

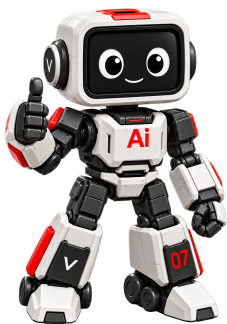


AI IN 7 DAYS

DAY 6 HANDBOOK - SELF-STUDY GUIDE

Building Modern AI Applications

6 VIDEOS	60 MINUTES	3 PRACTICAL OUTPUTS	1 APP ARCHITECTURE
-------------	---------------	------------------------	-----------------------



© 2026 AI PANTHIYA

අද ඔබ ඉගෙනගන්නේ

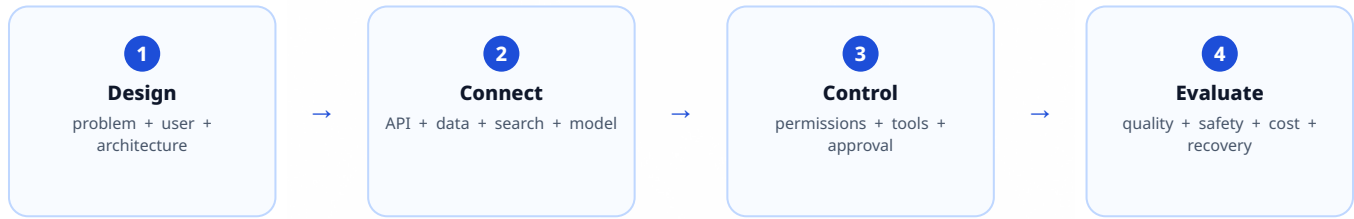
- AI Model එකකින් complete AI Application එකකට යන architecture එක
- Frontend, Backend, Databases, Authentication සහ Model Services
- APIs, JSON, Schema Validation, Latency, Cost සහ Reliability
- Search, Embeddings, Vector Databases සහ Hybrid Search
- RAG pipeline එක - documents සිට grounded answers දක්වා
- Chatbots, Workflows, Agents, Tools, Memory සහ Permissions
- Human Approval, Evaluation, Monitoring සහ Failure Recovery



Day 6 Main Task: ඔබේ AI solution එක සඳහා **AI Application Workflow Canvas + Permission Matrix + Evaluation Plan** එකක් සකස් කිරීම.

HOW TO USE THIS HANDBOOK

Day 6 එකේ focus එක model එකක් තෝරාගැනීමෙන් ඉදිරියට යාමයි. Real user කෙනෙක් භාවිත කරන AI application එකකට model එකට අමතරව interface, backend, data, search, permissions, tools, guardrails, verification, monitoring සහ failure recovery අවශ්‍ය වෙනවා.



Day 6 Learning Outcomes

- AI Model සහ complete AI Application අතර difference පැහැදිලි කිරීම
- No-code, Low-code සහ Full-code build routes compare කිරීම
- API request/response cycle සහ performance/cost concepts තේරුම්ගැනීම
- Keyword, Semantic, Vector, Hybrid සහ Graph search වෙන්කර හඳුනාගැනීම
- Complete RAG ingestion + query pipeline trace කිරීම
- Chatbot, Copilot, Workflow, Agent සහ Autonomous System අතර autonomy difference තේරුම්ගැනීම
- State, History, Short-term Memory, Long-term Memory සහ External Knowledge වෙනස් කිරීම
- Least Privilege, Human Approval, Audit Logs සහ Recovery design කිරීම

Today's 6 Video Map

1. From Model to Complete AI Application
2. APIs, Inputs, Outputs, Performance and Cost
3. Search, Embeddings and Vector Databases
4. RAG from Documents to Grounded Answers
5. Chatbots, Workflows, Agents, Tools and Memory
6. Permissions, Human Approval and Application Evaluation

Core rule: Model එක application එකේ එක් එක් component එකක් විතරයි. Useful AI product එකක් සඳහා **right data + right architecture + right permissions + verification + monitoring** එකට වැඩ කරන්න ඕන.

From Model to Complete AI Application

AI model එකක් userට භාවිත කළ හැකි product එකක් වෙන්නේ system components එකට වැඩි කළාමයි.

Complete Application Flow

User / Interface



Auth + Backend



Data / Retrieval



Model Service



Guardrails



Output +
Monitoring

No-code

Visual interfaces සහ ready-made blocks භාවිත කර ඉක්මනින් prototype කරන්න. Beginner friendly, නමුත් advanced control සීමා විය හැක.

Low-code

Ready-made components + custom code. Speed සහ flexibility අතර middle path එකක්.

Full-code

Frontend, backend, APIs, database සහ security custom code වලින් build කිරීම. Maximum control, වැඩි skill/time අවශ්‍යයි.

