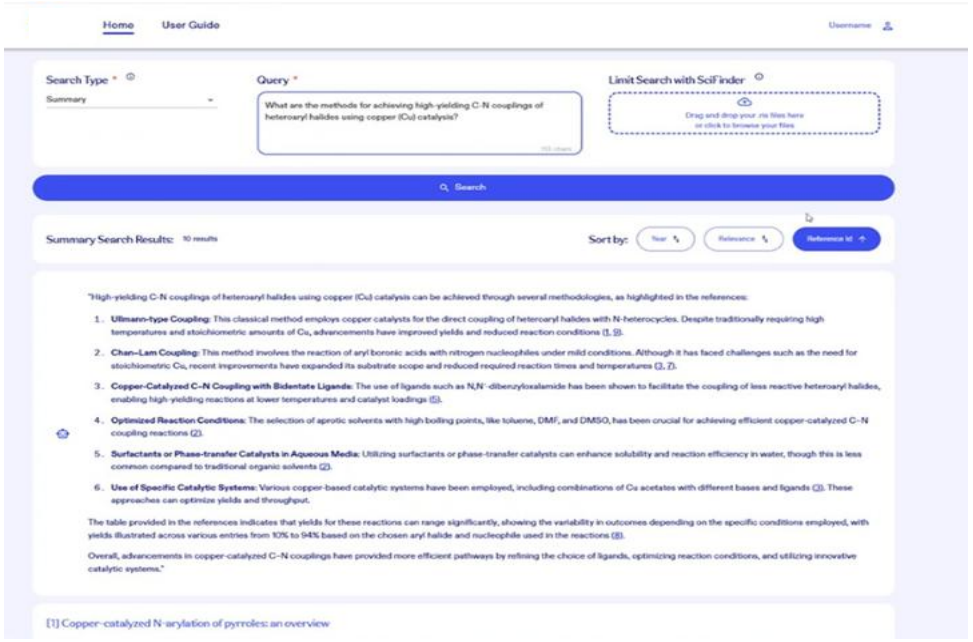
Agentic data science (FUTURE)

November 4-5, 2026

Real world RAG deployments in Pharma

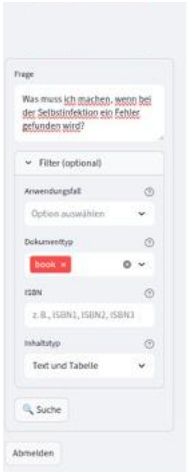
Scientific discovery platform

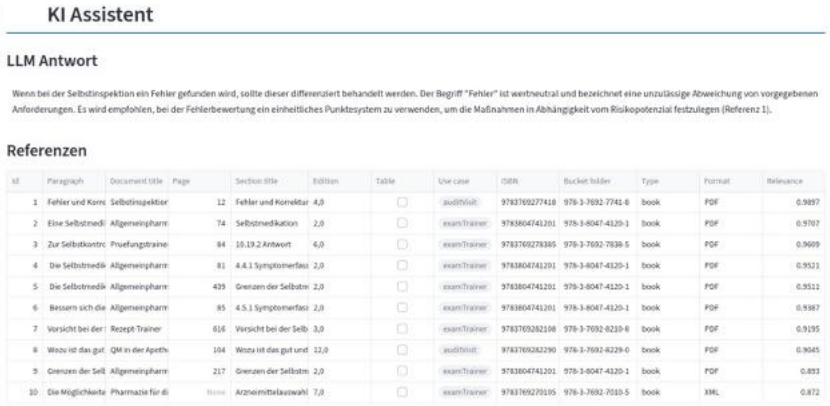
Help medicinal chemists keep track of advances in scientific literature



