Development Individual Project

Intelligent Forensics Agent – Practical Development (Unit 6 Design)

MSc Artificial Intelligence – IA_PCOM7E (July 2025 A

Murthy Kanuri - Student ID:12696139
Date: 10 October 2025

PROBLEM & OBJECTIVES

Problem :

- Manual triage is slow and inconsistent.
- File extensions can be misleading. Detect type from the file's bytes *(Dubettier et al., 2023).*
- Cross-platform quirks (paths, mounts, protected directories) may cause misses or errors.
- Evidence integrity and privacy require minimal, auditable capture *(ICO, 2024).*

Objectives :

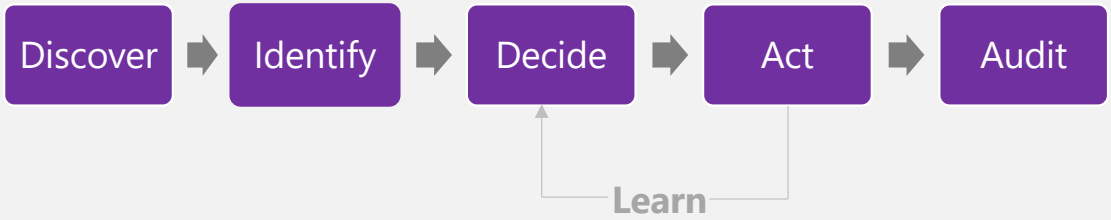
- Automate **safe discovery** (platform-aware exclusions; **read-only**).
- Identify **by content first** (**python-magic** → **filetype** → extension fallback) *(Hupp, 2022; Aparicio, 2022).*
- Capture **essential metadata only** (SHA-256, size, **timestamps**, MIME) and store in **SQLite** (*SQLite Consortium, 2025).*
- Make runs **reproducible** (CLI-driven with a consistent summary) *(ICO, 2024).*
- Lay the groundwork for **automation** (small supervised learner: archive recommendations).

References : Dubettier et al., 2023; Hupp, 2022; Aparicio, 2022; ICO, 2024; SQLite Consortium, 2025.

ARCHITECTURE

Architecture: Hybrid Reactive Pipeline

- Agent 1(File Locator): Discovery, Identification, Metadata Extraction, Store in SQLite, Audit via CLI summary.



- Agent 2(File Archiver): Predictive archiving based on learned human behaviour.

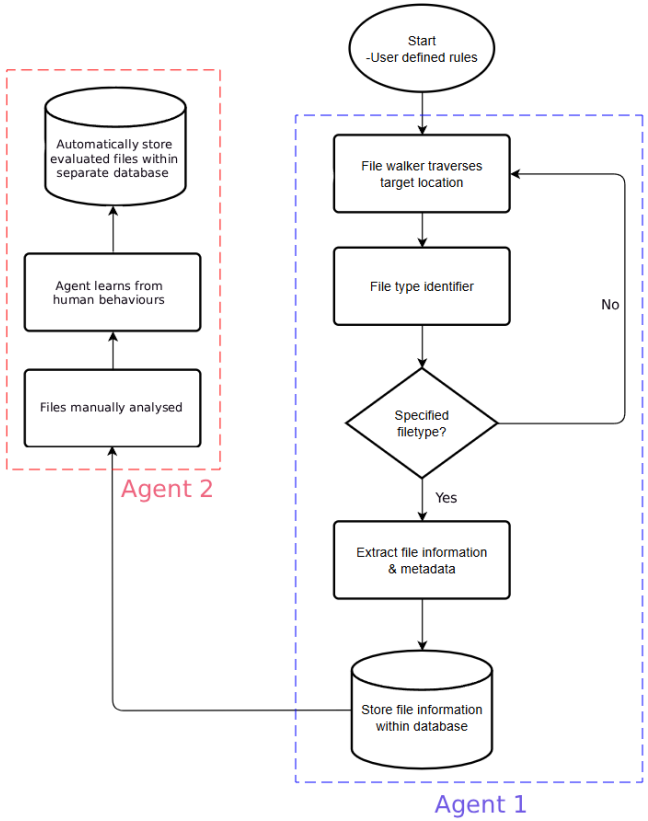
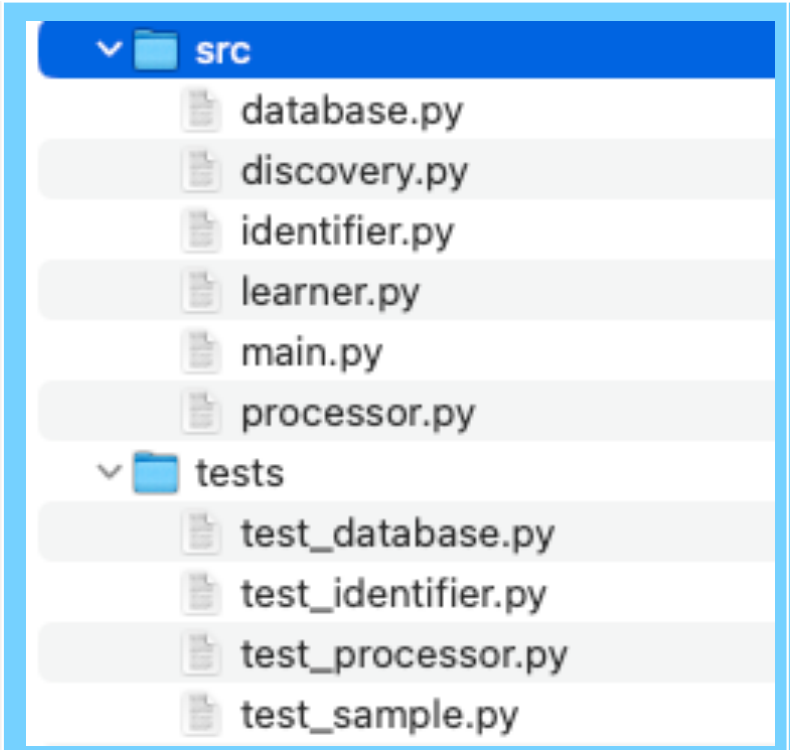


