A Trust-Aware Multimodal GenAI Digital Twin for UMKC Campus Information & Decision Support

Team Name: *HappyGroup*

Team Members:
Ailing Nan (GitHub: https://github.com/ailingnan)
Lyza Iamrache (GitHub: https://github.com/LyzaIamrache)
Gia Huynh (GitHub: https://github.com/JimHuynh25)

Group Repository: https://github.com/ailingnan/CS5588-SmartCampus

Primary Contact: *Ailing Nan*

1. Problem Statement & Real-World Impact

Universities publish a large volume of heterogeneous information, including policy documents, campus maps, transportation schedules, pricing tables, and annual reports. While these resources are publicly available, students and visitors frequently struggle to locate accurate, up-to-date, and trustworthy answers to practical questions such as where to park, how to navigate campus, or whether they can catch a shuttle at a given time.

Existing solutions—static websites, keyword-based PDF search, or generic chatbots—often return isolated facts without context, fail to integrate multiple modalities (text, tables, images), or generate plausible but unsupported answers. This creates friction in everyday decision-making and reduces trust in AI-assisted systems.

Our project addresses this gap by proposing a GenAI-dri3ven, multimodal digital twin of UMKC campus information. Rather than acting as a traditional chatbot, the system is designed to support decision-oriented workflows by grounding responses in authoritative documents, exposing evidence transparently, and explicitly refusing to answer when required information is missing.

If deployed, this system would reduce time-to-information, improve trust through citation and refusal behavior, and lower organizational risk from misinformation.

2. Project Direction & Digital Twin Framing

2.1 Beyond a Traditional Chatbot

A core motivation of this project is to move beyond the limitations of traditional chatbots, which typically:

- answer what something is, but not what will happen if a user takes a certain action,
- ignore temporal and contextual constraints (e.g., current time, applicability of rules),
- treat space as static text rather than an actionable environment.

2.2 Digital Twin Perspective

In this project, we interpret a digital twin as a dynamic, query-driven simulation layer built on top of authoritative campus data. The twin mirrors institutional knowledge (maps, schedules, policies) and enable short-horizon, rule-constrained reasoning over realistic scenarios.

Example decision-oriented questions include:

- If I do not enroll in this course this semester, will it delay my graduation timeline?
- If I leave now, can I realistically catch the next shuttle bus?
- Where exactly is a classroom located, and how can I navigate there efficiently?

Rather than long-term forecasting, the digital twin focuses on near-term feasibility and consequences, combining retrieved rules and constraints with LLM-based reasoning. This framing aligns with real-world expectations of reliability and interpretability.

3. GenAI System Architecture & Pipeline

3.1 Data Ingestion & Knowledge Layer

The system ingests a diverse set of campus artifacts, including:

- PDFs (student handbook, parking permits, shuttle schedules, catalogs),
- images and figures (campus maps),
- tables embedded within documents.

Processing steps include:

- PDF text extraction,
- OCR applied to images and scanned tables,
- page-level chunking with metadata (document name, page number, modality).

All content is versioned and linked to its provenance to support traceability and governance.

3.2 Retrieval, RAG Design & Grounded Generation

The system employs a hybrid RAG architecture:

- Sparse retrieval (BM25) for keyword-sensitive policy documents,
- Dense retrieval (FAISS embeddings) for semantic similarity,
- Hybrid fusion to improve evidence coverage across modalities,
- Optional reranking for top-k results.

Generation follows strict grounded QA rules:

- answers must cite retrieved evidence,
- missing required evidence triggers partial grounding labels or explicit refusal.

This design prioritizes trust calibration over fluent but unsupported responses.

3.3 Scenario Reasoning & Short-Horizon Simulation

To support digital twin behavior, the system enables scenario-based reasoning by combining retrieved documents with LLM reasoning. For example:

- shuttle feasibility questions integrate current time and schedule constraints,
- academic planning questions reference catalog rules and prerequisites.

Predictions are intentionally limited to short-horizon, rule-constrained outcomes, avoiding overclaiming long-term accuracy.

3.4 Interface & Human-in-the-Loop

In Phase 1, the system is demonstrated via an interactive notebook interface that surfaces:
- retrieved evidence packs,
- citations,
- refusal messages.

This design allows users to verify outputs and supports future extensions to chat interfaces or campus-facing applications.

