Project Plan — Integrating Knowledge Graphs with Multi-Modal Machine Learning (MMML)

Glossary of Abbreviations

- **KG** – Knowledge Graph
- **MMML** – Multi-Modal Machine Learning
- **VL** – Vision–Language (used broadly for multi-modal encoders)
- **SRS** – Semantic Retention Score (composite metric for relation/type/hierarchy/attribute preservation)
- **SLO** – Service Level Objective (e.g., latency budgets)
- **p50/p95/p99** – 50th/95th/99th percentile latency
- **GNN/KGE** – Graph Neural Network / Knowledge Graph Embedding
- **ANN** – Approximate Nearest Neighbour (vector index)

1) Change Log

What changed

- Narrowed scope from broad "KG + MMML" to **open-world retrieval & zero-shot classification** with **SEC EDGAR** as the primary, free data source.
- Architectural stance updated from "single store" to a **hybrid graph–vector design** (graph as semantic spine + vector index + cache/stream), with explicit **SLOs**, fall-backs and measurable targets.
- Evaluation broadened beyond accuracy to include **SRS**, **p95/p99 latency**, simple robustness checks and rationale-style explainability (via KG paths).
- **Decision gates** and **non-goals** added to keep the technically demanding scope feasible within module time.

Why it changed

- To act on feedback requesting a tighter focus, explicit trade-offs, operational realism, interpretability and measurable contributions. The revised literature review codifies these decisions; this plan implements them.

Impact on plan

- Milestones re-sequenced to **lock evaluation policy early**, introduce the hybrid pipeline sooner, and time-box ambitious items (joint objectives, large-scale latency) behind decision gates.

2) Project Summary (Specification)

Topic. Integrating a **knowledge graph** with **multi-modal/text encoders** for open-world retrieval/zero-shot classification, using a **hybrid graph–vector** architecture that prioritises semantic fidelity, operational performance and explainability.

Problem. Pure vector models can lose **relational structure** (types, hierarchies, directionality). Pure graphs can struggle with **open-set recall** and scale. We need a **balanced system** that preserves knowledge semantics (quantified), delivers acceptable latency (SLO-aware), and remains reproducible.

Objectives.

1. **O1 – Knowledge fidelity.** Improve SRS by ≥ **25%** over a no-KG baseline while maintaining or improving task accuracy.
2. **O2 – Task performance.** Achieve ≥ **+3 pp** micro-F1 (or equivalent retrieval metric) when adding KG-features to a text baseline.
3. **O3 – Operational SLOs.** Report p50/p95/p99 latency at corpus sizes 10^3 and 10^4 with reproducible harness.
4. **O4 – Robustness & honesty.** Show ≤ **10%** drop under a small robustness stress (taxonomy off / noisy units).
5. **O5 – Reproducibility.** Fixed seeds, pinned configs/snapshots, scripted data flow and transparent artefacts.

Research questions (abridged).

- RQ1: To what extent do lightweight KG signals improve semantic retention (SRS) over a text-only model?
- RQ2: When concatenated with text features, which KG representations (one-hot/hashed concepts vs minimal KGE) deliver the best **accuracy ↔ latency** trade-off?
- RQ3: What small set of **is-a** relations (auto-taxonomy) yields the highest SRS gain per engineering hour?

Significance. Demonstrates a pragmatic, **free-data**, evaluation-first approach to hybrid search/classification with measurable semantics and SLOs—relevant to industry retrieval systems and research on structure-aware MMML.

3) Scope & Guardrails

In scope (committed).

- **Free SEC CompanyFacts → facts.jsonl → KG snapshot** (Company, Filing, Concept, Unit, Period).
- **SRS** with **AtP, HP, AP** live; **RTF** added only if embeddings/probes are introduced later.
- **Baseline text task** (TF-IDF or one frozen encoder) and **KG-as-features** (one-hot/hashed concepts; optional tiny KGE).
- **Latency harness** at **10^3–10^4** scale (vector-only first), with p50/p95/p99.
- Full **reproducibility**: seeds, configs, scripted pipeline.

Out of scope (unless time allows).

- Raw instance XBRL parsing and deep statement-path mining.
- Very-large-scale latency ($\geq 10^5$–10^6) or heavy KGE sweeps/grid searches.
- Multi-dataset fusion beyond SEC EDGAR.

