An Agentic Multimodal Architecture for Personal Knowledge Compilation, Epistemic Evolution, and Expert-Level Reasoning

Alberto Espinosa

January 16, 2026

Abstract

Personal knowledge collections increasingly span heterogeneous sources and modalities, yet remain epistemically static, fragmented, and prone to hallucination when used in contemporary AI systems. This work proposes a rigorous agentic architecture for personal knowledge compilation and expert-level reasoning, extended with a deep external research agent to support epistemic evolution over time. The system explicitly distinguishes between shallow operational agents and deep epistemic agents, enforcing principled boundaries on reasoning depth, uncertainty management, and external inquiry. The result is a local-first cognitive infrastructure capable of knowledge mastery while remaining robust to temporal obsolescence and epistemic overconfidence.

1 Motivation and Problem Statement

The central challenge addressed by this research is not information retrieval, but epistemic integrity. Personal knowledge systems derived from saved references, papers, courses, and multimedia content are inherently incomplete, temporally bounded, and vulnerable to hallucination when treated as closed worlds.

Existing retrieval-augmented systems assume static corpora and lack mechanisms for:

- Epistemic uncertainty detection
- Controlled external inquiry
- Knowledge versioning and evolution

This project reframes personal AI systems as *epistemic infrastructures* rather than information tools.

2 Research Objectives

1. Design a universal ingestion and compilation architecture for heterogeneous personal knowledge artifacts.

2. Construct a multi-resolution internal knowledge representation supporting expert reasoning.
3. Explicitly model epistemic boundaries to minimize hallucination.
4. Introduce a principled deep external research mechanism for controlled knowledge evolution.
5. Demonstrate a local-first, multimodal implementation using agentic orchestration.

3 Conceptual Distinction: Shallow vs Deep Agents

A core contribution of this work is the explicit distinction between *shallow* and *deep* agents.

3.1 Shallow Agents

Shallow agents operate within a fixed epistemic boundary. They:

- Perform deterministic or bounded transformations
- Do not introduce new knowledge
- Do not reason under open uncertainty

Shallow agents are responsible for ingestion, expansion, scraping, structuring, and indexing.

3.2 Deep Agents

Deep agents operate across epistemic boundaries. They:

- Formulate research-grade questions
- Reason under uncertainty
- Compare competing knowledge claims
- Explicitly represent confidence and evidence

Design Principle: Only agents that cross epistemic boundaries are permitted to be deep.

4 High-Level Architecture

The system consists of layered shallow agent pipelines, complemented by a constrained deep external research agent. Agent orchestration is implemented via a graph-based execution framework.

5 Layer 0: Universal Ingestion (Shallow)

5.1 Epistemic Question

"What reference has the user saved, independent of format or source?"

5.2 Canonical Source Descriptor

```
{
  "source_id": "uuid",
  "origin": "linkedin | browser | manual | other",
  "raw_reference": "original input",
  "raw_text": "optional extracted text",
  "links": ["url_1", "url_2"],
  "content_hints": {
    "contains_pdf": false,
    "contains_video": true,
    "contains_external_links": true
  },
  "timestamp": "ISO-8601"
}
```

6 Layer 1: Exploration and Expansion (Shallow)

6.1 Epistemic Question

"What actual knowledge artifacts exist behind this reference?"

This agent identifies underlying papers, videos, courses, or datasets and dispatches specialized acquisition pipelines.

7 Layer 2: Knowledge Acquisition and Scraping (Shallow)

This layer obtains primary materials and performs multimodal extraction, segmentation, and structuring. It marks the transition from reference management to knowledge possession.

8 Layer 3: Multi-Resolution Knowledge Representation (Shallow)

Knowledge is stored explicitly at four levels:

- Level 0: Raw content chunks
- Level 1: Section-level structured representations
- Level 2: Document-level synthesis
- Level 3: Cross-document thematic synthesis

Lower levels support factual grounding; higher levels support reasoning and pedagogy.

9 Layer 4: Classification and Ontology Construction (Shallow)

This layer constructs topic hierarchies, dependency graphs, and redundancy mappings, enabling selective reasoning without global semantic search.

10 Layer 5: Expert Reasoning and Teaching (Shallow-Composite)

This agent composes explanations using precompiled representations but is explicitly constrained to the internal knowledge base. It cannot introduce external facts.

11 Layer 6: Deep External Research Agent (DERA)

11.1 Purpose

The Deep External Research Agent is introduced to prevent epistemic stagnation and hallucination under temporal or conceptual uncertainty.

11.2 Trigger Conditions

DERA is activated only under explicit conditions:

- Temporal obsolescence detection
- Internal explanatory insufficiency
- Explicit user authorization
- High epistemic uncertainty signals

11.3 Epistemic Question

> **"Does authoritative external knowledge exist that materially alters or invalidates the current internal understanding?"**

11.4 Operational Phases

1. Research-grade question formulation
2. External evidence gathering from authoritative sources
3. Triangulation and convergence analysis
4. Comparative epistemic reporting

11.5 Output Structure

```
{
  "internal_view": "description",
  "external_findings": [
    {
      "concept": "new approach",
      "evidence_strength": "low | medium | high",
      "source_count": 7,
      "status": "emerging | established"
    }
  ],
  "impact_assessment": "none | partial | superseding",
  "confidence": 0.0-1.0
}
```

DERA does not overwrite knowledge; it annotates, versions, and flags content for validation.

12 Technology Stack

- Agent orchestration: LangGraph
- Local execution: Ollama
- Text models: LLaMA 3.x, Qwen, Gemma
- Multimodal models: LLaMA 3.2 Vision
- Vector storage: FAISS or Qdrant
- Structured storage: relational and graph databases

13 Project Structure

```
project_root/
  ingestion/
  exploration/
  acquisition/
  representation/
  ontology/
  reasoning/
  deep_research/
  orchestration/
  storage/
```

14 Phased Research Plan

1. Shallow pipeline implementation
2. Multi-resolution representation
3. Teaching agent integration
4. Deep external research integration
5. Epistemic evaluation and stress testing

15 Conclusion

By explicitly distinguishing shallow operational agents from deep epistemic agents, this architecture avoids the common failure modes of hallucination, epistemic closure, and temporal stagnation. The introduction of a constrained Deep External Research Agent enables principled knowledge evolution while preserving epistemic integrity. The system thus constitutes a personal cognitive infrastructure capable of expertise, teaching, and responsible epistemic growth.