4. Data Sources

4.1 University Catalog 2025–2026

1. **Direct dataset link:**
   https://catalog.umkc.edu/pdf/2025-2026%20University%20Catalog_Archived%209-10-25.pdf

2. **Description:**
   The UMKC 2025–2026 University Catalog is the official academic catalog for the University of Missouri–Kansas City for the 2025–2026 academic year. It includes comprehensive program information, course requirements, academic policies, degree requirements, and academic regulations for undergraduate and graduate programs. Program listings cover majors, minors, required credits, curriculum descriptions, and degree pathways.
3. **Modality:**
   PDF / structured text and tables
4. **Use:**
   1. Retrieval & Grounding: Used in the RAG pipeline to answer program-related queries (e.g. “What are the requirements for a Psychology major?”).
   2. Fine-Tuning / Benchmarking: Provides rich structured and semi-structured data to fine-tune or evaluate the model’s ability to extract program requirements and curriculum details.
   3. Governance & Policy: Supports responses about academic rules and requirements.
   4. Acts as a key reference document for campus information queries beyond basic facts.
5. **Known limitations:**
   The catalog is updated annually and may not reflect mid-year changes (e.g., new courses or policy revisions that occur after publication).

   It is large and dense, requiring robust chunking and metadata tagging to make retrieval efficient.

4.2 UMKC Campus Maps

- **Direct dataset link:https://www.umkc.edu/documents/umkc-health-sciences-campus-map.pdf**

  https://www.umkc.edu/documents/umkc-volker-campus-map.pdf

- **Description:** PDF, includes building layouts, campus paths, and landmarks.
- **Use:**
  - RAG: Provide visual grounding for navigation queries (e.g., “Where is the library?”).
  - Fine-tuning: Teach the model to interpret campus spatial layouts.
  - Benchmarking: Test query responses against official maps for correctness.
- **Known limitations:** Static snapshot; may not reflect recent construction or temporary closures.

4.3 Shuttle Bus Schedules

- **Direct dataset link:** https://www.umkc.edu/transportation/docs/2026-spring-shuttle-schedule.pdf
- **Description:** PDF, contains timetable data with multiple routes, stops, and operating hours.
- **Use:**
  - RAG: Answer shuttle-related queries with grounding in official schedules.
  - Prediction: Train a model to forecast shuttle occupancy or arrival delays.
  - Benchmarking: Compare predicted vs. actual shuttle availability.
- **Known limitations:** Schedule updates only per semester; may not account for temporary changes or delays.

4.4 Parking Permits & Pricing Documents

- **Direct dataset link:** https://www.umkc.edu/parking/parking-options/student-permits.html
- **Description:** Web page/PDF, lot availability, cost, and permit types.
- **Use:**
  - RAG: Answer parking-related queries with evidence from official documents.
  - Prediction: Forecast parking demand for different lots or time periods.
  - Fine-tuning: Improve the model’s ability to interpret numeric tables.
- **Known limitations:** Policies may change mid-year; coverage limited to student permits.

4.5 Student Handbook / Annual Reports

- **Direct dataset link:** **https://www.umkc.edu/student-affairs/docs/umkc-student-handbook.pdf**
- **Description:** PDF, 57 pages, text, includes rules, policies, and procedures.
- **Use:**
  - RAG: Ground responses for policy or procedure questions.
  - Fine-tuning: Teach LLM domain-specific language and structure.
  - Benchmarking: Verify accuracy and trustworthiness of policy-based answers.
- **Known limitations:** Updates yearly; may contain ambiguities that require human verification.

4.6 Visual Identity Guidelines

- **Direct dataset link:** https://www.umkc.edu/mcom/docs/visual-identity-guidelines.pdf
- **Description:** PDF with design guidelines, logos, and branding rules.
- **Use:**
  - Interface design: Ensure generated outputs and dashboards comply with branding.
  - Fine-tuning: Train model to recognize design terms and guidelines.
- **Known limitations:** Mostly static; does not cover every future branding scenario.

4.7 Campus Crime Report

- **Direct dataset link: https://www.umkc.edu/police/docs/2025ccfsr.pdf**
- **Description:** PDF containing campus crime statistics, Clery Act reports, tables, and figures.
- **Use:**
- RAG / Retrieval: Provide evidence-backed answers for safety- and security-related campus questions.
- Benchmarking: Test the model's ability to summarize and accurately cite numeric and tabular data**.**
- **Known limitations:** Updated annually; may not reflect real-time crime incidents or minor infractions.

