AI 工具箱

數據分析與預測



AI Agentic Systems for Civil Services



Dr. Wai Mo Suen

CEO, ClusterTech Limited

Professor, Washington University

Honorary Professor, Chinese University of Hong Kong

6 Oct, 2026

聯科集團 CLUSTERTECH

AI Toolbox

01/3

Introducing ClusterTech

Providing HPC/AI Services to HK/Mainland/East Asia since year 2000



Our Clients (partial)

Our clients include **20+** companies in the Fortune 500, **100+** national and governmental departments, **150+** universities (including all top 20 universities in mainland China) and **100+** top research institutes (including **30+** institutes of the Chinese Academy of Sciences), on High Performance Computing/Intelligence Computing, AI, Big Data solutions.

Banking, Finance & Insurance

Education & Research

Telecommunications

Energy & Utilities

Retail & Tech

Public Sector

Aero Space

Others

Our HK Government Clients (26 departments, hundreds of projects):

1. Regulatory Intelligence: Hong Kong Monetary Authority, Financial Services and the Treasury Bureau, Environmental Protection Department, Digital Policy Office, Architectural Services Department, Buildings Department, Planning Department
2. Investigation Intelligence: Immigration Department, Hong Kong Police Force, Securities & Futures Commission of Hong Kong
3. Public Safety: Hong Kong Observatory, Airport Authority Hong Kong, Hospital Authority
4. Predictive Services: Environmental Protection Department, Electrical and Mechanical Services Department, Hong Kong Observatory
5. Security: Hong Kong Police Force, Correctional Services Department
6. Service Automation: Social Welfare Department, Efficiency Office
7. Education: Hong Kong Productivity Council/HKJC



AI Development of ClusterTech

2001 → Present

聯科集團 CLUSTERTECH

AI Toolbox

Both quantitative data intelligence and text intelligence now evolve into agentic production AI

Data Mining & ML

Prediction • Risk • Quant

2001 → **Fintech ML**

HSBC and banks: FX prediction using Bayesian neural nets

2002 → **Risk & Pricing**

DBS and banks: Risk management and derivative pricing with stochastic PDE + NN

2003 → **Quant Funds**

Neural-network time-series prediction for quant trading, decentralized finance

Now **Agentic Data AI**

Agents wrap predictiv models, call tools and validate outputs

Text Mining & NLP

Language • Entities • RAG

2011 → **PRISMA (ICT Award)**

Cantonese NLP, text classification and sentiment analysis

2020 → **ML Deployments**

Entity extraction, relationship mapping, predictive analytics

2023 → **LLM Integration**

RAG pipelines, private cloud deployments, secure environments

Now **Agentic KB/RAG AI**

KB/RAG agents retrieve, reason, draft and route to review

AGENTIC TREATMENT LAYER

Data intelligence and text intelligence thru agent-orchestrated workflows

- Data Agents – forecast, price, detect risk, call analytics tools
- Knowledge / RAG Agents – retrieve, reason, extract, explain and report
- Review Agents – route outputs for human review and exception handling
- Compliance Agents – validate, audit, cite and enforce controls
- Software Engineering / Coding Agents – code generation, legacy code understanding, API integration, testing, documentation, DevOps automation and secure production deployment

Agentic data intelligence

Quantitative ML is wrapped with agents that select models, call data tools, run checks and produce risk-aware outputs.

Agentic text intelligence

NLP and LLM capabilities are grounded in KB/RAG agents that retrieve evidence, reason, draft and route to review.

Unified agentic production AI

Data and text streams converge into secure, compliant, auditable agent workflows.

From "ML models" and "LLM components" to "AI Agent–Enabled production systems"

Trusted • Data-secure • Compliant • Tool-using • Human-in-the-loop • Enterprise deployable

Data Agents **Knowledge / RAG Agents**

Review Agents **Compliance Agents**

Software Engineering / Coding Agents

Strategic message: the company's AI capability began with financial data mining in 2001 and text mining in 2011, now developing into agent-based systems for reliable enterprise production deployment.

聯科集團 CLUSTERTECH

AI Toolbox

02/3

Our AI Product:

Agentic AI System

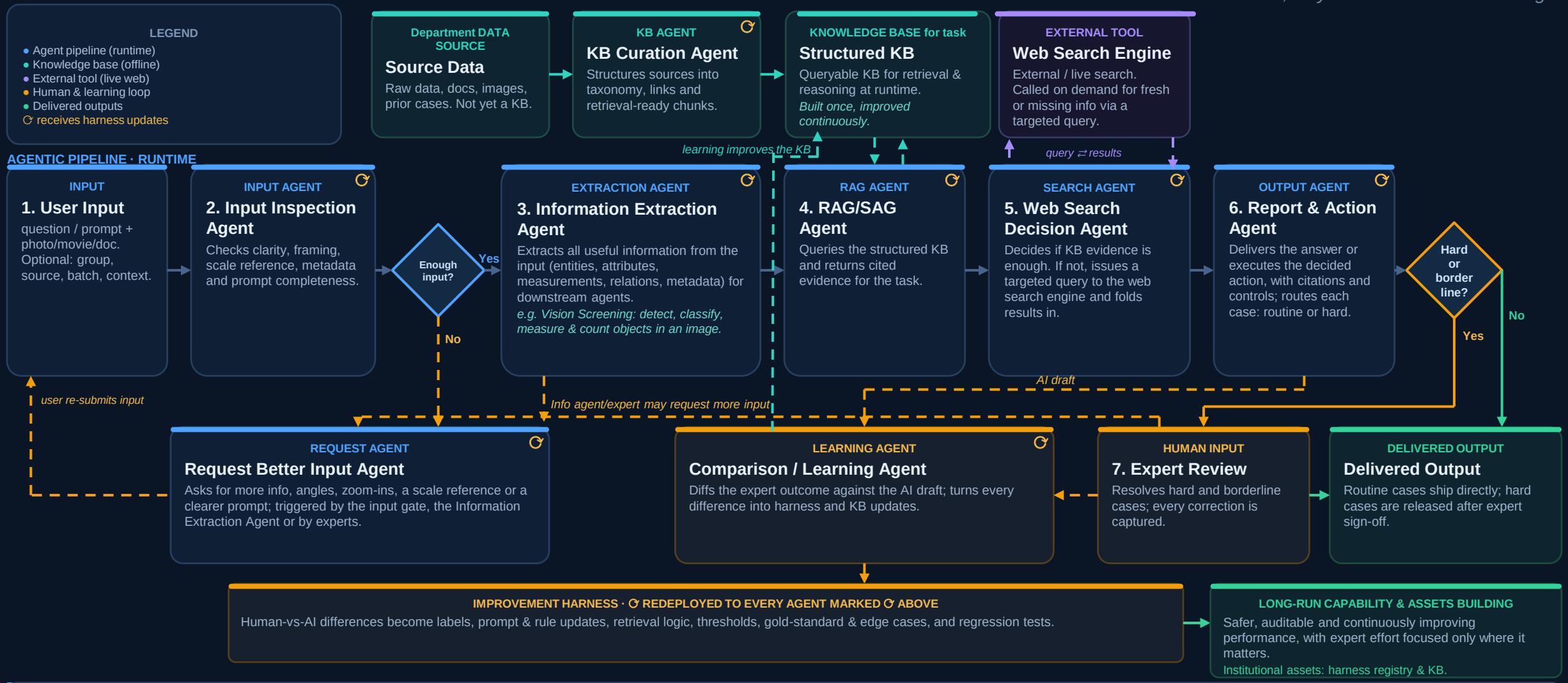
From Isolated AI tools to a
Reliable Orchestrated-Agentic System:
Reliability, Data Security, Local Deployment



Agentic Workflow System (Locally deployable depending on Data Security Requirement)