Diagram (right) adapted from Unit 6 Group E report: High-Level Logic Flow (Kanuri et al., 2025).
Reference : Shoham, 1993

COMPONENTS & CODE STRUCTURE



src/

```
discovery.py  – safe traversal (psutil; platform-aware exclusions)
identifier.py – content based MIME (python-magic → filetype → extension )
processor.py  – SHA-256, size, timestamps (data minimisation)
database.py   – SQLite upsert + indexes (idempotent reruns)
learner.py    – features + small logistic regression (joblib)
main.py       – CLI orchestrator; prints Run Summary
```

tests/

```
test_discovery.py  test_identifier.py
test_processor.py  test_database.py
```

Pipeline map: Discover ↔ discovery.py | Identify ↔ identifier.py | Decide ↔ learner.py + rules | Act ↔ processor.py + database.py | Audit ↔ main.py (CLI Run Summary)

Libraries referenced: psutil; python-magic/libmagic; filetype; Python 3.11 stdlib (sqlite3, hashlib).

KEY LIBRARIES & JUSTIFICATION

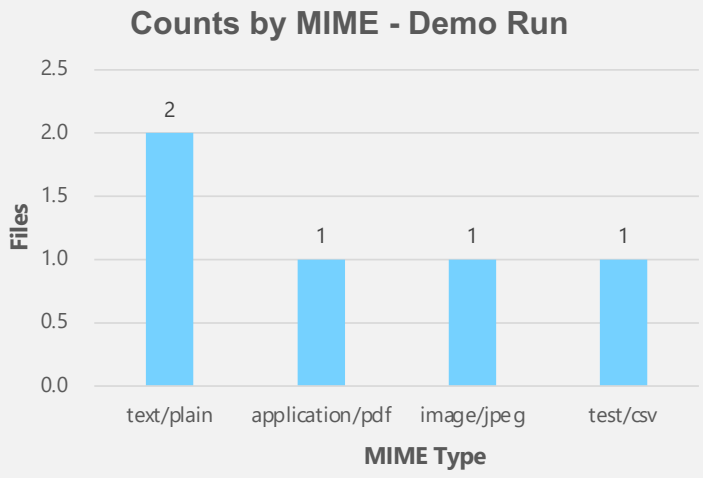
- **psutil** — **Environment-aware discovery** (mounted volumes, platform quirks); enables **safe exclusions** during traversal *(Rodola, 2024)*.
- **python-magic** — **Content-based MIME** detection (more accurate than extensions); primary identifier *(Hupp, 2022)*.
- **filetype** — Lightweight fallback when libmagic/python-magic isn't available *(Aparicio, 2022)*.
- **sqlite3 / SQLite** — Portable, single-file DB; **upsert + indexes** for **idempotent, auditable runs** (no server*) (SQLite Consortium, 2025).*
- **hashlib (SHA-256)** — Evidence integrity (stable content hash) *(Python Software Foundation, 2024)*.
- **scikit-learn + joblib** — Tiny **logistic regression** for "archive?" recommendations; persisted for reproducibility *(Pedregosa et al., 2011; Joblib Developers, 2024)*.
- **mimetypes (stdlib)** — Final safety net if both detectors fail *(Python Software Foundation, 2024)*.