5. Related Work (GenAI Systems & Evaluation)

5.1 GFM-RAG: Graph Foundation Model for Retrieval-Augmented Generation

Paper Source: NeurIPS 2025 Proceedings (listing of accepted papers including *GFM-RAG*)
Paper Listing: *GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation*
DOI: https://doi.org/10.48550/arXiv.2502.01113
Summary: The NeurIPS 2025 program includes *GFM-RAG*, a retrieval-augmented generation model that fuses graph-based representations with RAG to support structured reasoning on complex data. This approach enhances retrieval and grounding by incorporating graph structures into the retrieval process.
Relevance: Incorporating graph-based retrieval insights parallels our interest in **multimodal campus information grounding** (maps, schedules, policies) and highlights how structured retrieval can improve answer fidelity and context coherence, informing future enhancements to our RAG knowledge layer.

5.2 KnowTrace: Iterative Retrieval-Augmented Generation with Structured Knowledge Tracing

Paper: *KnowTrace: Bootstrapping Iterative Retrieval-Augmented Generation with Structured Knowledge Tracing*
KDD 2025 Proceedings: DOI: https://doi.org/10.1145/3711896.3737015
Summary: KnowTrace proposes an iterative RAG method that uses structured knowledge tracing to successively refine retrieved context and improve generation over multiple passes. The paper addresses challenges in aligning retrieved evidence with generation, especially under complex query dependencies.
Relevance: This work directly informs our **hybrid retrieval and evidence pack design**,

especially for iterative refinement of answers where multiple related campus documents must be correlated (e.g., parking rules + enforcement policies + maps), and reinforces our use of structured retrieval strategies.

5.3 Multi-Stage Verification-Centric Framework for Mitigating Hallucination in Multimodal RAG

Paper: *Multi-Stage Verification-Centric Framework for Mitigating Hallucination in Multi-Modal RAG* — KDD Cup 2025 Meta CRAG-MM Benchmark
Code: https://github.com/Breezelled/KDD-Cup-2025-Meta-CRAG-MM
Summary: This work presents a verification-centric multimodal RAG framework that integrates query routing, dual retrieval paths, and post-hoc verification to minimize hallucination in multimodal document QA. The framework was evaluated in the CRAG-MM challenge and provides a robust strategy for prioritizing factual consistency over completeness in generated answers.
Relevance: Our Smart Campus Assistant similarly targets **minimizing hallucination and enforcing evidence citations**, especially in multimodal settings (PDFs, maps, and OCR content). The verification pipeline from this paper guides our refusal logic and factual consistency evaluation plan.

6. Team Roles

6.1 Lyza Iamrache: Product Lead — Stakeholder Framing & Deployment Vision
Responsibilities:
- Defines the primary stakeholder personas (e.g., students, visitors, campus administrators) and the real-world decisions the system is designed to support.
- Articulates the product value proposition, including how the GenAI system improves decision quality, trust, and usability compared to traditional chatbots or search tools.
- Owns success metrics from a product perspective (e.g., time-to-information, perceived trust, usability).
- Leads the deployment vision, including the “If we shipped this” plan, outlining how the system could be integrated into existing campus workflows or user-facing interfaces.
- Ensures alignment between system capabilities and stakeholder expectations, preventing over-claiming of model intelligence.

6.2 Ailing Nan: Systems Lead — Multimodal Ingestion & RAG Pipeline
Responsibilities:
- Designs and implements the end-to-end multimodal GenAI pipeline, including PDF ingestion, OCR for images and tables, and metadata-aware chunking.
- Develops and maintains the retrieval architecture (BM25, dense embeddings, hybrid fusion, and optional reranking).
- Implements grounded generation logic, including citation enforcement and refusal behavior when required evidence is missing.
- Conducts system ablations and tradeoff analyses (e.g., dense vs. sparse vs. hybrid retrieval) to inform design decisions.
- Ensures the system architecture supports future extensions such as scenario reasoning and digital twin–style simulation.

