# AFL SGM Roadmap v0.1

(This is the markdown version of the living roadmap canvas)

**AFL SGM Project — Roadmap & Data Architecture (Living Doc)**

**Owner:** Ethan  
**Agents:** SGM Analysis Agent (engine + assistant)  
**Versioning:** Update this doc each time we change data, models, or outputs (see Change Log at bottom).  
**Purpose:** Single source of truth for how the system works now and where it’s going.

**0) North Star & Operating Model**

* **Chosen approach:** **Option B — Hybrid: Engine + Agent.**
  + **Engine** (deterministic): ETL → modelling → SGM generation → exports.
  + **Agent** (adaptive): data QA, fuzzy matching, correlation checks, rationale, weekly deltas, and research.
* **Cadence:** Nightly ETL snapshot; game‑day refresh 3–4 hours pre‑bounce; final **closing odds** archive at bounce‑time.
* **Success metrics:**
  + Calibration (Brier / log loss) for match/totals; hit‑rate and realized EV for SGMs; % data pipelines passing QA.

**1) Data Architecture (with Built‑in Archiving)**

**Layering: Bronze → Silver → Gold**

**Bronze (Raw, immutable, append‑only)**

* **Storage:** Parquet (preferred) or CSV, partitioned by snapshot\_ts and source.
* **Write policy:** Never overwrite; every run creates a new snapshot\_ts=YYYYMMDD\_HHMM folder.
* **Tables:**
  + odds\_raw (all books/markets; raw text kept)
  + weather\_raw (as scraped)
  + lineups\_raw (as scraped)
  + matches\_raw (historical results; when updated, store a new snapshot)
* **Keys:** (source, source\_event\_id, market\_group, market\_name, selection\_raw, snapshot\_ts).

**Silver (Normalized, tidy)**

* **Transform:** Type cleaning, mapping to internal keys, de‑duplication, **overround calculation & unvigging**.
* **Tables:**
  + games (game\_id, home, away, venue, kickoff\_utc, round, season)
  + markets (market\_id, market\_code, market\_name\_std; reference)
  + odds\_latest (game\_id, market\_id, selection\_id, line, price, **implied\_prob**, **implied\_prob\_unvig**, bookmaker, latest\_snapshot\_ts)
  + lineups (game\_id, team\_id, player\_id, player\_name\_std, status, source, latest\_snapshot\_ts)
  + weather (game\_id, temp\_c, rain\_prob, rain\_mm, wind\_kmh, gust\_kmh, msl\_hpa, cloud\_pct, latest\_snapshot\_ts)
* **Archiving detail:** Keep **all bronze snapshots**; in silver keep latest\_snapshot\_ts per entity + a monthly roll‑up archive for audits.

**Gold (Model‑ready features & outputs)**

* **Feature store:**
  + features\_game (venue baseline totals, recent form, rest days, travel distance, etc.)
  + features\_player (rates per 100 TOG, last n games, opponent/venue splits where stable)
* **Model outputs:**
  + pred\_match (home\_win\_prob, draw\_prob if modelled, calibration bins)
  + pred\_totals (mean\_total, std\_total, P(over x), sensitivity to weather)
  + pred\_player\_props (per market threshold, model\_prob)
* **SGM outputs:**
  + sgm\_candidates (game\_id, legs[], model\_prob, price\_vector, correlation\_flag)
  + sgm\_ranked (combined\_price, combined\_prob, EV, stake\_suggestion, rationale)
* **Archiving detail:** Persist **every run’s gold outputs** under run\_id (timestamp). Tag which run generated the weekly report.

**Closing Lines & Backtesting Archive (critical)**

* Capture a **closing odds snapshot** at scheduled "T‑0" (bounce) for every market we model.
* Persist as odds\_closing in bronze, and propagate to odds\_latest\_closing in silver for comparison.
* Backtesting tables:
  + backtest\_round (run\_id, round, season, generated\_on, models\_used)
  + backtest\_sgms (run\_id, game\_id, legs, model\_prob, price\_at\_gen, **price\_at\_close**, outcome, realized\_ev)
  + backtest\_metrics (per round & cumulative: Brier, log loss, hit‑rate, ROI proxy)