Sources: Rodola (2024); Hupp (2022); Aparicio (2022); Python Software Foundation (2024); Pedregosa et al. (2011); Joblib Developers (2024); SQLite Consortium (2025). Note : macOS: brew install libmagic; Windows: use python-magic-bin

DEMO : RUN SUMMARY

```
(venv) murthykanuri@MKanuri-MacBook-Pro ia_agent % python -m src.main --target sample_data --db agent.db | tee run_output.txt

=== Run Summary ===
Target: /Users/murthykanuri/Documents/ia_agent/sample_data
DB: /Users/murthykanuri/Documents/ia_agent/agent.db
Files processed: 5
Counts by MIME:
  text/plain: 2
  application/pdf: 1
  image/jpeg: 1
  text/csv: 1
Sample record:
{
  "path": "/Users/murthykanuri/Documents/ia_agent/sample_data/note1.txt",
  "size": 32,
  "mtime": 1759195654.0,
  "ctime": 1759200738.640116,
  "mime": "text/plain",
  "sha256": "8ffbe438c6ab1604bbcb942b55b4241df9d33baa19f0215aa9b72fe7ec7ccf93",
  "detector": "python-magic"
}
Elapsed: 0.03s
```



Click here to watch the demo video

TEST EVIDENCE

```
(venv) murthykanuri@MKanuri-MacBook-Pro ia_agent % coverage run -m unittest discover -s tests
coverage report -m | tee coverage_output.txt

...
----------------------------------------------------------------------
Ran 3 tests in 1.477s

OK
```

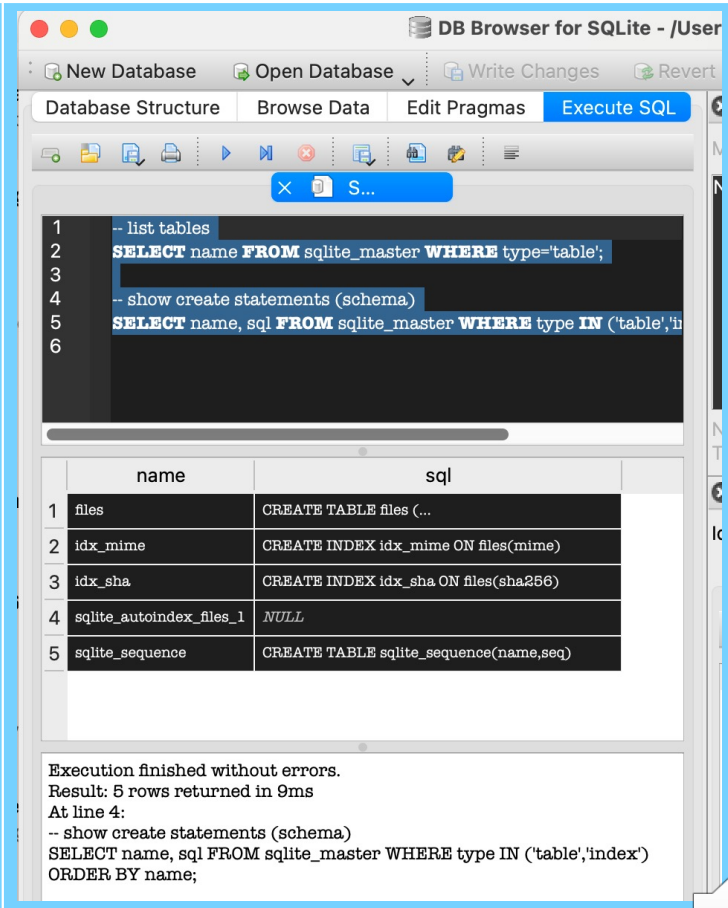
```
Name                  Stmts   Miss  Cover   Missing
-----------------------------------------------
src/database.py          21     14    33%   22-25, 28-45, 48-50
src/identifier.py        33     21    36%   12-13, 17-18, 21-37
src/processor.py         12      7    42%   8-12, 15-16
-----------------------------------------------
TOTAL                    66     42    36%
```