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



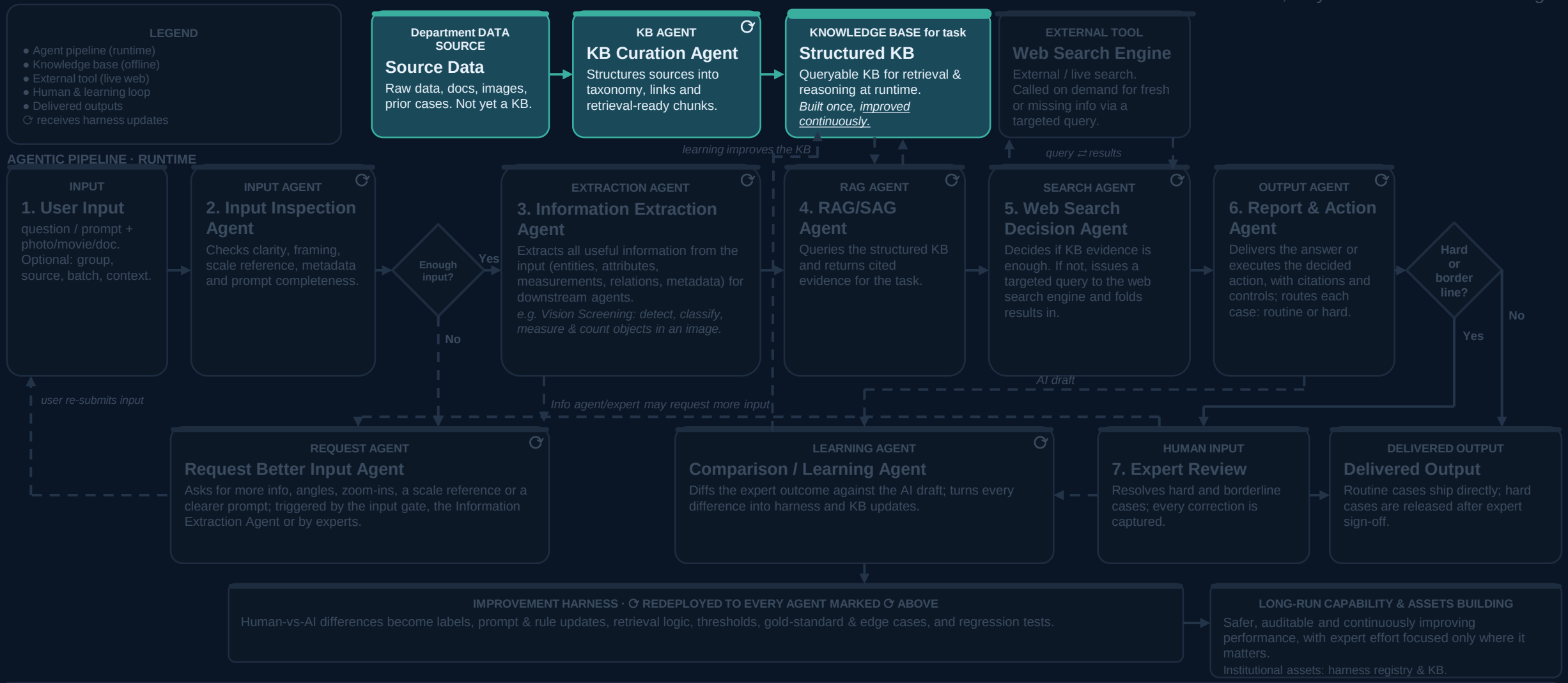
Core idea: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

Agentic Workflow System (Knowledge base construction)

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

AGENTIC FLOW

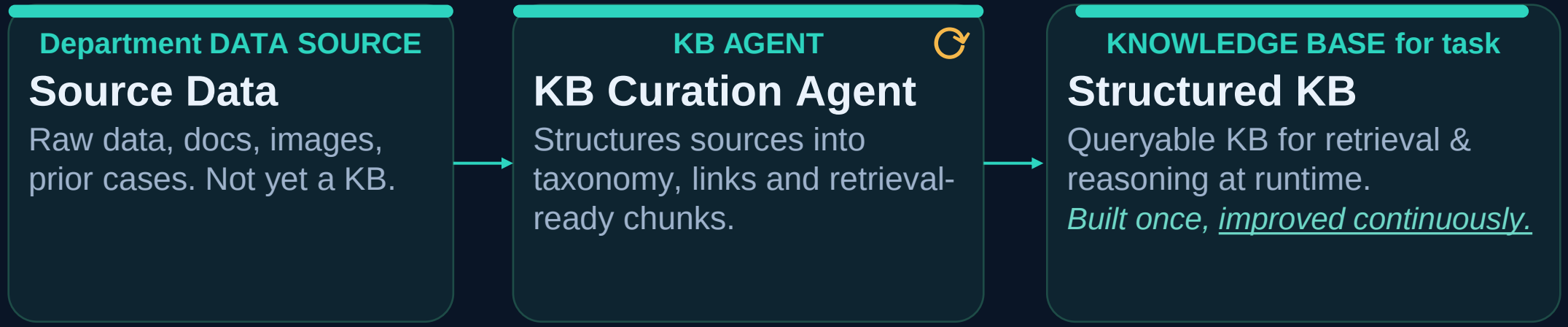
Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

Knowledge base construction (offline)

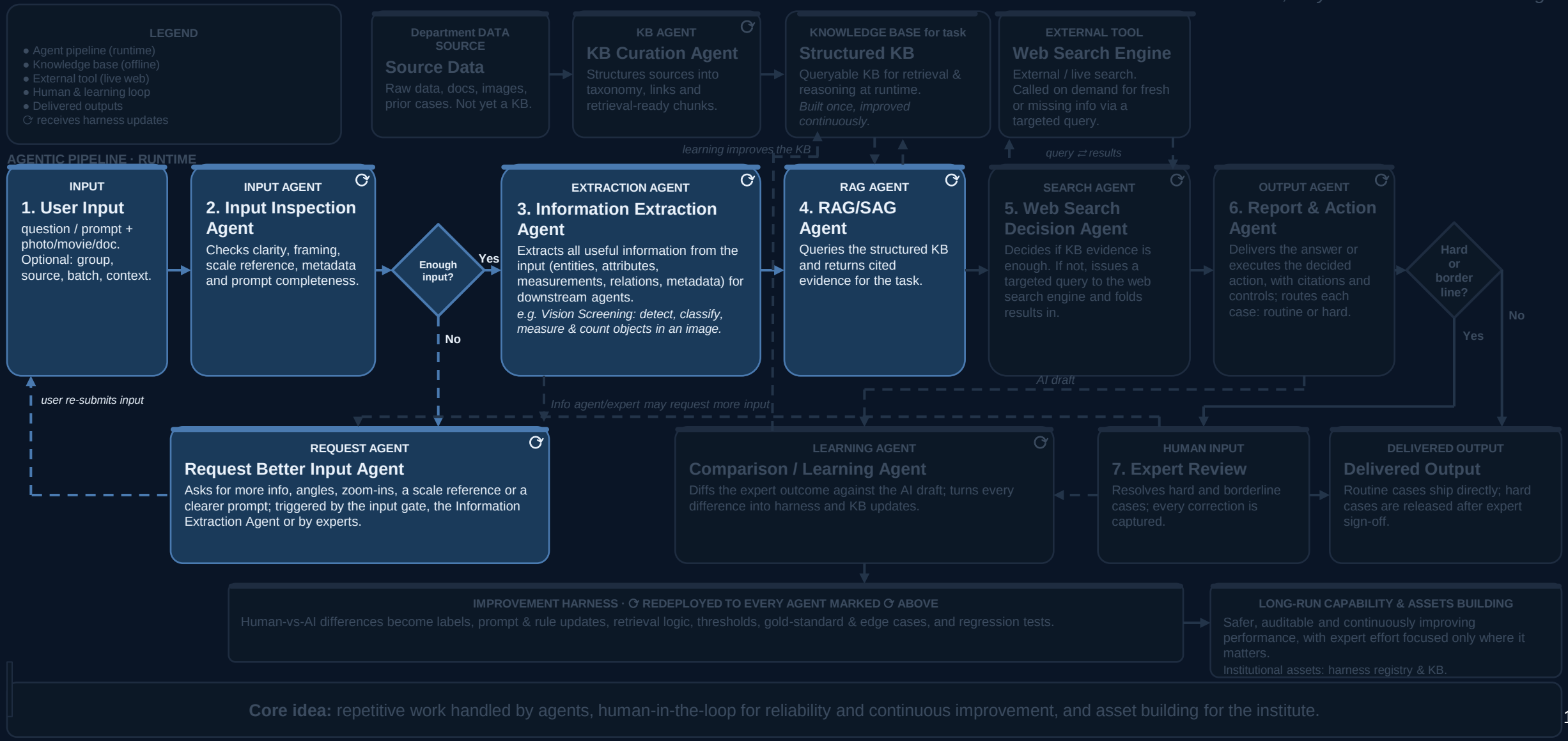
Same layout as the full diagram, zoomed in



Agentic Workflow System (understanding the input (runtime intake))

