RMIT UNIVERSITY · School of Science, Engineering & Technology

Industry partner: NEAX

Master of Artificial Intelligence · Capstone project 2025

ALM TrustGuard

A trust-control layer that scores every AI-generated engineering artifact before a human is asked to trust it.

Students: Nguyen Huu Nguyen · Pham Huy Minh · **Academic supervisor:** Dr Ginel Dorleon · **Industry supervisor:** Khanh Nguyen (NEAX Co., Ltd.)
School: Science, Engineering & Technology, RMIT University Vietnam

01 The problem

An **Application Lifecycle Management (ALM)** platform is the system of record for engineering a regulated product — every requirement, test case and traceability link lives there, and audits are run against it. That platform now *writes* those artifacts with AI. In **ISO 26262** and **FDA-regulated** environments an unverified AI artifact is a liability — yet no engineer can review AI output at the speed AI produces it.

Hallucination

Invented conditions, omitted edge cases and ambiguous safety language flow straight into the ALM baseline — product defects, ASIL violations, recalls.

False compliance

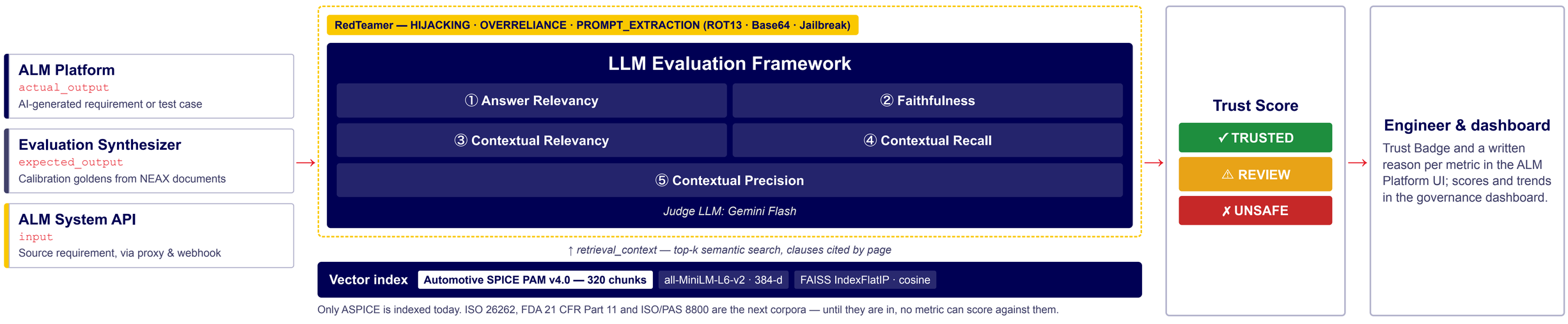
Artifacts pass superficial checks but are never measured against the ASPICE / FDA clauses they claim to satisfy — the gap surfaces months later, at review time.

Review bottleneck

Without a systematic trust signal, every artifact needs full manual review, which cancels out the entire productivity gain of using AI.

02 The solution — a non-invasive trust interceptor

TrustGuard sits between the ALM Platform and the engineer as a REST-API proxy and webhook listener. It **evaluates** AI output; it never generates or replaces it — and engineers change nothing about how they work.



Only ASPICE is indexed today. ISO 26262, FDA 21 CFR Part 11 and ISO/PAS 8800 are the next corpora — until they are in, no metric can score against them.

03 How the score is built

#	Metric	What it detects	Weight
1	**Faithfulness**	Claims not grounded in retrieved standards — hallucinated conditions	**0.30**
2	**Answer Relevancy**	Output drifting off the source requirement	**0.25**
3	**Contextual Relevancy**	Retrieved compliance context is not applicable	**0.20**
4	**Contextual Recall**	Coverage gaps in the retrieved clauses	**0.15**
5	**Contextual Precision**	Most relevant clauses not ranked highest	**0.10**

```
Trust Score = 100 × ( 0.30·Faithfulness
                    + 0.25·AnswerRelevancy
                    + 0.20·ContextualRelevancy
                    + 0.15·ContextualRecall
                    + 0.10·ContextualPrecision )
```

TRUSTED — 80 – 100 — Auto-approve and log

REVIEW — 50 – 79 — Flag + raise review ticket

UNSAFE — 0 – 49 — Block + request regeneration

04 Explainable, by design

No score without a reason

Every artifact carries all five metric scores, the retrieved ASPICE clauses they were judged against, and a written justification per metric — so a reviewer can see **why** an artifact was flagged, not just that it was. A number nobody can interrogate is a number nobody should act on.