✓ *3 tests passed in 1.48 s*

■ *Coverage baseline: 36% (database 33%, identifier 36%, processor 42%)*

→ *Next: add tests for python-magic fallback and upsert-conflict handling*

Click here to watch the demo video



Sources: Python unittest docs (PSF, 2025); coverage.py docs (Batchelder, 2025).

LEARNING COMPONENT (Toy Classifier)

- Goal: flag likely non-text files for archiving.
- Features (X): log(size), is_text, is_image, is_pdf.
- Labels (y): 1 = image/pdf (archive), 0 = text (keep).
- Data: derived from the same demo scan; no file contents read.
- Model: StandardScaler → LogisticRegression (class_weight="balanced").
- Outputs: demo_model.joblib + demo_model.sha256, learner_predictions.csv, learner_pred_counts.csv.
- Scope: advisory only (guidance, not an autonomous decision system).

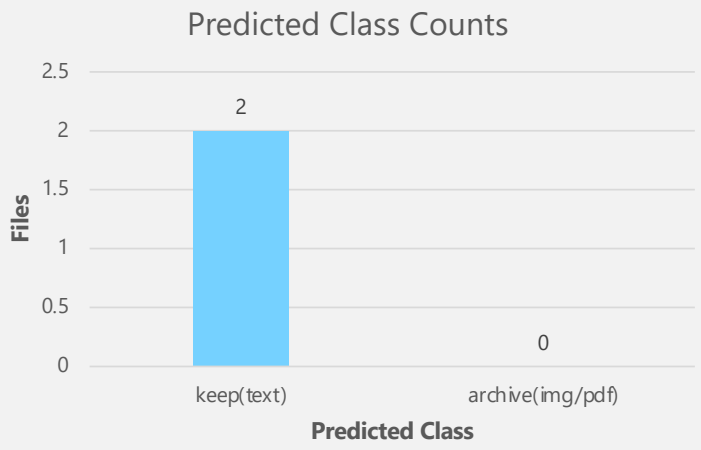
```
(.venv) murthykanuri@MKanuri-MacBook-Pro ia_agent % ls -lh demo_model.joblib demo_model.sha256 learner_predictions.csv learner_pred_counts.csv

# show first few predictions (path, mime, y_true, y_pred)
head -n 10 learner_predictions.csv

# show class counts (what your bar chart uses)
cat learner_pred_counts.csv

# (optional) verify the model hash with macOS shasum
shasum -a 256 demo_model.joblib | cut -d' ' -f1

-rw-r--r--@ 1 murthykanuri  staff   1.4K 11 Oct 19:37 demo_model.joblib
-rw-r--r--@ 1 murthykanuri  staff    65B 11 Oct 19:37 demo_model.sha256
-rw-r--r--@ 1 murthykanuri  staff    51B 11 Oct 19:37 learner_pred_counts.csv
-rw-r--r--@ 1 murthykanuri  staff   179B 11 Oct 17:37 learner_predictions.csv
zsh: number expected
path,mime,y_true,y_pred
/Users/murthykanuri/Documents/ia_agent/sample_data/photo.jpg,image/jpeg,1,0
/Users/murthykanuri/Documents/ia_agent/sample_data/script.py,text/plain,0,0
zsh: unknown file attribute: h
predicted,count
keep(text),2
archive(img/pdf),0
zsh: unknown sort specifier
e0b94679ea0480d18839c0c205a2509e3c15e43d2f05d02a87a2ac2da04a89af
(.venv) murthykanuri@MKanuri-MacBook-Pro ia_agent %
```



Note: Trained on the demo scan; counts reflect small sample size.

Click here to watch the demo video

References : Pedregosa et al. (2011); scikit-learn Developers (2025); Joblib Developers (2025).

