FACULTY OF COMPUTER SCIENCE AND ENGINEERING
National University of Computer and Emerging Science, Islamabad

Subject: Agenti AI (MS-DS) **Instructor: Dr. Usama**

Project Proposal

1. Project Title

Cricket Strategy Intelligence: Retrieval-Augmented Multi-Modal Analytics for Context-Aware Tactical Decision Support

2. Team Information

Student Name: Muhammad Usman Arif
Registration Number: 24i-8001
Email i248001@isb.nu.edu.pk

Student Name: Syed Shahbaz Ali Shah
Registration Number 24i-8002
Email i248002@isb.nu.edu.pk

Student Name: Hussain Abdullah
Registration Number(s): 24i-7612
Email: i247612@isb.nu.edu.pk

3. Problem Statement

Aspect	Details
Context of the Problem	Modern cricket strategy depends on synthesizing heterogeneous data: structured ball-by-ball logs, player/venue statistics, and unstructured sources (live commentary, post-match reports, analyst notes). Current analytics dashboards surface raw stats but lack context-aware tactical synthesis (e.g., "optimal bowling change given match phase, pitch trend, and batter style").
Importance	**1) Tactical Optimization:** Supports captains/coaches with situational insight (powerplay vs death overs). **2) Player Match-Up Intelligence:** Informs selection and real-time adjustments (e.g., left-arm pace vs top-order right-handers). **3) Broadcast / Fan Engagement:** Enhances narrative quality with evidence-backed micro-analyses. **4) Performance & Scouting:** Identifies emerging patterns (e.g., decline vs wrist spin at specific venues).
Challenges	• Data Heterogeneity (JSON/CSV/text commentary). • Noisy Commentary (colloquial language, sarcasm). • Entity Resolution (aliases, spelling variants of players/venues). • Temporal Context (overs, innings, phase segmentation). • Sparse Situational Samples (rare events like super overs). • Evaluation Difficulty (ground truth for "tactical insight" is subjective).

Aspect	Details
	• Latency & Cost (LLM inference + retrieval at query time). • Multi-Provider Variability (model drift across OpenAI/Groq/Gemini).
Risks	Inconsistent retrieval relevance, hallucinated tactical advice, ambiguous evaluation metrics for "quality of insight," and vendor lock-in if not abstracted.

4. Objectives

#	Objective
1	Unified ingestion & preprocessing (structured + unstructured).
2	Efficient cricket RAG (chunking + metadata filters).
3	Multi-provider LLM abstraction (OpenAI/Groq/Gemini).
4	LangGraph multi-hop tactical reasoning.
5	Evaluation suite (retrieval, grounding, rubric).
6	Latency & cost optimization (caching, reuse).
7	Reproducible insights (API + notebook).

5. Dataset(s)

Aspect	Details
Primary Structured	Cricsheet ball-by-ball (ODI, T20I, Tests, domestic T20)
Supplementary Stats	Kaggle Statsguru-derived aggregates
Unstructured	ESPN Cricinfo commentary, match reports, analyst summaries
Scale	2–5M deliveries; 0.5–1.2M commentary lines; 10–20K reports
Structured Features	Over, innings, players, dismissal, runs, derived phase, pressure
Unstructured Features	Text, timestamp, over marker, optional tone
Derived Labels	Phase, roles, bowler style, venue bias, run rate deltas
Quality Considerations	Noise, historic gaps, name variants

7. Proposed Methodology

- Ingestion & Normalization
  - Description: Load & unify structured + commentary + reports
  - Key Output: Unified parquet + mappings
- Derived Feature Engineering

    - Description: Phase, pressure, style, rotation stats
    - Key Output: Enriched dataset
- Chunking & Indexing
    - Description: Domain-aware splitting + metadata
    - Key Output: Persisted vector store
- Embeddings
    - Description: MiniLM baseline (upgradeable)
    - Key Output: Embedding cache
- Retrieval Layer
    - Description: Semantic + metadata filtering
    - Key Output: Retriever API
- Multi-Provider LLM Layer
    - Description: Provider-agnostic abstraction
    - Key Output: Pluggable LLM interface
- LangGraph Reasoning
    - Description: Multi-hop tactical pipeline
    - Key Output: Executable graph
- Prompt Engineering
    - Description: Structured tactical templates
    - Key Output: Versioned prompts
- Evaluation Framework
    - Description: Metrics + rubric + ground-truth queries
    - Key Output: Metrics dashboard
- API & Notebook
    - Description: FastAPI + reproducible walkthrough
    - Key Output: Usability layer
- Optimization
    - Description: Caching, graceful degradation
    - Key Output: Lower latency / resilience

8. Expected Outcomes

- High-Quality Retrieval
    - Indicator: Hit@5 ≥0.80
- Factual Grounding
    - Indicator: ≥85% claims source-backed
- Tactical Relevance
    - Indicator: Expert rubric ≥3.5/5
- Latency Control
    - Indicator: Median <3.0s
- Provider Flexibility
    - Indicator: Hot-swap via env only
- Robust Error Handling
    - Indicator: 0 unhandled exceptions (stress test)
- Insight Diversity
    - Indicator: ≥25% more contextual factors vs baseline