MatterWave — KG-CAE (Draft 1.0.1)

MatterWave — KG-CAE

Problem & Motivation

- LLMs often generate confident but incorrect assertions when asked about entities or facts.
- Existing entity linking datasets are limited in scale and diversity; manual labeling is costly.
- Goal: Build an automatic pipeline to produce KG-grounded examples and train a model that encodes text in a KG-aware embedding space for robust retrieval and entity disambiguation.

Model Overview

- KG-CAE: a single Transformer encoder with two outputs — a contextual embedding and a URI classifier — trained jointly so embeddings and predicted URIs align.
- Training objective: composite loss mixing a Triplet Margin Loss (factual consistency) with Cross-Entropy over subject URI IDs; optional knowledge-aware regularizers (KACR, MHSC) and curriculum-based dynamic negative hardening (DNH).
- Optional quantized bottleneck: RQ-VAE style quantization for compact, interpretable semantic IDs; useful for deployment and artifact reproducibility.

Data Pipeline

- Sources: DBpedia (RDF), multilingual Wikipedia abstracts, and optional Wikidata enrichment.
- Steps: entity discovery via SPARQL → 2-hop RDF enrichment → text aggregation and paraphrase augmentation → negative generation (entity swaps and graph-based negatives) → store in Apache Iceberg for versioning and reproducibility.
- For a formal specification of the theoretical framing and experimental recipes, see `Theoretical_Framing_and_Expected_Behaviors.md`. For the publication-oriented experiment plan and timeline, see `RESEARCH_PUBLICATION_PLAN.md`.

Experimental Protocol

- Datasets: AIDA-CoNLL, MSNBC, ACE2004, custom DBpedia-derived splits, FEVER for fact-checking.
- Metrics: Precision@1/5, MRR, MAP, nDCG, AUC-ROC for binary downstream tasks.
- Statistical tests: paired bootstrap, Wilcoxon signed-rank, BH correction.

Key Results (placeholder)

- Replace with final tables/figures after experiments. Use provided templates in `results_templates/`.

Ablations & Interpretability

- Ablations: classification head off, contrastive only, quantization on/off, negative-sample hardness.
- Interpretability: attention inspection, attribution on token spans, counterfactual edits, prototype nearest neighbors.

Limitations & Ethics

- License concerns with downstream data; ensure CC BY-SA compliance for Wikipedia-derived content.
- PII leakage risk when aggregating context; use redaction heuristics and human audit where necessary.

Next steps

- 1. Finalize dataset packaging and sample snapshot (50k-anchor DBpedia sample) and publish an artifact snapshot.
- 2. Run baseline and prioritized ablations per `RESEARCH_PUBLICATION_PLAN.md` (KACR first, then DNH/MHSC).
- 3. Prepare publication materials (figures, hyperparameter appendix, artifact README) and plan for artifact submission.

Acknowledgments

- Contributors and funding notes (populate `ACKS.md`).

Appendix: Speaker notes

- See `notes.md` for slide-by-slide speaker notes, linked examples, and references to the theoretical and publication docs.