Can Knowledge Graphs Make AI Faster AND Smarter?

Hybrid Architecture for Semantic Preservation and Operational Efficiency • Progress Report: Week 9-10 (M5 Complete)

✅ Milestone M5 Complete: Joint Objective Tested | SRS Gate Achieved (0.7571) | λ=0.0 Identified as Optimal | Week 8 Decision Gates: 3/4 Passed

Core Research Question

Can we build a hybrid system that preserves semantic structure while maintaining fast retrieval?

Project Objectives

- 🎯 **RQ1:** Semantic Fidelity - Achieve SRS ≥ 0.75 (AtP ≥ 0.95, HP ≥ 0.25, AP ≥ 0.99)
- 🎯 **RQ2:** Operational Latency - p99 < 150ms for two-hop-plus-vector queries
- 🎯 **RQ3:** Task Effectiveness - +3pp micro-F1 improvement over text baseline
- 🎯 **RQ4:** Robustness - ≤10% performance drop under stress (taxonomy off, noise)

Core Innovation: Preserve Structure, Keep Speed

Pure Vector	Pure Graph	✨ Hybrid (Our Approach)
❌ Loses hierarchy	✅ Preserves structure	✅ Preserves structure
❌ No directionality	✅ Explainable paths	✅ Explainable paths
✅ Fast (0.1ms)	❌ Slow (150ms+)	✅ Fast (0.037ms p99)
		Best of Both Worlds

Implemented System Architecture: Hybrid Pipeline

Data Source
SEC EDGAR
CompanyFacts API
1,891 relationships
Free & public

→

Graph Spine
Semantic structure
CSV snapshots
5 node types
4 edge types
is-a taxonomy

→

Feature Layer
Multi-resolution
TF-IDF (text)
Concept indicators
Auto-taxonomy

→

Retrieval
Annoy ANN
20 trees
SVD-256 dims
p99: 0.037ms

→

Learning
Joint Model
PyTorch + sklearn
λ=0.0 optimal
Seed=42

🎯 Semantic Retention Score (SRS) Breakdown - Gate Achievement

0.7571 ✅	0.9987 ✅	0.2726 ✅	1.0000 ✅	Pending
SRS (Overall)	AtP	HP	AP	RTF
Target: ≥0.75	Target: ≥0.95	Target: ≥0.25	Target: ≥0.99	W11-12 target

2370%
Hierarchy Precision
27.3% vs. 1.15% (no tax)

4000×
Latency Margin
0.037ms vs 150ms budget

99.68%
Micro-F1
vs 98.32% (text-only)

97.9%
Robustness
2.1% drop under stress

💡 Key Discoveries (W9-10)

Concept Features Matter
KG features: 99.68% accuracy (+1.36pp), +2.27pp macro-F1

Simple Beats Complex
λ>0 penalty adds 3-4× training time for worse results

Robust Design Validated
2.1% degradation under stress confirms stability

Research Problem & Context

Vector embeddings are fast but lose semantic meaning. Knowledge graphs preserve structure but lack scalability. Neither approach works well alone.

Data Source: SEC EDGAR CompanyFacts (free, public) with 1,891 namespace-aware relationships (us-gaap:*, dei:*)

Challenge: Preserve hierarchical structure and directionality while meeting real-time latency targets

Methodology

Three Integration Patterns Tested:

- ✓ KG-as-Features: Pre-computed embeddings (baseline)
- ✓ Joint KG-MM Objectives: Shared space with constraints
- ✓ Retrieval-time Routing: Hybrid architecture (adopted)

Auto-Taxonomy: Conservative is-a edges from regex + frequency rules over observed concepts

🚀 Progress: Milestones Achieved

W1-4 (M1-M2): ☑ Complete
LR finalized, data pipeline, SRS definition

W5-6 (M3): ☑ Complete
KG built, SRS implemented, baseline + KG-features

W7-8 (M4): ☑ 3/4 Gates
Auto-taxonomy, latency harness, gates passed

W9-10 (M5): ☑ Complete
Joint objective tested, λ=0.0 optimal identified

W11-12 (M6): ⏳ Next
RTF metrics, calibration, consolidation

Bibliography & References

Key Academic Sources

1. Chen, Z. et al. (2024) 'Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey', arXiv:2402.05391
2. Ji, S. et al. (2022) 'A survey on knowledge graphs: Representation, acquisition, and applications', IEEE TNNLS, 33(2), pp. 494–514
3. Nickel, M. et al. (2016) 'A review of relational machine learning for knowledge graphs', Proceedings of the IEEE, 104(1), pp. 11–33
4. Baltrušaitis, T., Ahuja, C. and Morency, L.P. (2019) 'Multimodal machine learning: A survey', IEEE TPAMI, 41(2), pp. 423–443

Technical Implementation

5. PyTorch (2024) Deep learning framework. Available at: https://pytorch.org
6. Scikit-learn (2024) Machine learning library. Available at: https://scikit-learn.org
7. Spotify Annoy (2024) Approximate Nearest Neighbors. Available at: https://github.com/spotify/annoy
8. SEC EDGAR (2024) CompanyFacts API. Available at: https://www.sec.gov/edgar/sec-api-documentation

Project Documentation

9. Mepani, N. (2024) 'Integrating Knowledge Graphs with Multi-Modal Machine Learning: A Literature Review', Keele University MSc Project, CSC-40098
10. Nash-79 Repository (2024) 'kg-mmml: Hybrid KG-MMML Implementation', GitHub. Available at: https://github.com/Nash-79/Nash-79/tree/KG-MMML/kg-mmml

KEELE UNIVERSITY

Naresh Mepani
MSc Project Poster [CSC-40098-2024-Y1-A]
Project Supervisor: Domingo Salgado

Integrating Knowledge Graphs with Multi-Modal Machine Learning