Building an AI-CRIS Platform: Multilingual, Contextual, and Predictive Research Data Access

Ecenaz Usta
Tugkan Tuglular
Gultekin Gurdal
Ahmet Semsettin Ozdemirden
Otmane Azeroual

The Problem

1. Expert-Only & Rigid Interfaces

THE BARRIER: Traditional CRIS platforms rely on faceted navigation and static keyword filters.

THE FLAW: They heavily assume that users already know complex relational data structures just to locate information.

2. Friction-Heavy & Multi-Step Workflows

THE BARRIER: Simple, natural questions like "Who has the highest citation growth?" cannot be typed directly.

THE FLAW: Finding an answer forces users through multiple tedious manual clicks, translations, and spreadsheet exports.

3. Isolated Data Silos & Static Reporting

THE BARRIER: Complex or multi-dimensional queries demand direct SQL database access or custom IT support.

THE FLAW: Operational dependency creates heavy cognitive overload and makes data inaccessible to non-technical stakeholders.

Our Purpose

Bridge the gap between complex research data and end-users by building an AI-CRIS platform that prioritizes intuitive, bilingual and context-driven exploration.

Democratizing Data Access

Contextual Intel

Verifiable Results

Our Solution: GCRIS AI Assistant

Bilingual

English & Turkish native support

Text-to-SQL

Natural language → PostgreSQL

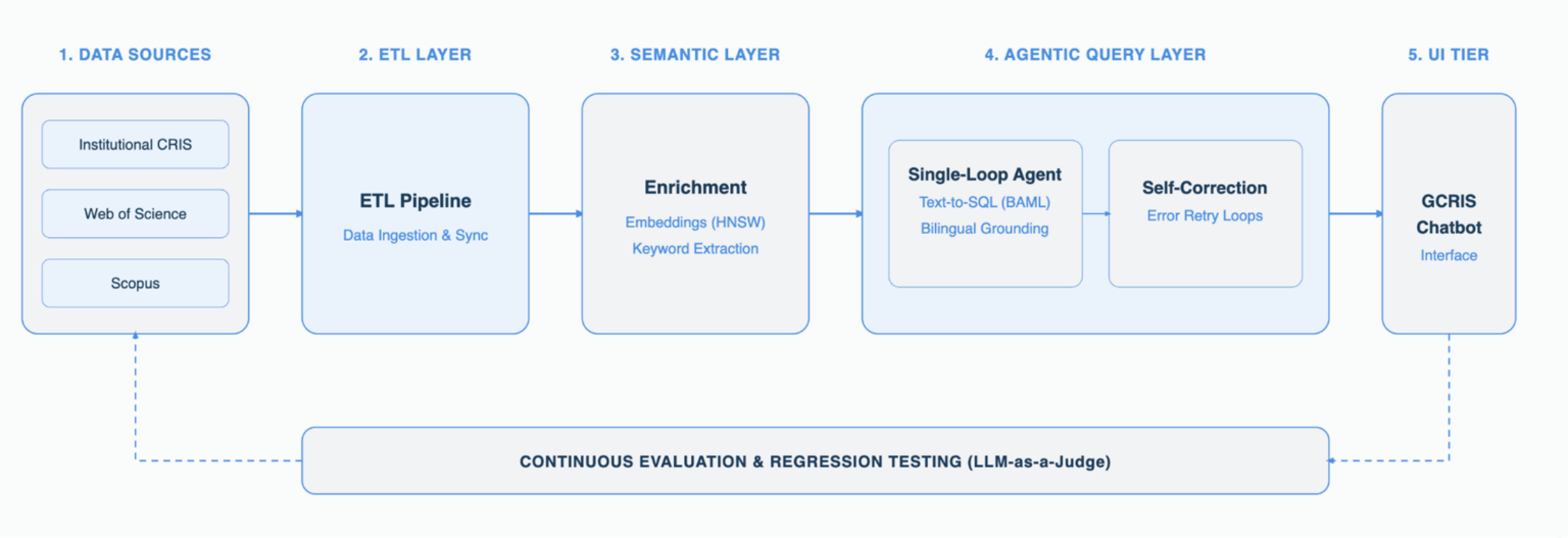
Secure

Parameterized, schema-grounded

Dynamic UI

Charts, tables via SSE streaming

Core Architecture



Data Engineering & Semantic Enrichment

Processing & ETL

Automated Ingestion

Kotlin/Spring Boot backend for high-concurrency data fetching from GCRIS, Scopus, and WoS.

Performance Optimization

GCC -O3 optimized shared libraries (via JNI) for lightning-fast text normalization and abstract concatenation.

Relational Grounding

Structured data storage in PostgreSQL, ensuring 100% data integrity for SQL execution.