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

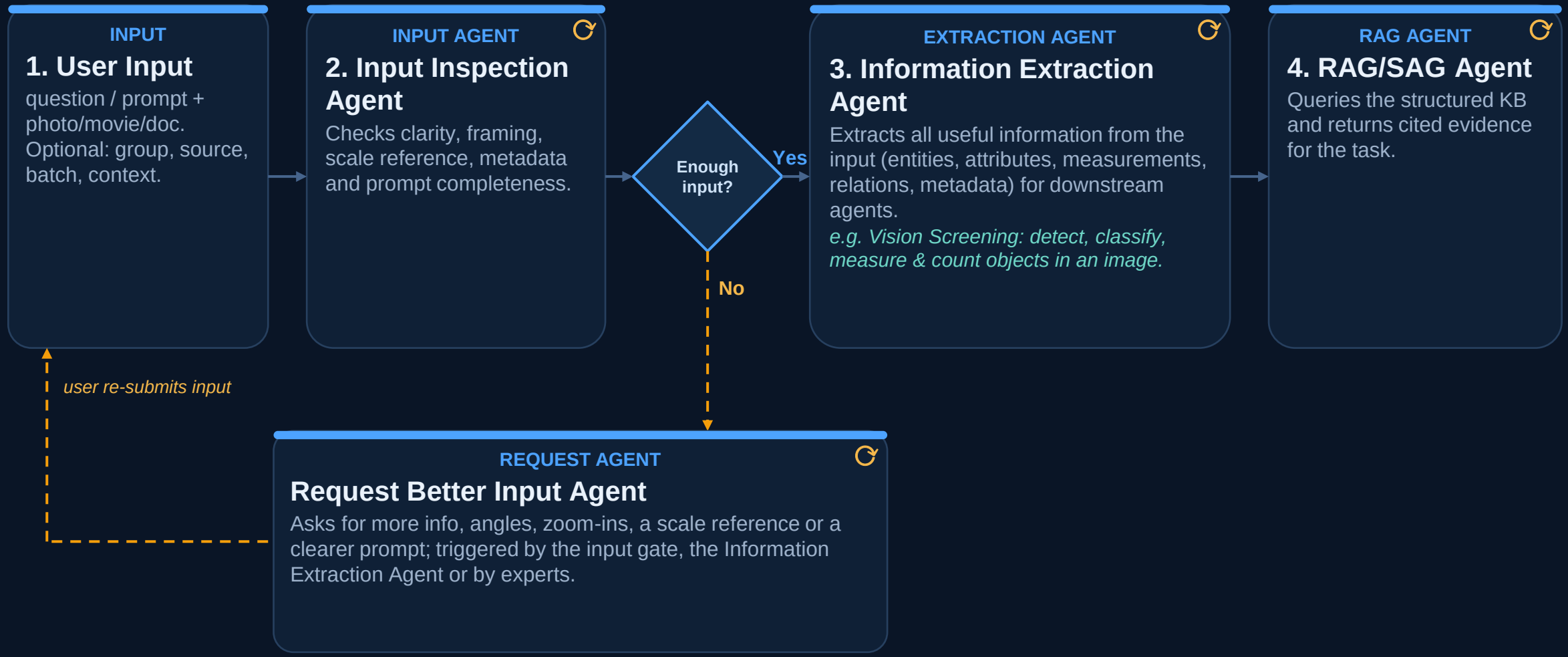
Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

Understanding the input (runtime intake)

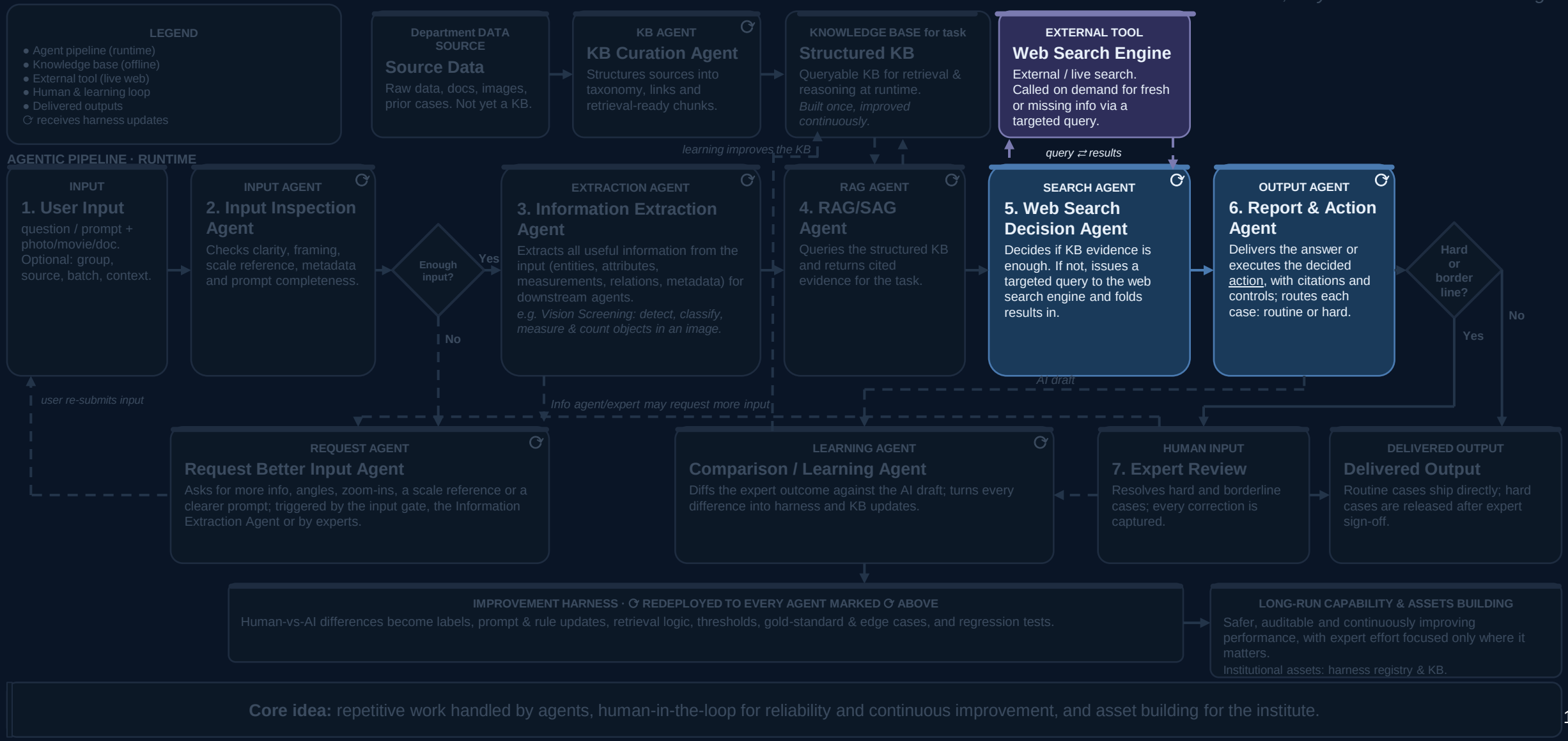
Same layout as the full diagram, zoomed in



Agentic Workflow System (Web-search augmentation, Routing Output)

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

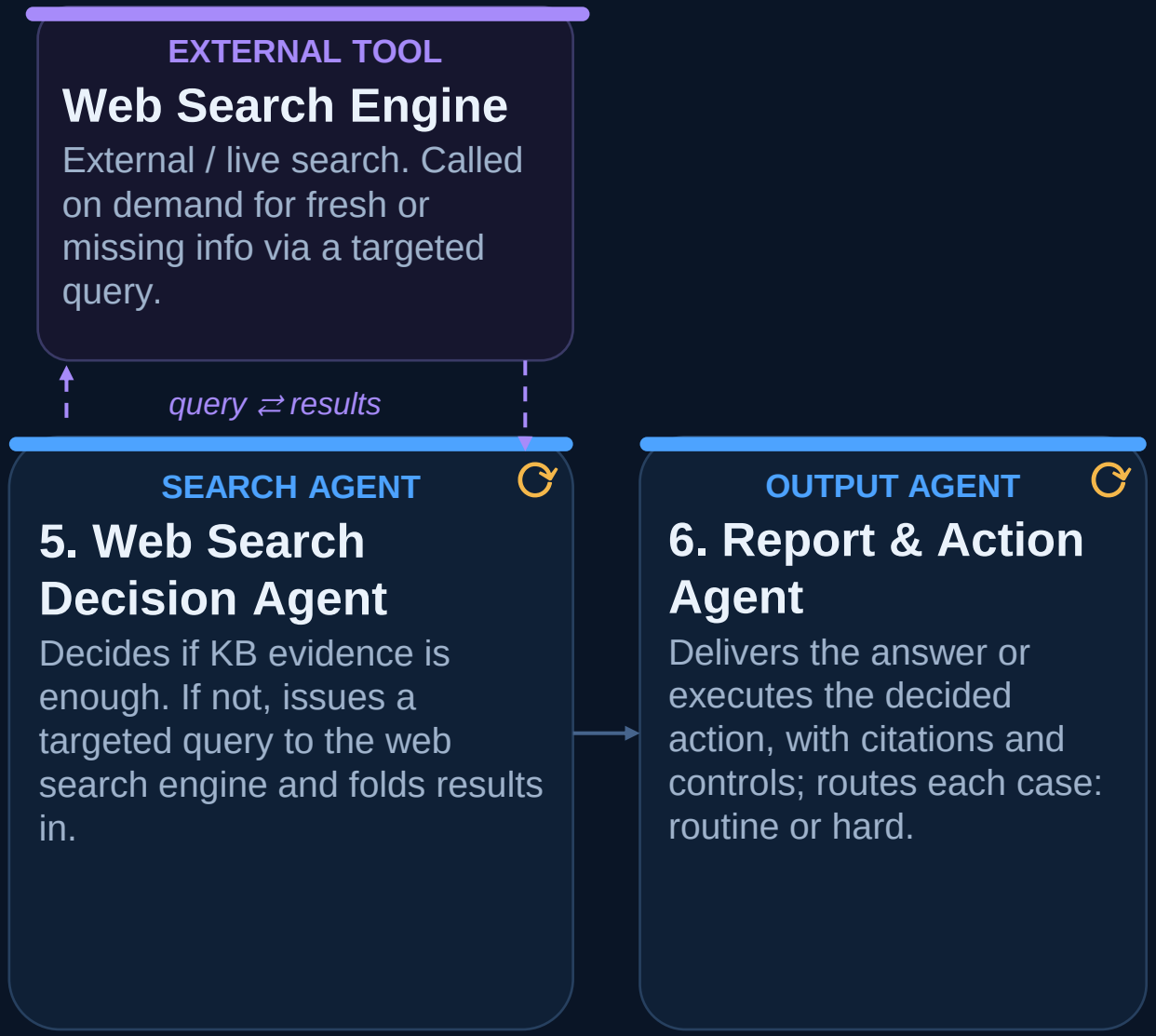
Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