6.3 Gia Huynh: Evaluation & Risk Lead — Metrics, Failure Analysis & Governance

Responsibilities:

- Designs and executes the system-level evaluation framework, including retrieval metrics (P@5, R@10), faithfulness labels, and failure case analysis.
- Analyzes hallucination risks and under-grounded answers, identifying where plausible outputs may lack sufficient evidence.
- Leads comparative evaluations against baseline systems (e.g., LLM without retrieval or grounding constraints).
- Develops governance and mitigation strategies, such as modality-aware gating, stricter citation requirements, and human-in-the-loop review.
- Documents risks, limitations, and ethical considerations to ensure the system is suitable for real-world deployment.

7. Methods, Technologies & Tools

This project adopts a **lightweight yet extensible GenAI system stack**, prioritizing transparency, modularity, and reproducibility over excessive model complexity. Tool choices are guided by the need to support rapid iteration, evidence grounding, and system-level evaluation.

- **LLMs:**
  OpenAI API is used for answer generation due to its strong instruction-following capability and stable integration with retrieval-augmented generation workflows.
- **Retrieval & Indexing:**
  Sparse retrieval (BM25) and dense retrieval (FAISS-based embeddings) are combined using hybrid fusion to balance keyword precision and semantic recall.
- **Multimodal Parsing & OCR:**
  PyMuPDF is used for PDF text extraction and page-level metadata tracking. OCR tools are applied to campus maps, tables, and visual documents to enable multimodal grounding.
- **Development Environment:**
  Python and Jupyter Notebooks are used for rapid prototyping, debugging, and transparent demonstration of system behavior.
- **Version Control & Reproducibility:**
  GitHub is used to manage code, data references, evaluation artifacts, and documentation, ensuring all system components are versioned and auditable.
- **Evaluation Framework:**
  Manual grounding analysis, retrieval metrics, and structured failure documentation are employed to assess system reliability and trustworthiness.

8. Evaluation & Impact Plan

Evaluation focuses on **system-level quality, grounding, and trust**, rather than answer plausibility or accuracy alone. The goal is to assess whether the GenAI system supports reliable decision-making in real-world campus workflows.

8.1 Evaluation Dimensions

- **Retrieval Quality:**
  Precision@5 (P@5) and Recall@10 (R@10) are used to evaluate whether relevant documents are retrieved at top ranks.
- **Grounding & Hallucination Analysis:**
  Citation correctness and unsupported claim rate are analyzed to identify cases where answers appear plausible but lack sufficient evidence.
- **Faithfulness Labels:**
  Each task output is manually categorized as:
  Yes (fully supported by retrieved evidence),
  Partial (plausible but missing required evidence),
  Refusal (correctly declines to answer due to insufficient evidence).
- **Failure Analysis:**
  Failure cases are documented to understand organizational and user risk, particularly in scenarios involving spatial reasoning or service availability.

8.2 Baselines

System performance is compared against:
- an off-the-shelf LLM without retrieval, and
- a retrieval-based system without grounding or refusal constraints.

These comparisons highlight the impact of grounding and governance mechanisms on trust and reliability.

9. Expected Outcomes & Deliverables

By the end of the project, we expect to deliver:

- a working **multimodal GenAI digital twin prototype** for campus information and decision support,
- a **live demo** illustrating stakeholder workflows such as navigation, transportation feasibility, and planning queries,
- a fully versioned **GitHub repository** containing data references, prompts, pipelines, evaluation scripts, and documentation,
- a **final technical report** describing system architecture, experiments, evaluation results, and lessons learned.

These deliverables are designed to demonstrate both technical feasibility and real-world applicability.

10. Phase 1 Plan & Milestones

Phase 1 (Weeks 1–3) focuses on establishing a reliable and evaluable system foundation.

Key objectives include:
- refining RAG grounding and explicit refusal behavior,
- formalizing evaluation and trust metrics,
- maintaining structured and versioned artifacts in the team GitHub repository.

Concrete deliverables for Phase 1 include:

- a stabilized multimodal RAG pipeline,
- initial retrieval and grounding evaluation results,
- a finalized project proposal and repository structure.

11. Closing Statement

This project treats GenAI as a decision-support system whose value depends on trust, grounding, and responsible deployment, not merely fluent text generation. By explicitly modeling failure cases and refusal behavior, our digital twin aims to reflect real-world GenAI product constraints rather than idealized performance.