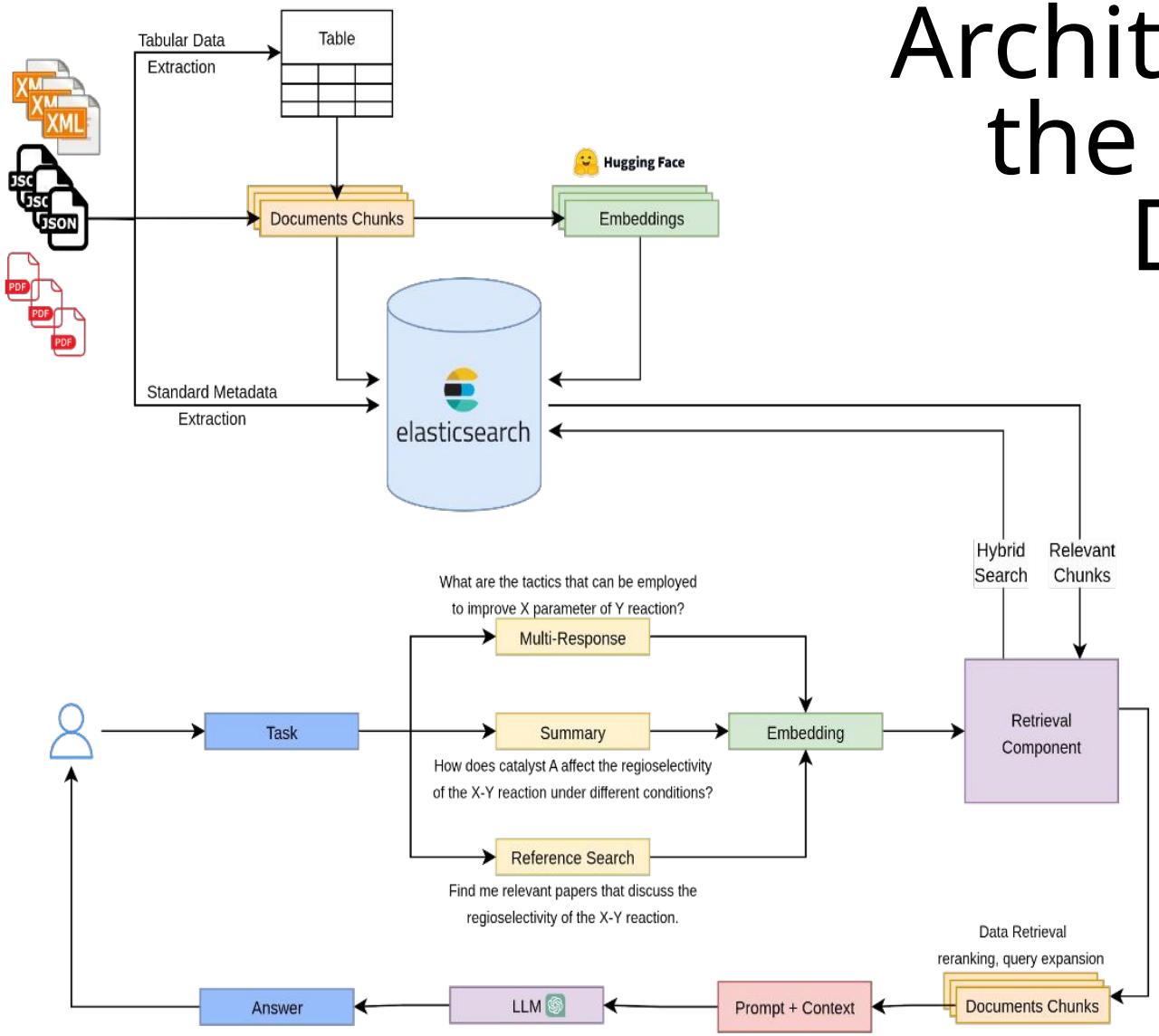
Unified RAG platform

Support multiple AI-driven business use cases (reference grounded QA)