RISKS, ETHICS & MITIGATIONS

Area	Risk / Concern	Mitigations
Operational	MIME edge cases / spoofed extensions → misclassification, wrong stats	Layered checks: python-magic → filetype → extension; flag unknown.
Operational	Platform variance (OS/paths/permissions) → crashes, inconsistent results	Pin deps; handle path/perm errors; pre-run smoke tests.
Operational	Large folders / long scans → timeouts, partial runs	Progress + Run Summary; safe error handling; optional file/time limits.
Operational	DB locks / corruption → data loss, failed writes	Single-writer process; integrity checks; SQLite WAL.
Operational → Ethical	Model misuse (toy labels) → false positives/negatives	Clearly label as demo; keep scope minimal; human review only.
Ethics / Compliance	Scope creep / lack of consent → privacy breach, policy violation	Explicit scope & consent; data minimisation; platform-aware excludes; least privilege; local-only; no content stored.
Ethics / Compliance	Path/filename sensitivity → indirect disclosure when sharing	Redact paths in logs/screenshots; share aggregates only.
Ethics / Audit	Audit gaps → poor traceability	Dry-run mode; signed audit log (hash + timestamp of Run Summary); capture version & config; retention window + purge command.

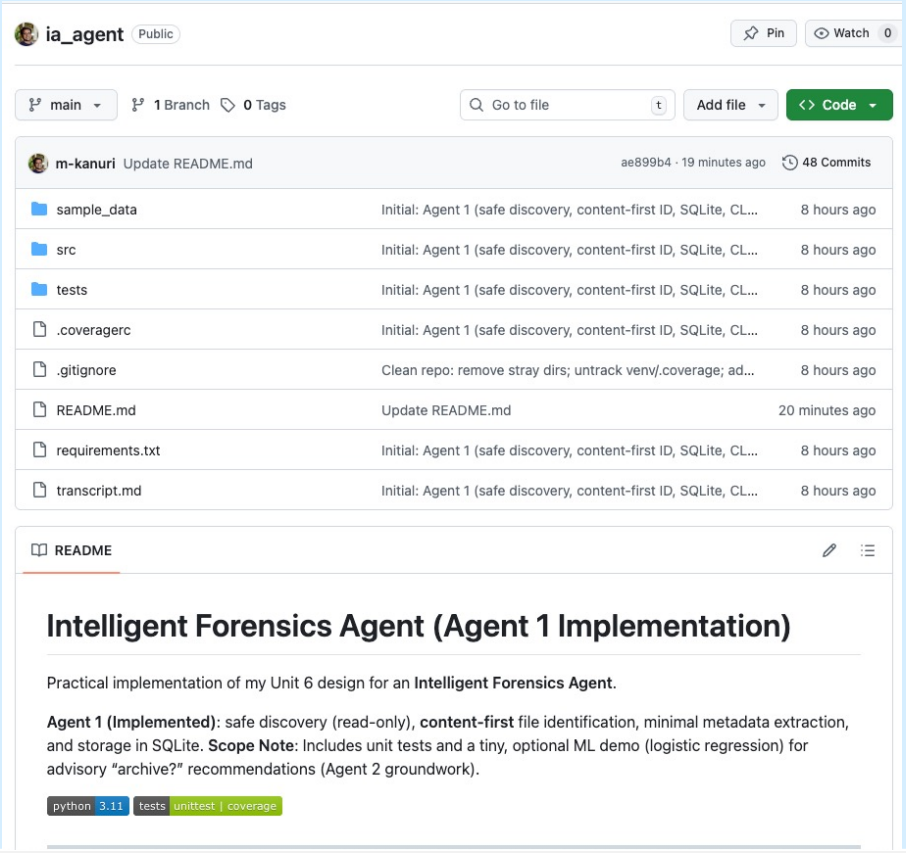
Source: ICO (2024) data minimisation & storage limitation; NIST SP 800-53 Rev.5 (least privilege & audit controls); OWASP Logging Cheat Sheet (2023) (redaction); SQLite docs (2025) (WAL); Kessler (2024) & MITRE ATT&CK T1036 (extension spoofing); Hupp (2022) & Aparicio (2022) (content-based type detection)

CONCLUSION & CODE

Conclusion

- **Implemented (Agent 1):** Safe discovery (read-only, excludes) → content-first ID (magic→filetype→ext) → metadata only (SHA-256, size, times, MIME) → SQLite upsert → CLI summary.
- **Evidence:** Demo run + MIME chart; 3/3 tests; 36% coverage; artefacts: agent.db, run_output.txt, coverage_output.txt, demo_model.joblib.
- **Design fit:** Meets Unit 6 — automation, robustness, data minimisation, reproducibility.
- **Limits:** Toy recommender; moderate coverage; MIME/archive edge cases; large trees slow.
- **Next (Agent 2 + polish):** Archive recommendations (active learning); rules+model; more tests (fallback/error); retention & purge; packaging; optional UI; CI.