© 2026 AI PANTHIYA

Main Components

- **Frontend:** userට පෙනෙන interaction layer
- **Backend:** workflow සහ business logic coordinate කරන layer
- **Authentication:** “ඔයා කවුද?”
- **Authorization:** “ඔයාට මොනවා කරන්න පුළුවන්ද?”
- **Guardrails:** input, output, permissions සහ safety controls

Right Storage for Right Data

- Structured records → Relational Database
- Flexible JSON-like data → Document / NoSQL Database
- PDFs, images, audio → Object Storage
- Embeddings → Vector Store
- Entities + relationships → Graph Database
- Frequently reused temporary result → Cache

Important: No-code tool එකක් use කළා කියලා security, privacy, permissions සහ evaluation responsibility නැති වෙන්නේ නැහැ.

APIs, Inputs, Outputs, Performance and Cost

Software systems එකිනෙකට communicate කරන basic contract එක API එකයි.

API Request / Response Cycle

Your App



Endpoint



Request + JSON



AI Service



Response



Schema Check

Input Validation

Required fields, type, length, allowed values සහ dangerous content check කරන්න.

Schema Validation

Output structure expected fields/data types වලට match වෙනවද බලන්න.

API Key

Secret credential එකක්. Frontend, screenshots හෝ public repository වල expose කරන්න එපා.

Latency / Throughput

Latency: one request wait time. **Throughput:** system එක time period එකක handle කරන total work.

Concurrency / Rate Limit

Concurrency: එකම වෙලාවේ active requests. **Rate Limit:** allowed usage limit.

Reliability

Timeout, Retry, Exponential Backoff, Idempotency, fallback සහ error paths කලින් design කරන්න.

Caching vs Batching

Caching: repeated result temporary store කර reuse කිරීම.

Batching: items කිහිපයක් group එකක් ලෙස process කිරීම.

Total Cost

Model tokens විතරක් නොව **embeddings + vector DB + storage + tools + search + cloud + monitoring** costsත් track කරන්න.

Remember: Generated JSON ≠ valid JSON. Valid JSON ≠ correct data. Structure එකත් meaning එකත් දෙකම validate කරන්න.

Search, Embeddings and Vector Databases

Search එක එකම method එකක් තෙවෙයි. Query type එක අනුව right strategy එක තෝරන්න.

Keyword Search

Exact words, codes, names හොයනවා. Policy number වගේ exact identifier සඳහා useful.

Semantic / Vector Search

Exact words වෙනස් වුණත් meaning similar content embeddings මත හොයනවා.

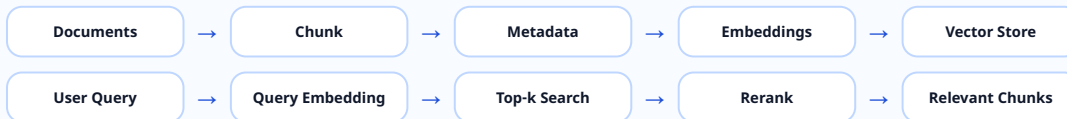
Hybrid Search

Keyword + semantic signals combine කරනවා. Exact term + meaning දෙකම වැදගත් query සඳහා.

Graph Search

Entities සහ relationships follow කරනවා. "Who is connected to what?" වගේ questions සඳහා.

Semantic Retrieval Pipeline



Chunking Trade-off

Too small → context කැඩෙන්න පුළුවන්. Too large → noise සහ token usage වැඩි. Overlap boundary context preserve කරන්න useful, නමුත් duplicates වැඩි කළ හැක.

Metadata + Versioning

Source, page, date, language, version, department, permissions වගේ metadata search filtering සහ citations සඳහා වැදගත්. Embedding model version එක වෙනස් වුණොත් re-embedding අවශ්‍ය විය හැක.

Critical rule: Similarity score එක relevance signal එකක් විතරයි. **Similarity ≠ Truth**. Retrieved source එක latest ද, reliable ද, permission තියෙනවද කියලා වෙනම check කරන්න.

RAG from Documents to Grounded Answers

RAG = external evidence retrieve කරලා model එකට context ලෙස දී source-grounded answer එකක් හදන architecture pattern එකක්.



Remember: Citation තිබීම claim supported බව automatically prove කරන්නේ නැහැ. **Claim** ↔ **exact source passage** connection එක verify කරන්න.

Chatbots, Workflows, Agents, Tools and Memory

Main difference එක autonomy level, tool use සහ "next step" තෝරන්නේ කවුද කියන එකයි.



Safety rule: Higher tool risk → stronger permission + approval + logging + recovery. Read-only calculator එක සහ money transfer එක එකම permission level එකකින් handle කරන්න එහ.

Permissions, Human Approval and Application Evaluation

Production-ready AI system එකකට success path එකත් failure path එකත් දෙකම design කරන්න ඕන.