Built on open, reproducible foundations

- **The LLM evaluation framework** — evaluation logic lives in metrics and GEval rubrics, so rubrics evolve by editing criteria, not by redeploying code.
- **FAISS index of ASPICE PAM v4.0** — 320 chunks of 180 words (35 overlap), embedded with all-MiniLM-L6-v2 and searched by cosine similarity; every retrieved clause keeps its page number.
- **The Synthesizer** generates goldens from real NEAX requirement documents to calibrate the score thresholds.
- **The RedTeamer** probes the pipeline for HIJACKING, OVERRELIANCE and PROMPT_EXTRACTION — target: zero successful bypasses.
- **~US$150** total cash cost: fully open-source stack on a free-tier judge LLM.

05 Results — 200-artifact PHEV propulsion evaluation

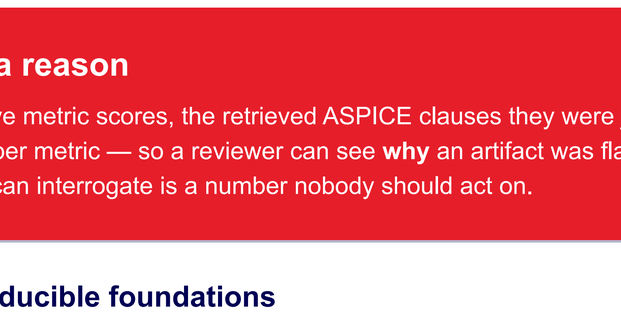
200 — AI-generated PHEV artifacts scored end-to-end

91.3 — Mean Trust Score (median 93.7)

1.00 — Mean Faithfulness — 197 of 200 fully grounded

176 — Artifacts auto-approved as TRUSTED — zero blocked

The finding: switching from ASPICE process-level boilerplate to each requirement's own **traceability chain** as retrieval context lifted all three RAG metrics from near-zero (relevancy 0.02 → 0.94, recall 0.10 → 0.86, precision 0.05 → 0.85). Generation quality remained perfect: faithfulness 1.00, answer relevancy 0.95. **176 of 200 artifacts scored TRUSTED** and were auto-approved; 24 went to a human for review, and none were blocked. The right retrieval context, not a better generator, was the lever that unlocked compliance-grade trust scores.



Distribution of Trust Scores across the 200-artifact PHEV propulsion-control dataset. 176 scored TRUSTED (≥ 80) and were auto-approved; 24 landed in REVIEW (50–79); none were blocked. Traceability-chain retrieval was the decisive improvement.

06 Why it matters & what's next

For engineers

A Trust Badge appears inside the ALM Platform with a collapsible per-metric report. Review effort concentrates on the artifacts that actually need it — target: **≥ 30 % review-efficiency gain** at **< 3 s** per artifact.

For quality leads

Scores aggregate into a governance dashboard: which datasets, requirement types and standards clauses the AI is weakest on — a measurement, not an opinion, of where AI can be trusted.

Next

Expand the standards corpus beyond ASPICE (ISO 26262, FDA 21 CFR Part 11, ISO/PAS 8800), re-chunk for clause-level retrieval, calibrate thresholds on ≥ 200 goldens, and complete the red-team pass.

Scan for the full project

ALM TrustGuard · RMIT Vietnam · School of Science, Engineering & Technology
Master of Artificial Intelligence Capstone, 2025 · in partnership with NEAX Co., Ltd.

Built with

LLM evaluation framework · Gemini Flash (judge LLM) · FAISS + MiniLM · Python · FastAPI · React + Recharts · ALM Platform REST API · ASPICE PAM v4.0