**Research Progress Report**

**Title:**

Detecting Narrative Bias in LLM-Generated IPL 2024 Analyses

**Objective:**

To collect, analyse, and validate data-driven narratives from multiple large language models (LLMs) on IPL 2024 team performance. The goal is to measure factual consistency, detect framing bias, and evaluate each model’s reliability against ground-truth statistics.

**1. Experiment Design**

* Queried **GPT-4**, **Claude 3**, and **Gemini 1.5** with three prompt framings:
  + *Neutral framing*: factual description only.
  + *Positive framing*: emphasizing team success.
  + *Critical framing*: emphasizing weaknesses.
* Each prompt variant was asked **5 times** to control for temperature-based randomness.

**2. Data Integrity**

* Responses saved with unique IDs and SHA-256 hash to ensure traceability.
* Bias-testing metadata included “model prompt condition” and “frame ID”.

**Quantitative Analysis**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Method** | **Finding Summary** |
| Entity mentions | Frequency count (Panda’s tokenization) | RCB & CSK appeared in >80 % responses → potential model focus bias |
| Sentiment polarity | VADER + manual cross-check | GPT-4 showed balanced tone; Claude slightly positive; Gemini more critical |
| Recommendation type | Rule-based classifier | 60 % team-level vs 40 % individual focus |
| Significance test | χ² test (p < 0.05) on sentiment counts | Statistically significant difference between framing conditions |

**2. Qualitative Analysis**

* **Pattern of language:**  
  GPT-4 used neutral performance phrasing; Claude added optimistic narratives; Gemini highlighted risks.
* **Hallucination checks:**  
  Gemini occasionally added non-existent player records; flagged as fabrication rate ≈ 12 %.
* **Selective data use:**  
  Claude omitted home-advantage data for Gujarat in some responses → narrative bias identified.

**3. Validation Against Ground Truth**

* Compared each response against verified dataset from Task 6 and Task 7 using Python/Pandas scripts.
* **Contradiction rate:** ≈ 8 % (GPT-4), 14 % (Claude), 19 % (Gemini).
* Fabrication flagging logic captured misstated statistics or invented players.