4) Success Criteria & Decision Gates

Week-8 decision gates (go/no-go):

- **HP ≥ 0.25**, **AtP ≥ 0.95**, **AP ≥ 0.99**; **SRS ≥ 0.75** (weighted over available components).
- **+3 pp** micro-F1 vs text baseline when adding KG-features.
- Latency table produced at $N \in \{10^3, 10^4\}$.
  If any gate misses: pause scope expansion; invest next sprint purely in the blocking area (e.g., auto-taxonomy coverage) until green.

5) Methodology Overview

Data. SEC CompanyFacts (free). Scripts: submissions/selection, facts normalisation, filters (namespace/units/forms/year), optional reduction to latest per key.

KG build. Nodes: Company, Filing, Concept, Unit, Period. Edges: reports (Company→Filing), measured-in (Concept→Unit), for-period (Concept→Period), is-a (Concept→Concept). Concept IDs are **namespace-aware** (e.g., us-gaap:Assets).

SRS.

- **AtP:** proportion of Concept nodes with a valid measured-in link.
- **HP:** proportion of Concept nodes with a parent via is-a (coverage proxy; later HP@k with embeddings).
- **AP:** absence of erroneous reverse edges on directional types.
- **RTF:** reserved for a light probe if/when embeddings are introduced.
- Output: CSV + debug JSON with counts and per-component scores.

Baselines & KG-as-features.

- Text baseline via TF-IDF (or single frozen encoder).
- KG-features from concept presence (one-hot/hashed top-K) per document; optional tiny KGE (e.g., TransE 64–128d).
- Compare **text** vs **text+concept** using micro/macro-F1; log to CSV.

Auto-taxonomy (HP uplift).

- Conservative regex rules over **observed** us-gaap:* concepts to generate is-a pairs safely, combined with a small curated slice.

Latency harness.

- ANN index, vector-only path first; measure p50/p95/p99 with warm caches at 10^3 and 10^4; log table.

Robustness.

- Toggle taxonomy (on/off) and apply small noise to units; report ≤10% drop.

Reproducibility.

- Deterministic configs, fixed seeds, single canonical dataset location, outputs under reports/.

6) 24-Week Workplan with Milestones (Sep 2025 → Mar 2026)

Phase A — Scope & LR lock (W1–W4)

- **W1–2:** Topic specification; Evaluation Sheet v1; (ethics/data management drafted separately). **M1.**
- **W3–4:** Data pipeline; SRS definition; repo hygiene; LR freeze. **M2.**
  Status: Delivered.

Phase B — Methods build (W5–W10)

- **W5–6:** Scale facts; rebuild KG; implement SRS (AtP/HP/AP); baseline TF-IDF; KG-as-features; comparison table; seeds/configs pinned. **M3.**
  Status: Delivered.
- **W7–8: Auto-taxonomy** to lift HP; **Latency harness** (vector-only) at 10^3–10^4; decision gates (HP/AtP/AP/SRS; +3 pp micro-F1; latency table). **M4.**
- **W9–10:** Minimal **joint objective** (consistency penalty on directional edges); one ablation; document trade-offs. **M5.**

Phase C — Evaluation & results (W11–W18)

- **W11–12:** Calibration; consolidated SRS + task + p50/p95/p99; tidy outputs. **M6.**
- **W13–14:** Robustness (taxonomy off / unit noise) with ≤10% drop; hard negatives. **M7.**
- **W15–16:** Scalability exploration (two-hop + vector) in miniature; domain vs generic ablation. **M8.**
- **W17–18:** Error taxonomy; Results & Discussion (draft). **M9.**

Phase D — Final analysis & submission (W19–W24)

- **W19–20:** Repeats (≥5 seeds), CIs, simple tests; archive artefacts. **M10.**
- **W21–22:** Conclusions; decision rules; appendices (configs, seeds, eval sheet, risk log). **M11.**
- **W23–24:** 5–8 min video demo; final QA & submission. **M12.**

7) Risks & Mitigations

Risk	Likely impact	Mitigation
Sparse taxonomy coverage at scale (HP low)	Depresses SRS	Auto-taxonomy over **observed** concepts; curated additions; measure HP weekly.
Baseline too weak to show +3 pp	Ambiguous contribution	If TF-IDF underperforms, switch to **one** frozen encoder while keeping scope tight.
API throttling / gaps	Slow refresh	Polite User-Agent; 250 ms sleeps; batch overnight if expanding CIKs.
Scope creep (XML parsing, big KGE)	Schedule slip	Enforce **non-goals** and **decision gates**; do not proceed if gates fail.
Latency variance in cloud dev env	Noisy results	Warm caches; repeat runs; report medians and p95/p99 with N and hardware notes.