System-Level Thinking

Problem + User



Inputs + Access



Data + Retrieval



Model + Tools



Controls



Evaluate + Monitor

LLM Evaluation

- Correctness
- Relevance
- Completeness
- Groundedness
- Safety
- Style / instruction following
- Latency + Cost

RAG Evaluation

- Retrieval Precision / Recall
- Context Relevance
- Citation Correctness
- Faithfulness
- Abstention when evidence is insufficient

Agent Evaluation

- Task Success
- Tool Selection
- Arguments Correctness
- Step Efficiency
- Recovery
- Safety + Approval
- Cost

Least Privilege + Human Approval

Task එකට අවශ්‍ය minimum permissions විතරක් දෙන්න. High-impact actions සඳහා execution කලින් authorised human approval gate එකක් දාන්න.

Example: Draft email → automatic විය හැක. Send email → approval useful. Money transfer / delete critical data → strong approval essential.

Failure Recovery

No source: guess නෙකර “evidence insufficient” කියන්න.

Model outage: fallback model හෝ safe error.

Tool failure: retry / stop / human route.

Partial action: rollback හෝ reconcile.

Production rule: Build → Deploy → Monitor → Detect → Recover → Improve. One good demo output එකක් production readiness prove කරන්නේ නැහැ.

DAY 6 PRACTICAL WORKSHEET

මෙම AI application idea එකක් කේරලා පහත canvas එක පුරවන්න. Short answers ලියන්න; architecture එක later refine කරන්න.

1. Problem

Application එක solve කරන actual problem එක?

2. Users

Application එක use කරන්නේ කවුද?

3. Inputs

Question, file, image, form, event?

4. Data

Documents, databases, APIs, metadata?

5. Search / Retrieval

Keyword, Vector, Hybrid, Graph?

6. Model

Hosted, open-weight, local, router?

7. Tools

Search, calculator, email, calendar, database?

8. Permissions

Who can read/write/execute what?

9. Memory

What is retained? For how long?

10. Verification

How are claims / actions checked?

11. Human Approval

Which actions require approval?

12. Evaluation

LLM, RAG, Agent metrics?

13. Output

Answer + sources + uncertainty + status?

14. Monitoring

Quality, latency, cost, errors, security?

15. Failure Paths

What happens if model/search/tool fails?

16. Cost

Model, embeddings, storage, tools, monitoring?

17. Main Risks

Privacy, hallucination, stale data, permission leakage?

18. Final Architecture

Draw: Input → Access → Data → Model → Tools → Verify → Output → Monitor

Permission Matrix - Quick Draft

Action	Risk	Approval?
Read public info	Low	No
Save preference	Medium	Consent
Send email	High	Usually Yes
Delete / transfer money	Very High	Strong Approval

Evaluation Plan - Quick Draft

- 10 normal questions
- 5 ambiguous / edge cases
- 5 unsafe or unauthorised requests
- 5 evidence-insufficient RAG cases
- Measure correctness, groundedness, latency, cost
- Record tool failures and approval behaviour

DAY 6 QUICK SELF-CHECK

Quiz Questions

1. AI Model සහ AI Application අතර difference එක?
2. Frontend සහ Backend roles මොනවාද?
3. Authentication vs Authorization?
4. API request/response cycle explain කරන්න.
5. Latency, Throughput, Concurrency වෙනස් වන්නේ කොහොමද?
6. Keyword vs Semantic vs Hybrid Search?
7. Embedding කියන්නේ මොකක්ද?
8. Chunking සහ Metadata RAG සඳහා වැදගත් ඇයි?
9. RAG offline සහ online stages මොනවාද?
10. Citation තිබීම claim supported බව prove කරනවාද?
11. Chatbot, Workflow සහ Agent අතර difference?
12. State, History, Memory සහ External Knowledge අතර වෙනස?
13. Least Privilege කියන්නේ මොකක්ද?
14. High-risk tool action එකකට controls මොනවාද?
15. LLM, RAG සහ Agent evaluation වෙන වෙනම අවශ්‍ය ඇයි?

Answer Hints

- Model = learned component; Application = complete working system.
- Frontend = user interaction; Backend = orchestration/business logic.
- Identity vs allowed actions.
- App → Endpoint/Request → Service → Response → Validation.
- Wait time / total work rate / simultaneous active requests.
- Exact terms / meaning / both signals.
- Numerical representation used for similarity/retrieval.
- Searchable context pieces + source/version/permission info.
- Prepare documents / answer user query in real time.
- නැහැ. Exact source passage verify කරන්න.
- Answer / predefined process / goal-driven action.
- Current status / past messages / retained info / retrieved source info.
- Minimum permissions only.
- Validation + permission + approval + logging + recovery.
- Different failure modes occur at model, retrieval and action layers.



Sampath Bandaranayake
Instructor - AI Panthiya

Need guided learning?

Online Zoom Classes - Contact Us on WhatsApp

WhatsApp: +81 70 2313 1981

MSc - AI and Digital Innovation (Final Year) - Tokyo International University, Japan

MBA - Cardiff Metropolitan University, UK

BSc - Applied Sciences, Sri Lanka

LinkedIn: <https://www.linkedin.com/in/sampathbandara/>

Website: <https://aipanthiya.com/>

Next step: **Day 7 - Deployment, MLOps, Security, Governance and Final AI Solution Design**