Web-search augmentation, Routing Output

Same layout as the full diagram, zoomed in

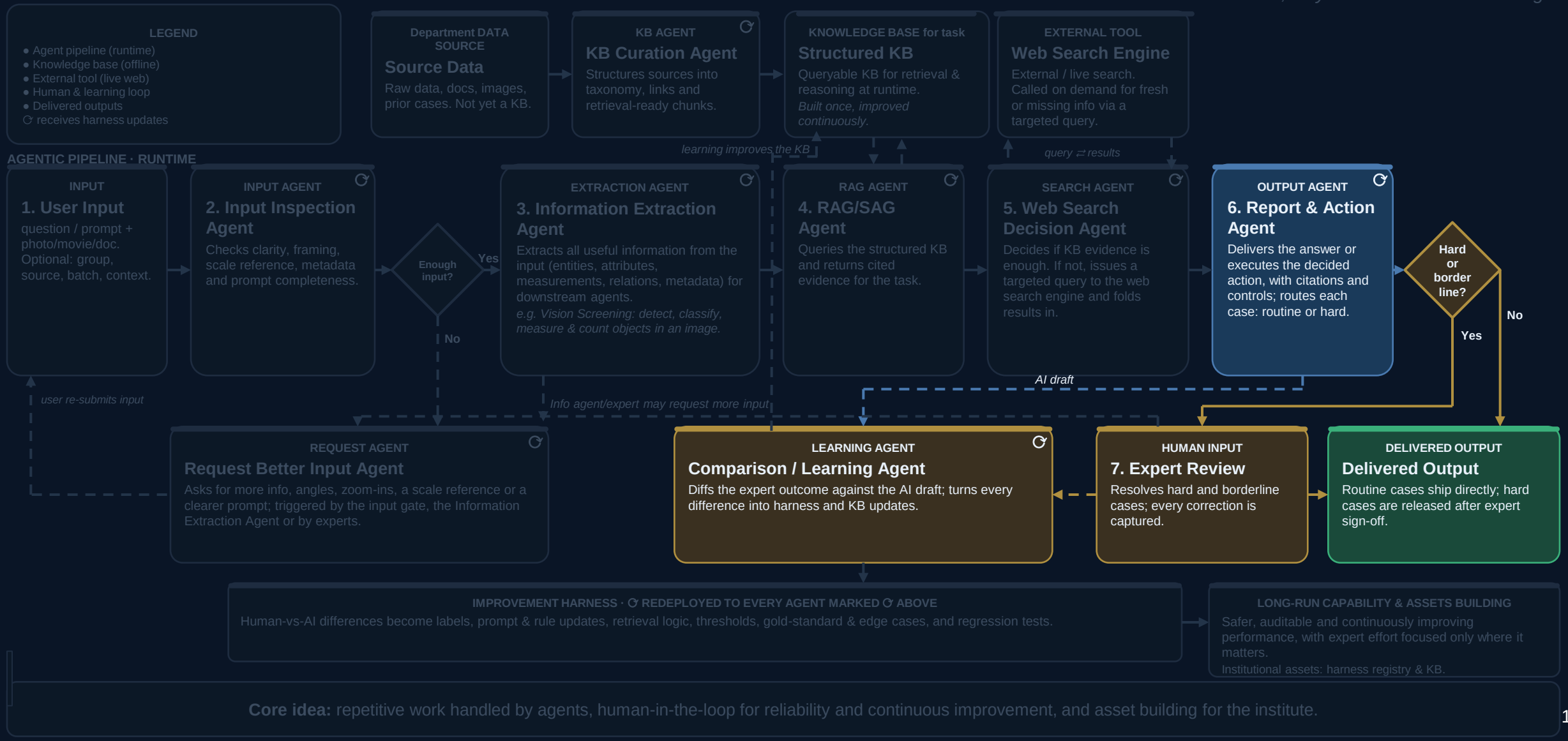


Agentic Workflow System (Decide, review, deliver & learn)