**2) Core Schemas (concise)**

**games**

* game\_id (pk), season, round, home, away, venue, kickoff\_utc, kickoff\_local\_awst

**odds\_latest**

* game\_id, bookmaker, market\_code, selection\_id, selection\_name\_std, line, price,  
  implied\_prob (=1/price), market\_overround, implied\_prob\_unvig (=implied\_prob / overround\_sum), latest\_snapshot\_ts

**weather**

* game\_id, temp\_c, rain\_prob, rain\_mm, wind\_kmh, gust\_kmh, msl\_hpa, cloud\_pct, latest\_snapshot\_ts

**lineups**

* game\_id, team, player\_id, player\_name\_std, status (in/out/sub/emergency), source, latest\_snapshot\_ts

**pred\_match / pred\_totals / pred\_player\_props**

* Include model\_name, model\_version, train\_window, run\_id, and features\_hash for reproducibility.

**sgm\_ranked**

* run\_id, game\_id, sgm\_text, legs[], combined\_prob, combined\_price, EV, stake, correlation\_adj, rationale

**3) ETL & Orchestration**

* **Scheduler:** cron / Windows Task Scheduler / GitHub Actions.
* **Job order:** scrape→bronze→silver→features→models→SGMs→exports→Excel refresh.
* **Idempotency:** each run writes a new bronze snapshot; silver/gold keyed by latest and by run\_id.
* **Retries & logging:** exponential backoff, structured logs, error budget and alerting (email/slack).

**4) Modelling Plan (minimum viable → advanced)**

* **Match result:** logistic/Elo hybrid; evaluate vs closing H2H.
* **Totals:** venue baseline + weather deltas (rain/wind); move to neg‑bin if over‑dispersion.
* **Player props:** neg‑bin (disposals/tackles), Poisson/ZIP (goals); ingest role/TOG if available.
* **Correlation:** start with independence × cap; progress to Monte Carlo joint simulation or copula‑lite.
* **Calibration:** reliability curves; isotonic/platt if needed.

**5) Agent Responsibilities (durable value)**

* Fuzzy match: lineup names ↔ historical players, assign player\_id.
* QA: overround sanity checks, venue mismatches, missing lines.
* Correlation guardrails: reject/adjust over‑correlated SGMs.
* Explainability: produce short Excel‑friendly tables + rationale per SGM.
* Weekly change log: “what changed” in features, data, or model outputs.

**6) Source Expansion (gaps to fill)**

* **Injuries & role changes:** scrape official club updates; store as injury\_reports (bronze) → injury\_events (silver).
* **Travel/rest:** compute features from fixtures; add to features\_game.
* **Consensus/closing lines:** at least one additional bookmaker feed for robustness.
* **Player minutes/TOG:** if available; otherwise approximate via role proxies.

**7) Excel/PowerQuery Interface (front‑end only)**

* Connect to **gold**:
  + sgm\_ranked, pred\_match, pred\_totals, pred\_player\_props, backtest\_metrics.
* Dashboards: Round summary, Top SGMs, EV & hit‑rate, Calibration plots (bucket tables).
* No scraping or heavy transforms in Excel.

**8) Governance & Versioning**

* **Change control:** every model/data change increments model\_version and logs in Change Log.
* **Repro:** every export tagged with run\_id, model\_version, and features\_hash.
* **Rollbacks:** keep last N gold runs; allow pinning a prior run for reporting.

**9) 90‑Day Milestones (high‑level)**

* **Weeks 1–2:** finalize schemas; implement overround removal and name reconciliation; wire Excel to gold.
* **Weeks 3–6:** totals & match models v1; SGM generator v1; backtest harness v0.
* **Weeks 7–10:** player prop models; correlation adj v2 (Monte Carlo); backtest metrics v1.
* **Weeks 11–13:** agent QA & rationale automation; drift alerts; publish weekly pack.

**10) How to Use/Store This Plan**

* Keep this as the **living roadmap**. Update as you add sources, fields, or models.
* Optionally mirror in Google Drive (doc or markdown) and in a Git repo for version history.
* Reference this doc’s **section anchors** when asking the agent to update/execute tasks.

**Change Log**

* **v0.1 (today):** Initial roadmap + schemas + archiving strategy.