Code in GitHub



Code Reference : github.com/m-kanuri/ia_agent

REFERENCES

- **Aparicio, T. (2022)** *filetype: infer file type and MIME*. Available at: https://github.com/h2non/filetype (Accessed: 11 October 2025).
- **Batchelder, N. (2025)** *coverage.py — Code coverage for Python*. Available at: https://coverage.readthedocs.io/ (Accessed: 11 October 2025).
- **Dubettier, A., et al. (2023)** 'File type identification tools for digital investigations', Forensic Science International: Digital Investigation, 46, p. 301574. doi:10.1016/j.fsidi.2023.301574.
- **Hupp, A. (2022)** *python-magic documentation*. Available at: https://github.com/ahupp/python-magic (Accessed: 11 October 2025).
- **Information Commissioner's Office (ICO) (2024)** 'UK GDPR: Principle (c) Data minimisation'. Available at: https://ico.org.uk/ (Accessed: 11 October 2025).
- **Joblib Developers (2025)** *Joblib documentation*. Available at: https://joblib.readthedocs.io/ (Accessed: 11 October 2025).
- m-kanuri (2025) *ia_agent*. GitHub repository. Available at: https://github.com/m-kanuri/ia_agent (Accessed: 12 October 2025).
- **Kanuri, M., Espag, J., Kirwan, F. and Jittipattanakulchai, G. (2025)** *Intelligent Forensics Agent – Group E Design Report (Unit 6)*. University of Essex Online (unpublished coursework).
- **Kessler, G.C. (2024)** *File Signature Table*. Available at: https://www.garykessler.net/library/file_sigs.html (Accessed: 11 October 2025).
- **MITRE (2024)** *MITRE ATT&CK®: Masquerading (T1036)*. Available at: https://attack.mitre.org/techniques/T1036/ (Accessed: 11 October 2025).
- **National Institute of Standards and Technology (NIST) (2020)** *Security and Privacy Controls for Information Systems and Organizations (SP 800-53 Rev. 5)*. Gaithersburg, MD: NIST.
- **OWASP Foundation (2023)** *Logging Cheat Sheet*. Available at: https://cheatsheetseries.owasp.org/ (Accessed: 11 October 2025).
- **Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011)** 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- **PyPI (2024)** *python-magic-bin (Windows prebuilt)*. Available at: https://pypi.org/project/python-magic-bin/ (Accessed: 11 October 2025).
- **Python Software Foundation (PSF) (2025a)** *hashlib — Secure hashes and message digests*. Available at: https://docs.python.org/3/library/hashlib.html (Accessed: 11 October 2025).
- **Python Software Foundation (PSF) (2025b)** *mimetypes — Map filenames to MIME types*. Available at: https://docs.python.org/3/library/mimetypes.html (Accessed: 11 October 2025).
- **Python Software Foundation (PSF) (2025c)** *sqlite3 — DB-API 2.0 interface*. Available at: https://docs.python.org/3/library/sqlite3.html (Accessed: 11 October 2025).
- **Python Software Foundation (PSF) (2025d)** *unittest — Unit testing framework*. Available at: https://docs.python.org/3/library/unittest.html (Accessed: 11 October 2025).
- **Rodolà, G. (2025)** *psutil documentation*. Available at: https://psutil.readthedocs.io/ (Accessed: 11 October 2025).
- **scikit-learn Developers (2025)** *scikit-learn User Guide*. Available at: https://scikit-learn.org/stable/ (Accessed: 11 October 2025).
- **SQLite Consortium (2025)** *SQLite documentation*. Available at: https://sqlite.org/docs.html (Accessed: 11 October 2025).
- **Shoham, Y. (1993)** 'Agent-oriented programming', *Artificial Intelligence*, 60(1), pp. 51–92.

Thank You