The "Expertise" Brain

Bilingual Enrichment

Python agents extracting domain keywords and summaries in both Turkish & English (BAML-orchestrated).

Vectorization

Generating 768-dimensional expertise embeddings using open-source models (Ollama / Nomic).

Sub-second Retrieval

HNSW Indexing via pgvector for rapid semantic search and researcher discovery.

The Reasoning Engine: Agentic Text-to-SQL

Step-by-Step Logic (The Process)

1. **Intent Interpretation**
   User goals and language preferences identified.
2. **Semantic Grounding**
   Concepts mapped to specific database schemas.
3. **SQL Generation**
   BAML-orchestrated query synthesis via LLM APIs.
4. **Security & Whitelist Validation**
   Hard-checks ensuring read-only & authorized access.

The Self-Correction Advantage

- **Autonomous Refinement**
  If validation fails, agent analyzes error fingerprint.
- **Iterative Feedback**
  Continuous loops automatically re-generate the query.
- **Performance Leap**
  Valid SQL rate increases from 68.2% to 91.3%.
- **Efficiency**
  92% of failures resolved within max 3 cycles.

91.3%	92%	+23.1pp
Valid SQL Rate	Failures Resolved	Accuracy Lift
After autonomous self-correction	Within ≤ 3 cycles max	Compared to 68.2% baseline

Continuous Evaluation & A/B Testing

Evaluation & Guardrails

1. **Continuous Benchmarking**
   Automated production logs evaluation against golden datasets.
2. **Ground-Truth Comparison**
   User feedback, errors, and latency logs update benchmarks.
3. **The Guardrail Effect**
   Eval Agents instantly block anomalous query structures.

A/B Testing Framework

- **Live Traffic Routing**
  Safely splits user cohorts to evaluate prompt optimizations.
- **Parallel Benchmarking**
  Simultaneously measures cost, latency, and quality drift.
- **Shadow Deployments**
  Tests updated models in background before full rollout.

500+
Integration Tests
Executed per system deployment

94%
Success Rate
Overall system operational accuracy

3
Models Benchmarked
Evaluated in parallel workflows

Infrastructure Economics & Scaling Thresholds

Phase 1: Cloud API Sweet Spot

Proprietary LLM Deployment (Claude Sonnet 4.6)

Low Volume Efficiency
Highly cost-effective for pilot phases
Optimal: < 700 queries/month

Financial Profile (CapEx)
$0 upfront infrastructure investment
Pay-as-you-go: ~$0.08 / query

Token Transparency
Avg. 16,954 prompt tokens/query
Driven by complex schema injection

BREAK-EVEN
~700 Q / Month

Inflection Threshold: Cumulative token expenses cross over local hardware deployment costs.

Phase 2: Self-Hosting Transition

Financially Sustainable Open-Source Infrastructure (Grok 4.1 Fast / DeepSeek V3.2)

16x Inference Cost Cut
Slashes operational budgets
Annual API Savings: ~$620

Strict Data Sovereignty
No sensitive metadata or metrics
Ever leave institutional perimeter

Consortium Break-Even
RunPod RTX 4090 node deployment
Highly feasible split across 5+ orgs

Example Outputs

Generate a narrative CV for Gültekin Gürdal in PDF format.

The **Narrative CV for Gültekin Gürdal** has been generated successfully! 🎉

Your PDF is ready to download. Here's a summary of what's included:

Detail	Value
Name	Gültekin Gürdal
ORCID	0000-0001-7259-8134
Scopus ID	55512630900
WoS ID	G-1580-2015
Total Scholarly Outputs	121
Articles	17
Open Access Publications	119
Scopus Citations	21
Scopus H-Index	3
WoS H-Index	2

The PDF contains **7 sections** covering the researcher's academic profile, publication metrics, indexing details, and more.

📩 Download Narrative CV – Gültekin Gürdal

Show Gültekin Gürdal's publications by year | Compare with other researchers in the same department | Show citation trends for Gültekin Gürdal

Generate a narrative CV for Gültekin Gürdal in PDF format.

- **Actionable Workflows**
  On-demand PDF generation and administrative reporting for users.
- **Intelligent Follow-ups**
  Anticipates user needs and provides clarifying questions by guessing the research intent.

Example Outputs

Give the open access rates of publications in computer engineering for the last 5 years.

Here are the open access rates for Computer Engineering publications over the last 5 years:

Year	Open Access	Total Publications	Open Access Rate
2021	36	48	75.0%
2022	39	47	83.0%
2023	27	53	50.9%
2024	16	45	35.6%
2025	12	38	31.6%

Key observations:

- The open access rate peaked in **2022 at ~83%**, which was the highest in the 5-year period.
- There has been a **significant downward trend** since 2022, dropping to ~31.6% in 2025.
- The total number of publications has remained relatively stable (38–53 per year), so the decline in open access rate reflects fewer publications being made openly accessible rather than a drop in total output.
- 2025 data may be partial as it is the current year.