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

AGENTIC FLOW

Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

Decide, review, deliver & learn

Same layout as the full diagram, zoomed in

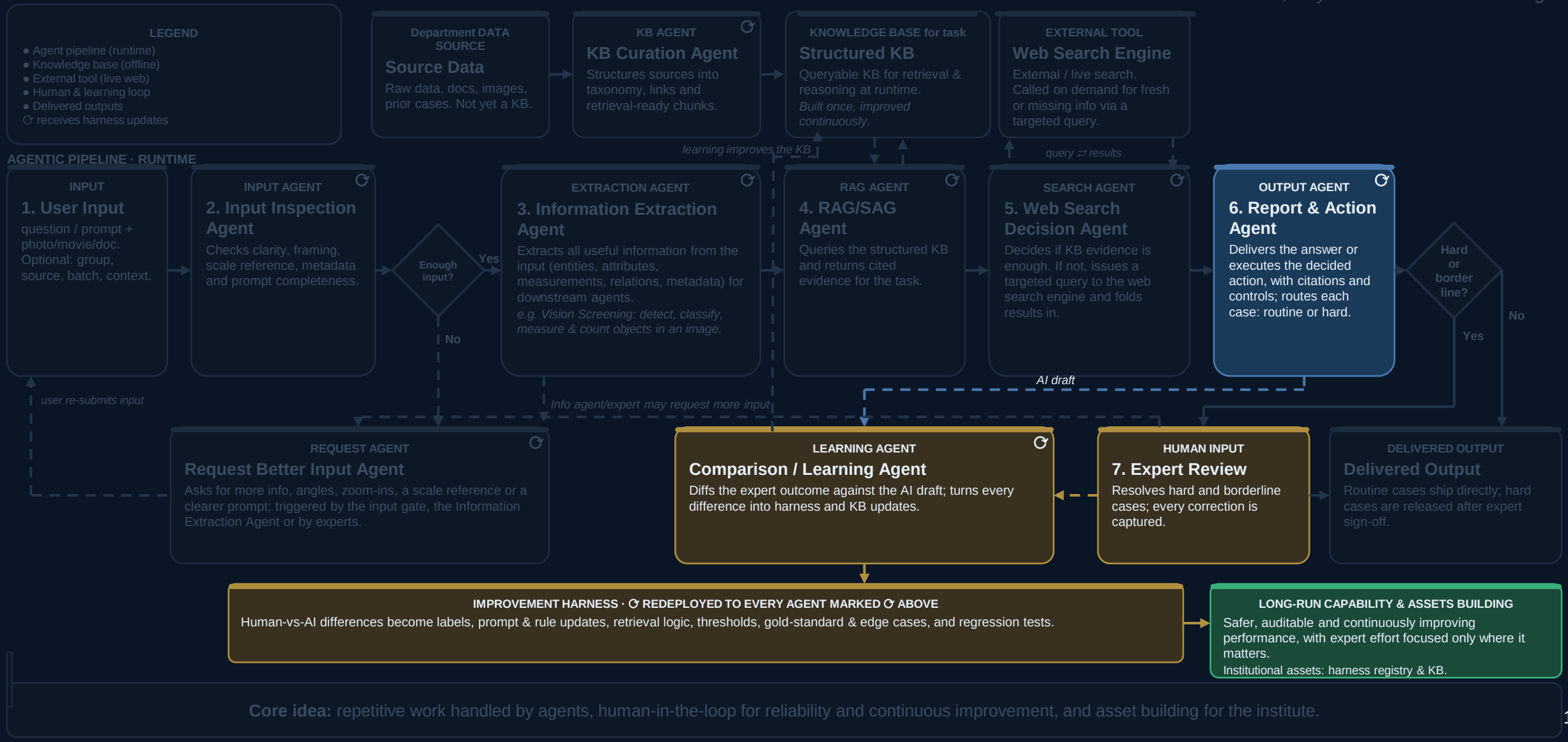


Agentic Workflow System

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

AGENTIC FLOW

Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

The improvement process

How human corrections turn into system-wide improvements

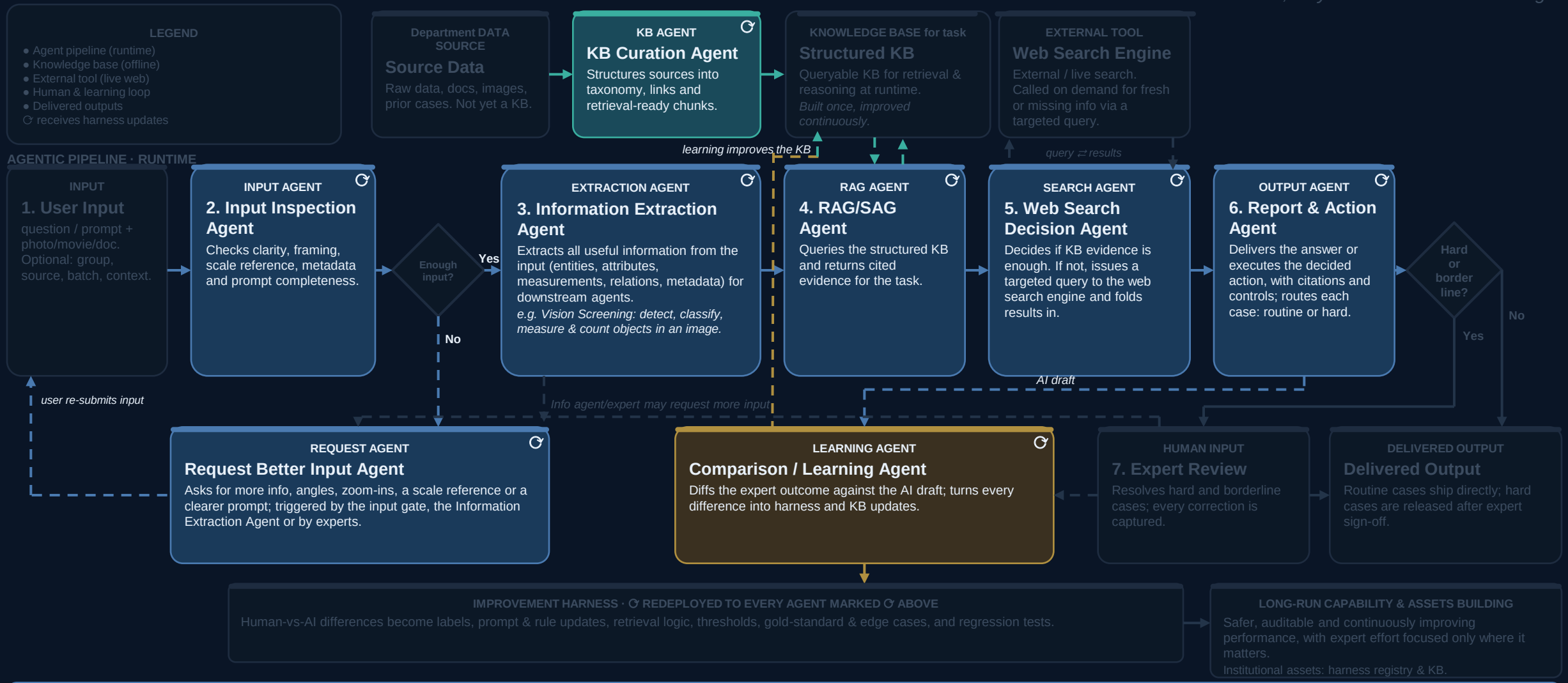


Agentic Workflow System

A general-purpose agentic pipeline with a human-in-the-loop improvement cycle

AGENTIC FLOW

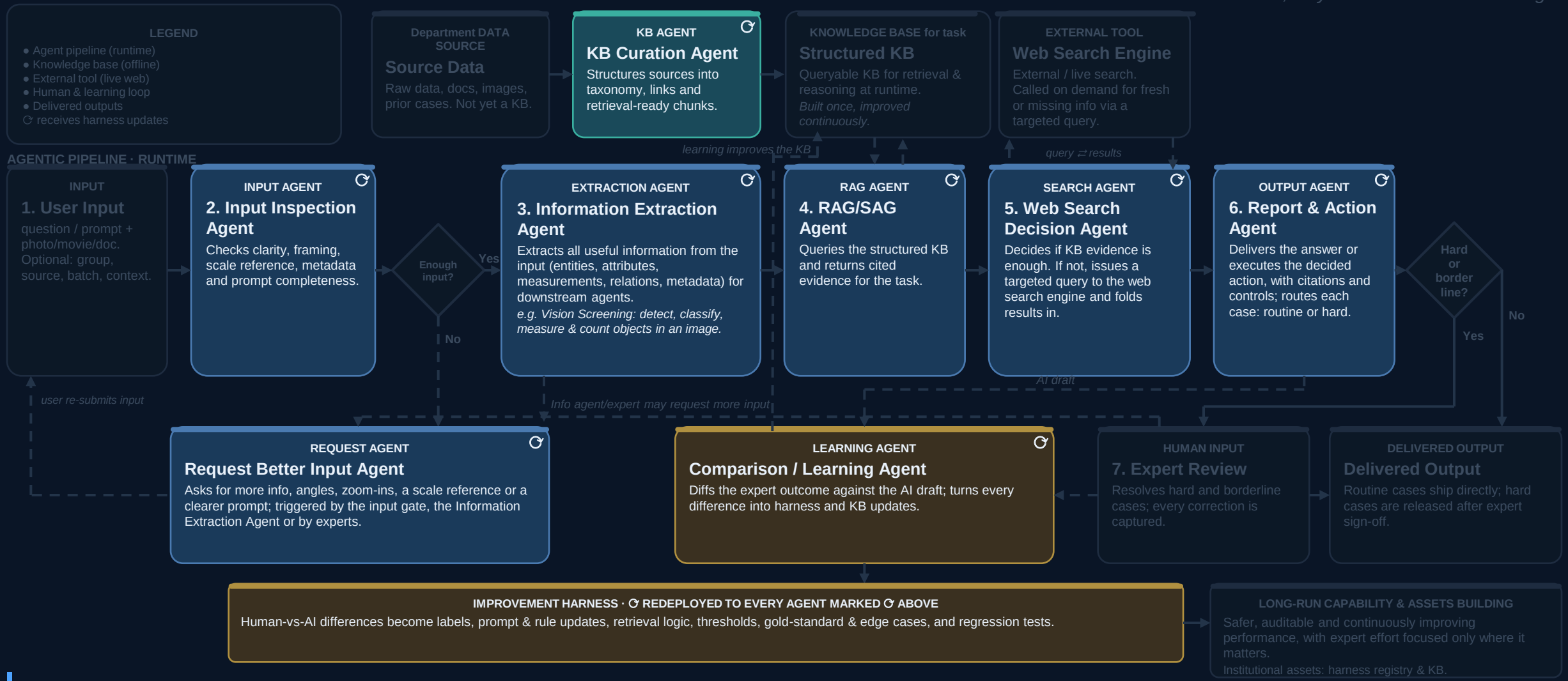
Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea 1: repetitive work handled by agents, human-in-the-loop for reliability and continuous improvement, and asset building for the institute.

Every Agent is LLM-Powered — Local Deployment for Data Security

Offline KB is curated and versioned; web search is live and on-demand, only when the KB is not enough.



Core idea 2: Every highlighted agent runs on LLM inference. The full LLM stack can be deployed locally on your infrastructure when data security requires it.

Core ideas

Core idea 1

- Agents handle the repetition, people retain accountability.
- Each tool can be upgraded conveniently with the latest tech development.
- Every usage/expert correction makes the system smarter, building a compounding institutional asset.

Core idea 2

- When data security requires it, the full LLM stack can run locally on your own infrastructure. No data leaves your environment.
- Full data sovereignty, by design.

