Technical Appendix: Provider Profile & Peer-Group Database

To facilitate advanced forensic reasoning in the Foundation Model, a specialized **Provider Profile Database** was engineered. This database acts as a "Semantic Bridge," collapsing granular service-line claims into high-density NPI-level summaries optimized for **Retrieval-Augmented Generation (RAG)**.

1. Aggregation & Dimensionality Reduction

The raw CMS Medicare Provider Utilization data is transformed from service-level granularity (HCPCS) to unique provider profiles.

- **Total Providers:** 1,535,547 unique providers aggregated by National Provider Identifier (NPI).
- **Volume Metrics:** Sum of total services and distinct beneficiaries to establish the provider's operational scale.
- **Financial Benchmarking:** Calculation of both **Mean** and **Max** submitted charges.
- **Forensic Note:** Retaining the "Max" charge is a critical forensic requirement for detecting "unbundling" schemes where fraudulent costs are hidden within high-volume claims.
- **Geographic Footprint:** Unique count of states billed to identify "multi-latitude" anomalies.

2. Forensic Feature Engineering

The database computes normalized "Forensic Features" that serve as primary indicators for the LLM "Prosecutor" agent.

- **The Greed Ratio (`charge_Ratio`):** Isolates the specific markup of a provider's submitted charges against Medicare allowed amounts.
- **The Risk Score (`Type_risk_score`):** Assigns a prior probability of fraud based on historical exclusion rates within the provider's specific medical specialty.

3. Statistical Peer-Group Benchmarking

To prevent the "Specialist Bias" observed in traditional models (Experiment I), the database establishes what "normal" looks like at the provider-type level, allowing outliers to stand out mathematically.

Provider Identity & Scale

- **Rndrng_Prvdr_Type:** The primary grouping key (e.g., Cardiology, Internal Medicine) used because billing patterns for surgeons differ naturally from physical therapists.
- **Provider_Count:** The total number of unique NPIs within a specialty, serving as a reliability metric for statistical stability.

Absolute Billing Metrics

- **Avg_Total_Charge:** The arithmetic mean of submitted charges for the group, representing the "Average Annual Ask" for the field.
- **Charge_StdDev:** Measures the "spread" in billing; high deviation suggests widely different volumes, while low deviation indicates highly similar billing.
- **Suspicious_Charge_Threshold:** Calculated as **Avg + (2 * StdDev)**. Providers billing above this number are in the top 2.5% of their peers and are primary candidates for further investigation.

Relative Markup Metrics

- **Avg_Charge_Ratio:** The average "Markup" (Submitted Charge / Medicare Allowed Amount). A value of 5.0 indicates the specialty typically asks for 5x the Medicare payment.
- **Ratio_StdDev:** Measures markup variation; low deviation could signal "industry standard" pricing.
- **Suspicious_Ratio_Threshold:** A custom "Markup Ceiling." Ratios exceeding this threshold identify providers asking for significantly more money relative to the fee schedule than their peers, a red flag for aggressive revenue cycling.

The Ground Truth Label

- **Spec_Fraud_Rate:** The percentage of providers in a specific specialty appearing on the LEIE (Exclusion List).

4. Computational Offloading

By pre-calculating these baselines and individual **Z-scores** (statistical distance from the peer group), the architecture offloads mathematical heavy lifting to Python. This allows the Foundation Model to focus exclusively on interpreting the **semantic meaning** of the anomaly rather than performing raw arithmetic.