Give the open access rates of publications in computer engineering for the last 5 years.

- **On-Demand Report Generation**
  Instantly exports complex query results and structured research data into production-ready PDF, CSV, and Markdown formats.

- **Administrative Integration**
  Streamlines academic decision-making by producing high-integrity documentation directly aligned with institutional institutional reporting pipelines.

Example Outputs

Who is the rector of IZTECH, and what is his/her bibliometric information?

Rector of IZTECH: Prof. Dr. Yusuf Baran

The current Rector of İzmir Institute of Technology (IZTECH) is **Prof. Dr. Yusuf Baran**.

Bibliometric Profile

Below are his key research metrics from the GCRIS database:

Metric	Value
ORCID	0000-0002-1056-4673
Total Scholarly Outputs	185
Articles	94
Open Access Publications	146
Scopus Citations	4,428
WoS Citations	4,093
Scopus H-Index	33
WoS H-Index	32
Scopus Indexed Publications	96
WoS Indexed Publications	156

Prof. Dr. Baran has a strong international research presence with over **4,400 Scopus citations** and an **H-index of 33**, reflecting a highly impactful publication record.

Who is the rector of IZTECH, and what is his/her bibliometric information?

- **Cross-Domain Discovery**

When a query requires information outside the core CRIS schema, the agent dynamically triggers a custom Institution-Base Tool to identify non-CRIS external entities (e.g., locating the current Rector or university administration data).

- **Contextual Data Enrichment**

Once the external entity is retrieved, the agent automatically pivots back to the primary GCRIS Core Database, mapping this fresh context against live publication records, citations, and researcher profiles to deliver a fully enriched, validated synthesis.

Limitations

Agentic Reasoning Latency — UX and Processing Speed

THE CHALLENGE

Multi-step reasoning loops — intent interpretation, Text-to-SQL generation, schema validation, and self-correction introduce delays.

15-45 sec end-to-end latency

OUR MITIGATIONS

- **Thinking Scratchpad:** Agent's internal thoughts streamed.
- **Polymorphic Retry Cap:** Max 3 cycles limits infinite loops.

Context Window Pressure — Token & Memory Management

THE CHALLENGE

Injecting raw database schemas, complex entity relationships, and session history exerts massive pressure on the context window.

Avg. 15,954 tokens / query

OUR MITIGATIONS

- **Dynamic Schema Pruning:** Injects only relevant tables.
- **Context Compression:** Sliding-window strips metadata.

Unpredictable Cloud API Cost Scaling — The Budget Ceiling Problem

THE CHALLENGE

Relying solely on proprietary cloud models scales monthly token costs linearly with volume, creating an uncapped financial burden.

Linear cost growth with volume

OUR MITIGATIONS

- **700-Query Cost Ceiling:** Breaks even with on-premise hardware.
- **Infrastructure Hard Cap:** Blueprint for self-hosting on RTX 4090.

Future Work

1. Ingesting Horizon Europe Data & Funding Insights

THE VISION: Integrating global funding and project repositories directly into the data processing pipeline.

THE MILESTONE: By ingesting Horizon Europe project data, call histories, and successfully funded consortia structures, the agent will dynamically cross-reference internal institutional metadata to recommend specific open funding calls tailored to a researcher's active expertise profile.

2. Advanced AI-Driven Collaboration Recommender

THE VISION: Transitioning from simple keyword matching to predictive partnership building.

THE MILESTONE: Implementing predictive multi-agent systems that analyze historical co-authorship networks, active citation trajectories, and newly ingested Horizon data to surface advanced interdisciplinary opportunities before consortium search begins.

3. Nationwide Multi-Institutional Consortium Scaling

THE VISION: Breaking data silos across the national higher education landscape.

THE MILESTONE: Scaling the infrastructure into a unified, multi-tenant nationwide consortium architecture. The platform will concurrently query and synthesize anonymous metadata across all Turkish universities, protecting data sovereignty via shared GPU clusters.

Conclusion

1. Democratizing Complex CRIS Architecture

Making complex database hierarchies and relational tables fully accessible to non-technical leaders and researchers through intuitive, natural language querying—removing the technical barrier between data and insight.

2. Continuous Data Verification & Enrichment

Deploying a dedicated verification layer that eliminates AI hallucinations by cross-referencing raw CRIS data, cleaning anomalies, and enriching metadata to ensure a high-fidelity research intelligence platform.

3. Actionable Academic Workflows (PDF, CSV, MD)

Beyond chat: A full-scale reporting engine capable of generating instant, multi-language official academic reports in diverse formats, perfectly integrated into existing institutional documentation and administrative pipelines.

Thank You For Your Interest