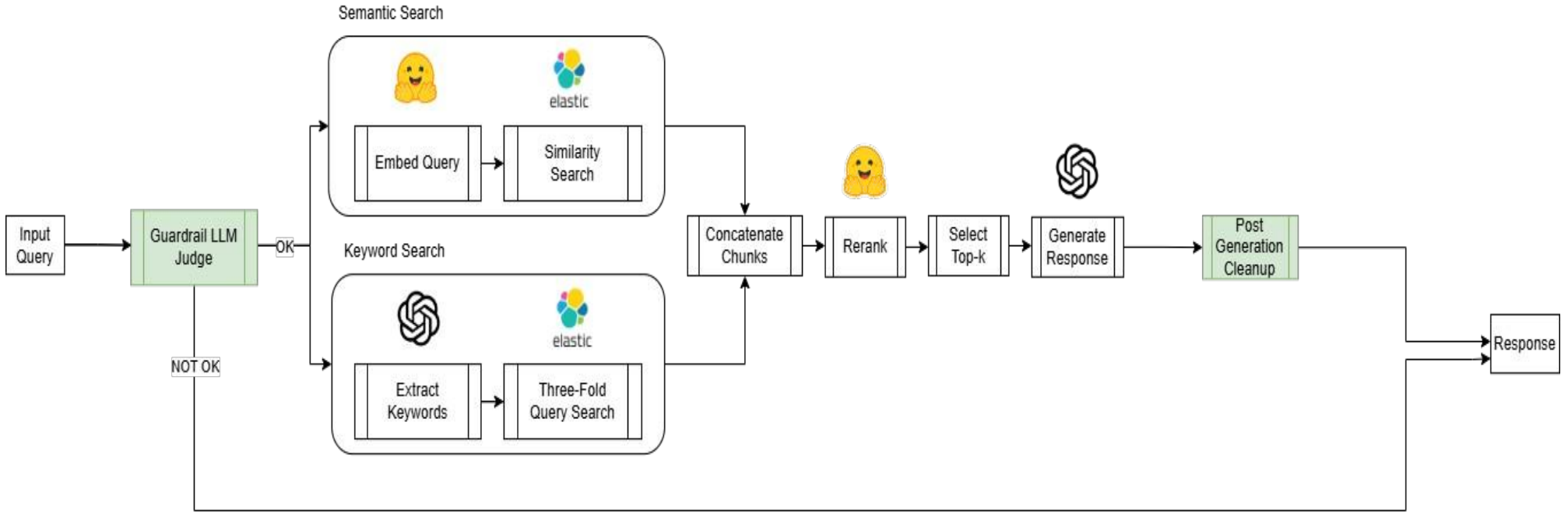




Architecture of the Scientific Discovery Platform



Unified RAG Platform Operational Architecture



Lego AI Assistant



What is Lego AI Assistant?

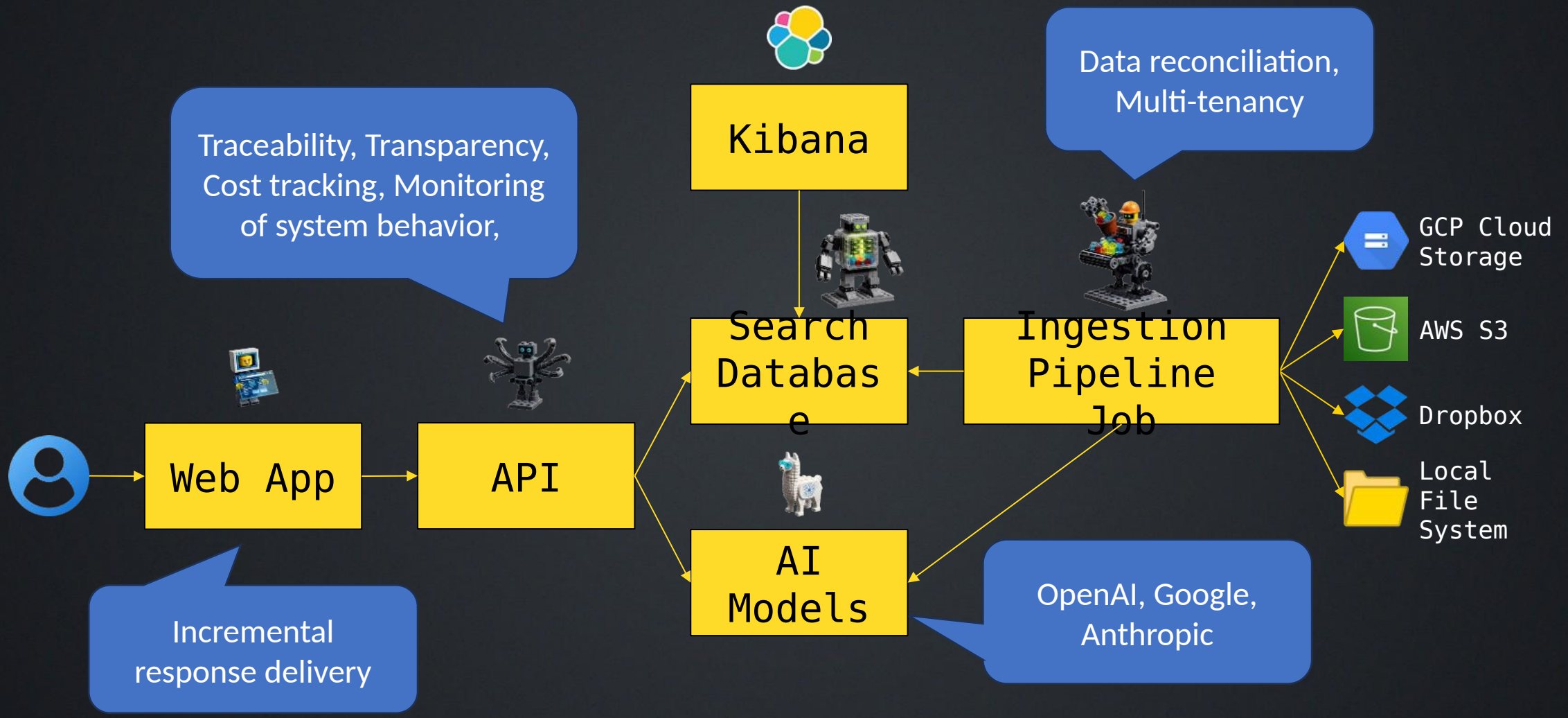
Goal:

Provide a reusable, modular RAG platform for building AI assistants with domain-specific knowledge.

Scope:

- E2E pipeline: collection → ingestion → retrieval → reasoning → UI
- Modular: Each component runs independently
- Configurable: Behavior controlled via JSON configs
- Model-Agnostic: Plug & play models without code changes
- Deployment-flexible: On-premise or GCP
- Documented: Clear technical documentation per module
- Secure Configuration: Sensitive variables in .env
- Dev Containers: Reproducible development environments

High-Level Architecture



* All components are **Dockerized** and **environment-configurable**.

FUTURE

Facilities for Unified Next-generation Wind Tunnel Testing and Research in Europe

Project details

6/2026 – 05/2030

15 million Euros

HORIZON-CL5-20025-04-D5-08

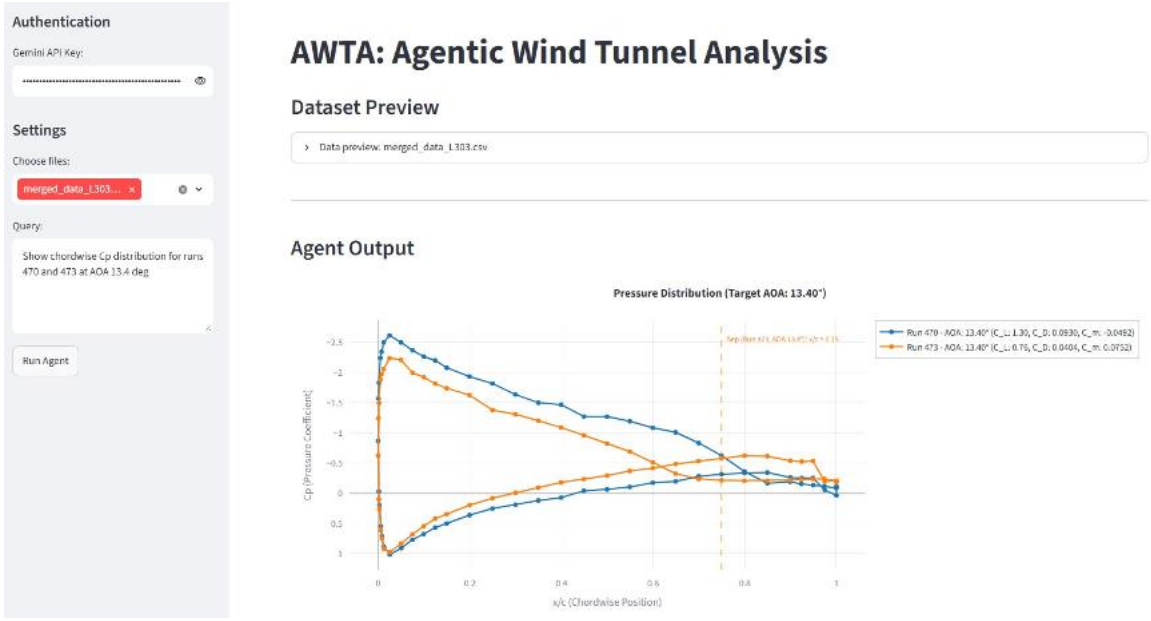
- Next generation testing capabilities in strategic EU wind tunnels

Strategic EU Wind tunnel partners

- DE-NL Wind Tunnels (**NLR**, DLR, DNW)
- Office National d'Etudes et Recherches Aérospatiales (ONERA)
- European Transonic Wind tunnel (ETW)

Medoid AI's role

No code AI assistant for analyzing wind tunnel test data



FUTURE

Facilities for Unified Next-generation Wind Tunnel Testing and Research in Europe