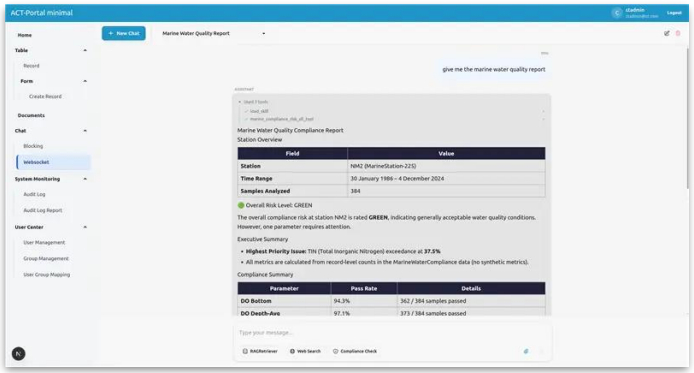
ClusterTech has deployed **1,000+** high performance computing clusters for hundreds of institutes, with experience in Nvidia (B200, H200, H100, H800, H20…), and domestic 华为910, 384超节点, 沐曦C500…

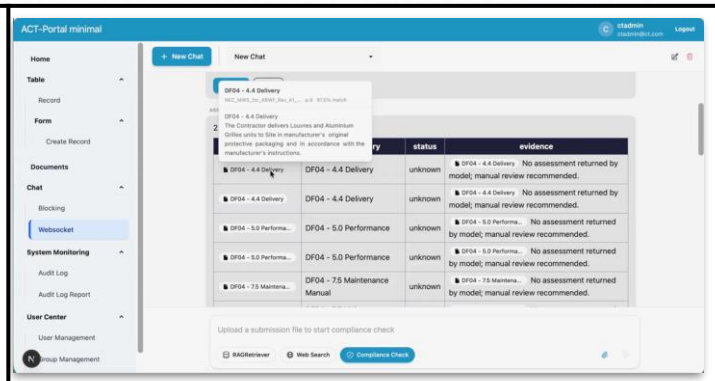
Choices of deployable LLM's: Kimi, GLM, SeepSeek, MiniMax, Qwen, Gemma, gpt-oss, Llama…depending on task requirements

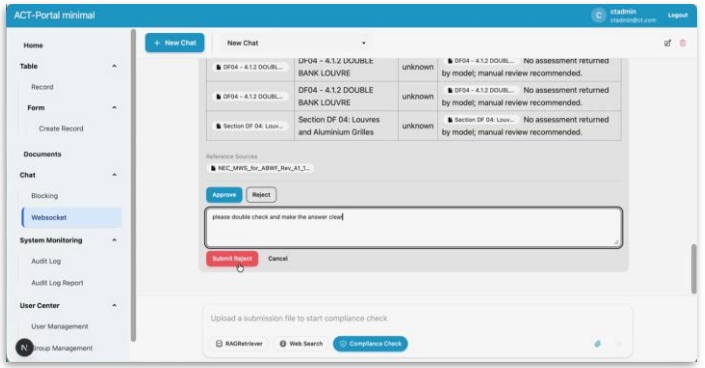
Agentic AI System Feature Samples

Chat Interface

1. Tool Calling Trace
2. Response with citations
3.Human-in-the-loop verification

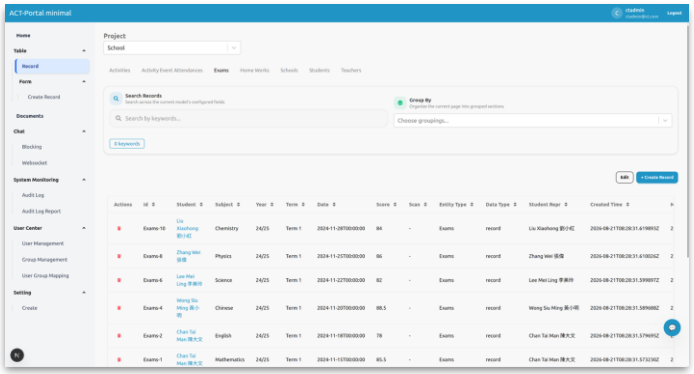


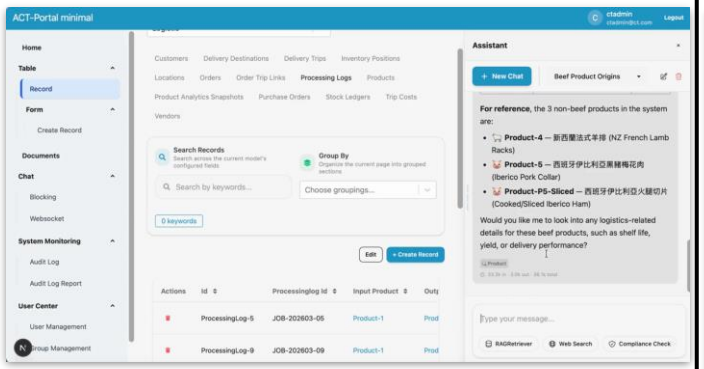


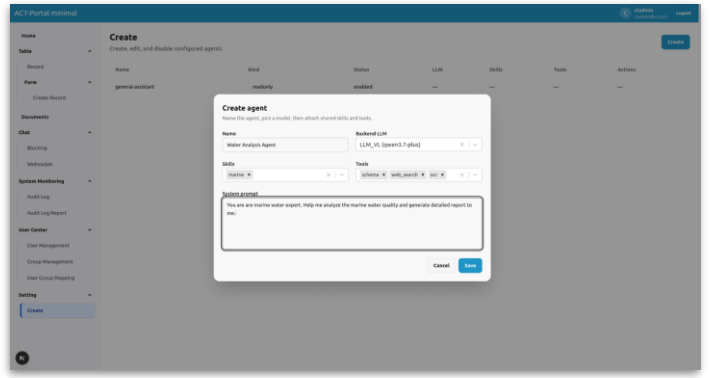


KB and Agent Management

1. KB Management
2. KB+Chat
3. Agent Management

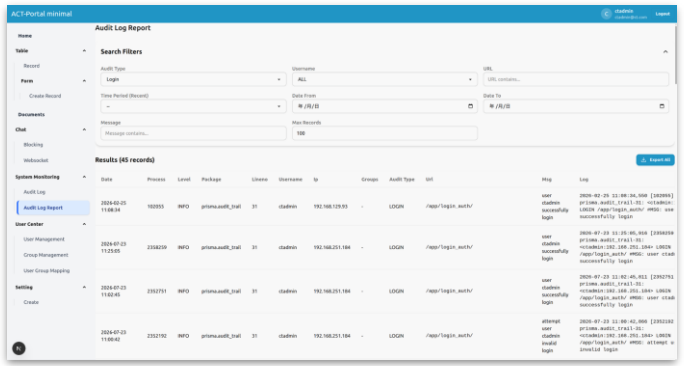


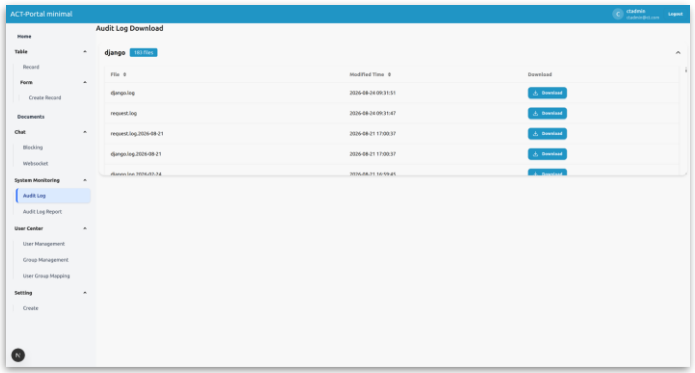


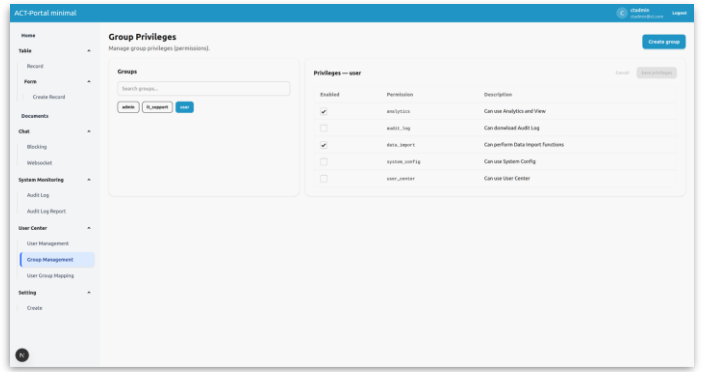


Audit Log and User Management

1. Audit Log
2. System Monitoring Report
3. User Privilege Management







聯科集團 CLUSTERTECH

AI Toolbox

03/3

Civil Services Use Cases



Safer, auditable, continuously improving performance - with expert attention focused where it matters.