Demo paper

A Tool-Augmented Agent for Wind Tunnel Data Analysis

Ege Kots[1], Tatiana Passali[1,2], Francesco Martone[3], Carmelo Izzo[3], Grigorios Tsoumakas[1,2]

[1]Aristotle University of Thessaloniki, Greece
[2]Medoid AI, Greece,
[3] Italian Aerospace Research Centre (CIRA), Italy
Correspondence: ekotsb@csd.auth.gr; scpassali@csd.auth.gr

Video

Demo

Abstract

Recent advancements in sensor technology have fundamentally shifted the primary challenge in wind tunnel testing from data acquisition to data processing. Where test campaigns once yielded megabytes of static force balances, modern digital diagnostics routinely generate terabytes of unsteady flow-field data. Traditional analysis remains highly manual, requiring domain experts to process complex telemetry and compute aerodynamic coefficients. We present AWTA, a web-based interactive system that automates applied aerodynamics data science tasks based on a specialized agentic

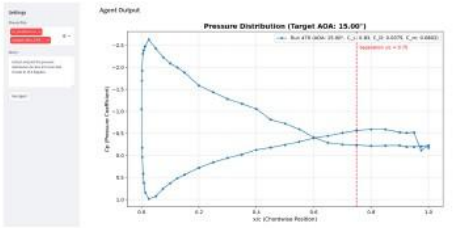
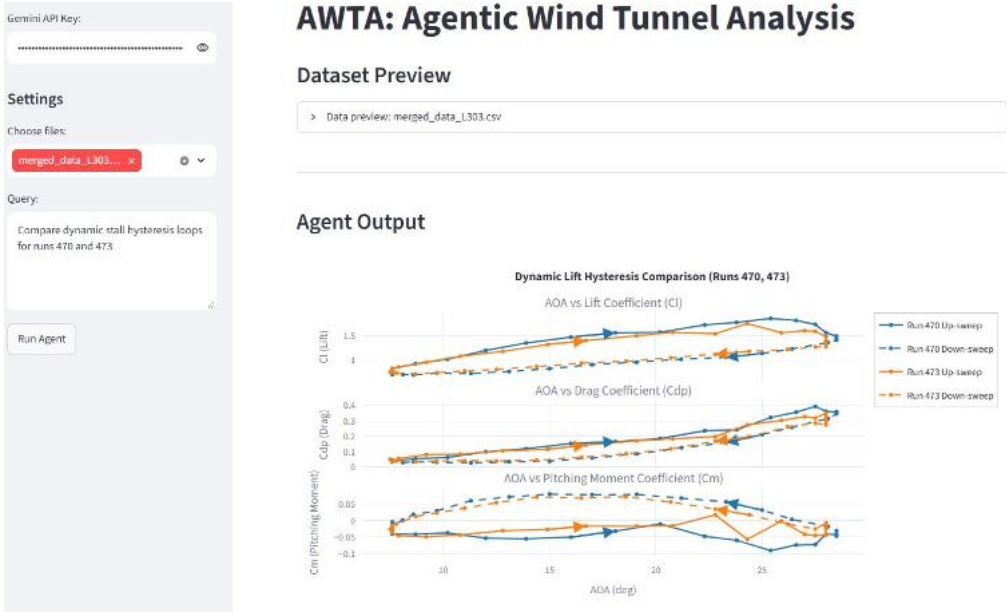


Figure 1: The interactive Streamlit-based user interface of AWTA. The system processes a natural language query (left panel) to generate an analytical visualization, such as a pressure distribution plot with automated boundary layer separation detection (right panel).

Medoid AI's role

No code AI assistant for analyzing wind tunnel test data



A unified framework for continual, explainable and embodied General-Purpose Artificial Intelligence

Project details

09/2026 – 08/2030

15 million Euros

HORIZON-CL4-2025-04-DIGITAL-EMERGING-07:

- Enhanced Learning Strategies for General-Purpose Artificial Intelligence

WP3: Continual & Evolutionary Learning Engine

Validation suite of lifelong learning benchmarks

Specialist model ecosystem

Multi-agent RL orchestrator

Continual learning pipeline that mitigates catastrophic forgetting, detects drift and consolidates knowledge

Evolutionary optimization toolkit

ARISTOTLE UNIVERSITY OF THESSALONIKI

Medoid AI





Thank You

Grigorios Tsoumakas

greg@csd.auth.gr

greg@medoid.ai