Civil Services Core Demands – *the everyday reality of civil service*

Volume & Constraints

- High-volume repetitive work & document processing
- Cross-department information silos
- Staffing constraints under growing caseloads
- Strict privacy & confidentiality requirements

Quality & Accountability

- Rigid process compliance
  - every step governed
- Evidence-based decisions
  - every answer with citation
- High accuracy
  - public-facing errors carry real consequences
- Consistent service
  - same question, same answer

Governance & Institutional Memory

- Full auditability & traceability
  - who did what, when
- Multi-tier approval hierarchy
  - different sign-off levels
- Expert knowledge retention
  - experience must not walk out the door

Highlight - Anatomy of an agentic use case for civil services

Not one model, but an orchestrated team of specialist agents.
Every case that follows runs the same governed pipeline; only the lit agents change.

Orchestrated specialists 01
Ten specialist agents run one pipeline, not a monolith, so each step can be inspected and improved separately.

Cited evidence 02
Every answer carries sources from a versioned KB.

Confidence routing 03
Routine cases ship; hard cases go to Expert Review.

Actions, not just reports 04
Report & Action delivers the output or executes the step.

Continuous improvement 05
Comparison / Learning and the harness update agents.

Governed by design 06
RBAC, audit trail, PII control, private cloud, on-premises.

LEGEND Runtime agents · Knowledge, review and learning · Security Control

TRUST worthy BY DESIGN

- Cited
- RBAC
- Audit trail
- Private cloud
- Human-approved

Use Case 1 - Regulatory & Risk Intelligence

JOB TO BE DONE sense-making at scale: surface what matters, with sources attached

PAIN

Signals are buried in daily volume; manual monitoring delays escalation and leaves coverage uneven.

OUTCOME

A prioritized, risk-classified briefing where every item is traceable to its source.

AGENTS ON THIS CASE

- **KB CURATION** structured KB
- **INPUT INSPECTION**
- **INFORMATION EXTRACTION**
- **LINK ANALYSIS** entity resolution
- **RAG / SAG** cited evidence
- **WEB SEARCH**
- **REQUEST BETTER INPUT**
- **REPORT & ACTION**
- **EXPERT REVIEW**
- **COMPARISON / LEARNIN**

EXAMPLE INPUTS

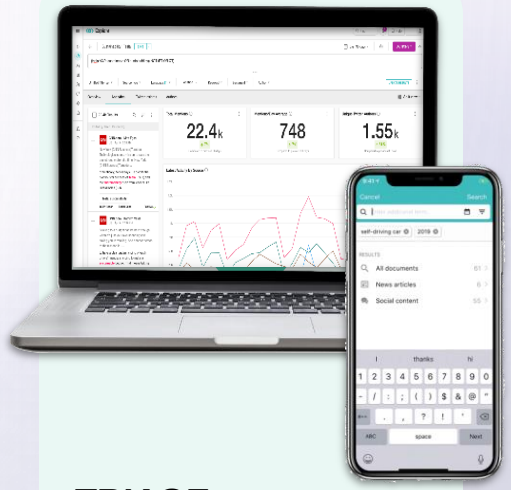
Regulatory circulars | financial news | social posts
Internal watchlists | entity and topic taxonomies

EXAMPLE OUTPUTS

Prioritized briefing | risk classification
Affected entities | cited source pack

WHY AGENTIC HERE

- Cited prioritization: RAG / SAG returns evidence with sources, so every alert can be defended.
- Live search only when the knowledge base is insufficient; the Web Search Decision Agent makes that call.
- Learns your risk appetite: expert corrections return through Comparison / Learning and the harness.



TRIAGE
hours to minutes

COVERAGE
multilingual

EVIDENCE
every alert cited

TRUSTED BY DESIGN

Cited | RBAC | Audit trail | Private cloud | Human-approved

——Use Case 2 - Application & Casework Review

JOB TO BE DONE do the busywork, keep judgment human

PAIN

Officers re-extract the same fields and chase completeness across formats and systems.

OUTCOME

A structured review pack with version-pinned policy citations; approval stays with the officer.

AGENTS ON THIS CASE

- **KB CURATION** structured KB
- **INPUT INSPECTION**
- **INFORMATION EXTRACTION**
- **LINK ANALYSIS** entity resolution
- **RAG / SAG** cited evidence
- **WEB SEARCH DECISION**
- **REQUEST BETTER INPUT**
- **REPORT & ACTION**
- **EXPERT REVIEW**
- **COMPARISON / LEARNING**

EXAMPLE INPUTS

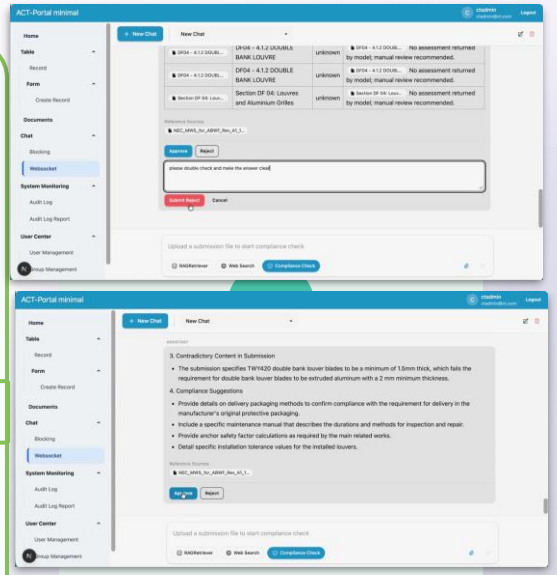
Application forms | PDF and Word attachments
Departmental guidelines | authorized records

EXAMPLE OUTPUTS

Case summary | missing-information flags
Cited rules and precedents | draft questions

WHY AGENTIC HERE

- Repeatable multimodal extraction: OCR, tables and stamps handled the same way on every case.
- Version-pinned cited rules, so a review can be defended against the policy in force at the time.
- A human approval gate with a full audit trail, and the pack improves case over case.



MANUAL WORK
much less of it

POLICY BASIS
every rule cited

ACCOUNTABILITY
human approval

TRUSTED BY DESIGN Cited | RBAC | Audit trail | Private cloud | Human-approved

Use Case 3 - Investigation & Link Analysis

JOB TO BE DONE cross-silo links you can defend

PAIN

Records, reports and media sit in separate silos; hidden relationships are hard to explain.

OUTCOME

A relationship map and timeline where every link carries evidence, confidence and a review trail.

AGENTS ON THIS CASE

- **KB CURATION** structured KB
- **INPUT INSPECTION**
- **INFORMATION EXTRACTION**
- **LINK ANALYSIS** entity resolution
- **RAG / SAG** cited evidence
- **COMPARISON / LEARNING**
- **EXPERT REVIEW**
- **REPORT & ACTION**
- **REQUEST BETTER INPUT**
- **WEB SEARCH DECISION**

EXAMPLE INPUTS

Case reports and statements | entity lists
Authorized records | images and extracted text

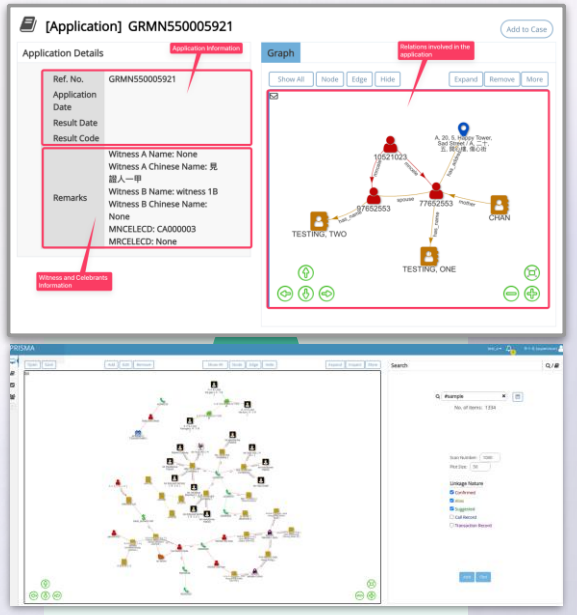
EXAMPLE OUTPUTS

Relationship map | event timeline
Linked-case candidates | expert-review queue

WHY AGENTIC HERE

- Alias and identity resolution runs as a named agent, not a hidden heuristic.
- Every link is evidence-backed and confidence-scored, so leads survive challenge.
- Expert verification and chain of custody, inside private-cloud deployment and RBAC.

TRUSTED BY DESIGN Cited | RBAC | Audit trail | Private cloud | Human-approved



REACH
links across silos

EVERY LEAD
evidence-backed

CONFIDENCE
scored, reviewable

Use Case 4 - Public Enquiry & Service Triage

JOB TO BE DONE a consistent, cited, safe front door

PAIN

Enquiries arrive in varied language and leave out details; staff repeat lookups and route by hand.

OUTCOME

Consistent cited guidance in the citizen's language, with low-confidence cases handed to staff.

AGENTS ON THIS CASE

- **KB CURATION** structured KB
- **INPUT INSPECTION**
- **INFORMATION EXTRACTION**
- **LINK ANALYSIS** entity resolution
- **RAG / SAG** cited evidence
- **WEB SEARCH DECISION**
- **REQUEST BETTER INPUT**
- **REPORT & ACTION**
- **EXPERT REVIEW**
- **COMPARISON / LEARNING**

EXAMPLE INPUTS

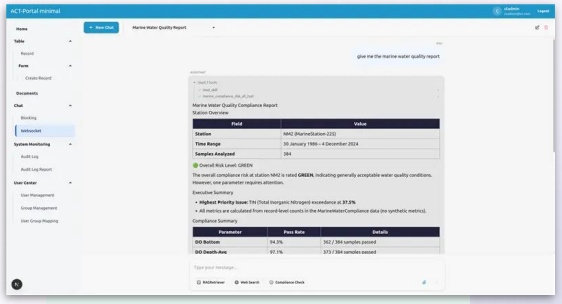
Enquiry in Traditional Chinese or English
Uploaded form or photo | service guides | rules

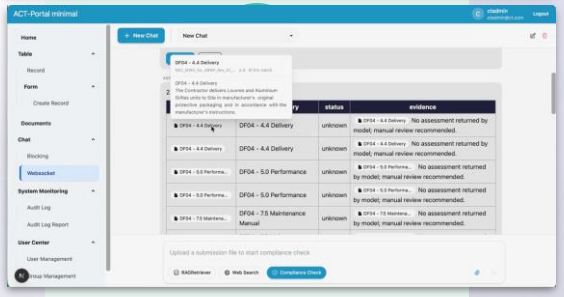
EXAMPLE OUTPUTS

Cited answer | document checklist
Recommended channel | staff hand-off summary

WHY AGENTIC HERE

- Consistent cited guidance around the clock from a versioned knowledge base.
- Safe by design: clarify when intent is unclear, refuse when it should refuse.
- Confidence-based hand-off to staff, and every hand-off feeds the learning loop.





ROUTINE
answered and cited

HARD CASES
escalated in context

AVAILABILITY
consistent, 24/7

TRUSTED BY DESIGN Cited | RBAC | Audit trail | Private cloud | Human-approved

One pipeline, many missions

REGULATORY & RISK INTELLIGENCE | **APPLICATION & CASEWORK REVIEW** | **INVESTIGATION & LINK ANALYSIS** | **PUBLIC ENQUIRY & SERVICE TRIAGE**

ONE GOVERNED PIPELINE — specialist agents | cited evidence | human approval

THE IMPROVEMENT FLYWHEEL

every reviewed case upgrades every agent

AGENTS DO THE WORK routine cases ship → **EXPERT REVIEW** resolves hard cases → **COMPARISON / LEARNING** AI draft vs expert → **IMPROVEMENT HARNESS** prompts and rules → **REDEPLOYED TO EVERY AGENT** and the KB

BETTER NEXT TIME

START ONE PILOT

One use case | one fixed timebox | agreed evidence and accuracy criteria | agreed human approval points

TRUSTED BY DESIGN

Cited | RBAC | Audit trail | Private cloud | Human-approved

WHY IT COMPOUNDS

Every correction becomes an asset in the knowledge base and harness.

NEXT STEP

one use case,
one fixed timebox.

Our IT Infrastructure Service Team

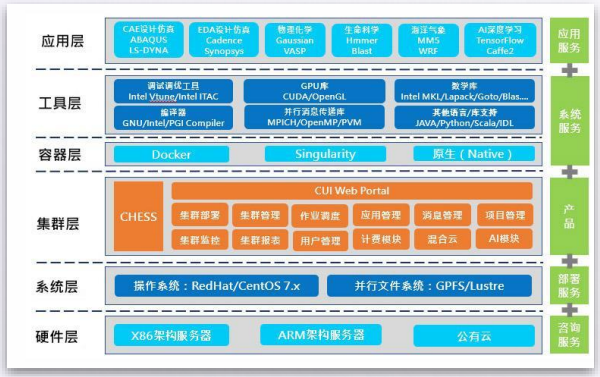
On-the-Ground Infrastructure Engineering Capacity in Hong Kong

Private Cloud
Self-hosted or client on-premises deployment; full data sovereignty.

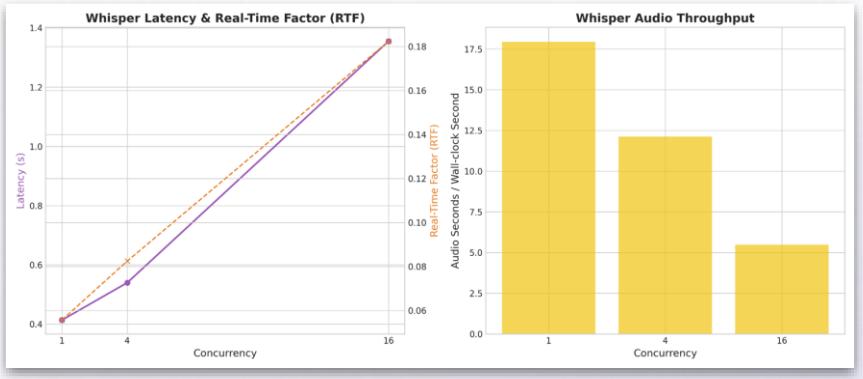
HPC Environment
CHESS cluster management; EaaS for dynamic compute allocation.

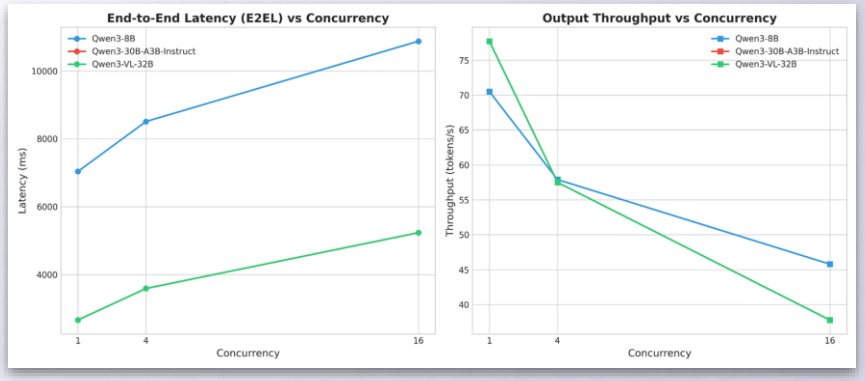
API Gateway Architecture
Proxy layer for AI model traffic inspection and control.

Support & Operations
Ongoing maintenance, incident response, system updates.



CHESS





Benchmarking Multi-Modal LLMs on MetaX C500

What This Enables

- AI intra design and construction
- Local, same-time-zone communication
- Long-term operational sustainability

Infrastructure Support

Capacity

- Daily index size
- Multiple document repository sources
- Multiple file types
- Large file quantity

Security

- Doc access control
- Private data
- Sensitive info preprocess/filter

Availability

- Embedding model selection
- VectorDB selection

Flexibility

- Throughput
- Response time
- Scale out /up
- High availability

Compatibility

- On-Prem deployment
- Domestic GPU server
- Sizing/ Cost
- Public LLM service

Environment Cloud On-premises HPC Slurm Kubernetes VMs Bare metal

and more…

Agent Access Control & Governance

Bounded Scope by Design

- ➔• Each agent is scoped to a specific task with a clear boundary
- ➔• Cannot access tools/information outside its boundary

Human Approval Gates for Agent

- ➔• High-risk actions require explicit sign-off
- ➔• Configurable per use case / department

RBAC per Agent

- ➔• Role-based access to each agent
- ➔• User groups see only authorized tools

Audit Trail of Agent Decisions

- ➔• Every invocation, prompt, and output logged
- ➔• Traceable to user, role, and timestamp

Private Cloud / On-Prem Boundary

- ➔• Data and agents stay inside your network
- ➔• Full sovereignty over model traffic

TRUSTED BY DESIGN

Cited | RBAC | Audit trail | Private cloud | Human-approved

In Summary

1. One pipeline, ten specialist agents, deployable across civil-service missions.
2. Governed and Trustworthy by design.
3. Sovereign by deployment.
4. Growing smarter as your institutional asset.

THANK YOU

ClusterTech Limited

Units 211 - 213, Lakeside 1, No. 8 Science Park West Avenue, Hong Kong Science Park, N.T., Hong Kong
Tel: +852 2655 6100
Email: enquiry@clustertech.com
Website: www.clustertech.com