NFL Dissertation + System-of-Systems — Master TODOs

Global Coordination

- [P0] Freeze **scope** and **chapter list** (incl. Evaluation & Calibration; Uncertainty & Risk; Simulation-Based Strategy Testing; SoS Governance).
- [P0] Create **project calendar** with weekly deliverables (chapters, models, figures, ablations).
- [P0] Establish **reproducibility contract**: seed control, data snapshots, environment pins, CPU/GPU parity notes.
- [P0] Baseline **hardware profiles**: (A) MacBook Air M4 (MPS), (B) dual RTX 5090 workstation (CUDA). Document expected batch sizes / epoch times.

Data Foundations (2015–2025 core; optional 1999–2014; selective priors pre-1999)

Acquisition & Storage

- [P0] Ingest **nflfastR** play-by-play 2015–2025; option flags for 1999–2014.
- [P1] Odds snapshots via **TheOddsAPI** every 10–15 min; persist to `odds_snapshots` (book, timestamp, market, price, rule hints).
- [P1] Weather joins (Open-Meteo/NOAA), stadium roof/surface map, geocoded stadium coords.
- [P1] Schedule context: rest days, travel distance (Haversine), time zones crossed, primetime flags.
- [P1] Injury integration: QB-out binary, team AGL index, cumulative starters out; weekly status (out/doubtful/questionable) encoder.
- [P2] Referee crew assignments; pace/penalty tendencies.

Feature Engineering (team-week / game-week grain)

- [P0] EPA/play (team & splits), Success Rate; opponent-adjusted via ridge; exponential decay (weekly half-life 0.6).
- [P0] PROE (pass rate over expected), neutral pace (sec/play), red-zone finishing (regressed).
- [P1] Trench proxies: pressure allowed/created, quick-pressure%, adjusted line yards proxy, stuff rate.
- [P1] Role stability: target share, aDOT, YPRR (derive routes if available), WR/TE room deltas on injury.
- [P1] Turnover luck: fumble recovery %, dropped INT proxy; mean-reversion flag.
- [P1] Discrete-margin model: fit key-number masses $P(M = n)$; expose as features (3, 6, 7, 10, $\dots$).
- [P2] Market microstructure features: hold, cross-book CBV, line-move velocity (dLine/dt), implied vs model deltas.

Data Quality & Testing

- [P0] Schema contracts; NOT NULLs; FK constraints; de-dupe policies for odds.
- [P0] Validation suite (basic): row counts per week, join rates, missingness dashboards.
- [P1] Statistical validation (Great-Expectations-style): value ranges, distribution drift monitors (weekly).

- [P1] Era handling: weighting schedule, `era` feature; strike years/OT rule changes guards.

Baseline Models (Classical)

Implementations

- [P0] **GLM**: spread → win prob (logit), home-field fixed effect; injury/weather interactions.
- [P0] **Stern (1991)** normal mapping sanity checks; calibrate σ seasonally.
- [P1] **State-space ratings** (Glickman–Stern): weekly θ for team strength via Kalman/Stan; posteriors.
- [P1] **Bivariate Poisson / Skellam** (Dixon–Coles; Karlis–Ntzoufras): score distribution, low-score dependence tweak; dynamic intensities (Koopman et al.).
- [P1] **In-play RF** (Lock–Nettleton) scaffolding for live WP (optional, keep modular).

Calibration & Outputs

- [P0] Brier & LogLoss vs. holdout; reliability diagrams; PIT for score distro.
- [P0] Vegas comparison: error vs closing spread; ATS/ML hit rates; CLV differentials.

RL Capstone

Agent Design

- [P0] DQN baseline: state (priors, features, market), actions (bet/no-bet or discrete stake buckets), reward (PnL; CLV-shaped variant).
- [P1] PPO actor-critic for richer actions (alt-lines/teasers/staking); entropy reg; clipping.
- [P1] Offline RL dataset (historic games as trajectories); behavior policy notes.

Training & Scaling

- [P0] Mac MPS config; CUDA config; batch/episode knobs for scale-up/down.
- [P1] Experience replay buffers; target networks (DQN); advantage normalization (PPO).
- [P1] Evaluation protocols: fixed-season rolling windows; no leakage; ATS/ROI metrics.

Ensembles & Comparative Backtesting

- [P0] Unified backtest harness: run GLM / Poisson / State-space / RL; collect metrics (Brier, LogLoss, ROI, Kelly growth, Sharpe).
- [P1] Simple ensembles (avg / logistic stack); Bayesian model averaging (optional).
- [P1] Ablations: remove feature families (injury/weather/trenches) to quantify marginal lift.

Uncertainty & Risk

- [P0] Posterior distributions (state-space, Poisson); bootstrap ensembles for GLM.
- [P0] Fractional Kelly module; bankroll simulator; risk-of-ruin & max drawdown analytics.
- [P1] Uncertainty-aware policy: downweight bets under wide posterior intervals.

Simulation-Based Strategy Testing

- [P0] Monte Carlo engine: simulate margins via fitted discrete distro; correlate with totals.
- [P0] Price teasers/alt-spreads via integer-crossing sums; EV curves vs book pricing.
- [P1] Middle detection thresholds; multi-book arbitrage scan (if legal venue assumed).

Narrative & Explainability

- [P1] SHAP for GLM/trees; factor attributions for game-level predictions.
- [P1] Rule miner: situational tags (short rest + cross-timezone + TNF).
- [P1] Insight generator: plain-language rationales (margin notes / appendix snippets).

Evaluation & Calibration (Dissertation Chapter)

- [P0] Define metrics: Brier, LogLoss, AUC, Accuracy, ROI, Kelly growth, Sharpe.
- [P0] Reliability diagrams; PIT histograms; CLV sparklines (margin figures).
- [P1] Vegas baseline tables; head-to-head model comparisons.

System-of-Systems Governance

- [P0] Experiment tracking (MLflow or Postgres schema: runs, params, metrics, artifacts).
- [P0] Model registry & promotion policy; semantic versioning; rollback plan.
- [P1] Pipeline DAG diagram; data lineage; environment manifests (Docker + native).

Writing & Figures

- [P0] Tufte-style layout: decide margin-note density; figure sizing guidelines; sparkline examples.
- [P0] "How we chose the timeframe" section with era weighting rationale.
- [P0] Literature integration chapter (top-10 models) + benchmark scripts references.
- [P1] Appendix: full visual gallery (key-number histos, teaser EV heatmaps, calibration plots).

Bibliography & Citations

- [P0] Maintain single `references.bib`; keep keys stable; add DOIs/URLs where missing.
- [P0] Audit all `\cite{}` have corresponding entries; compile warnings = 0.

Quality Gates (per milestone)

- Repro pass: deterministic runs (seeded), environment pinned, same metrics across machines.
- Validity pass: calibration in tolerance; Vegas comparison documented.
- Docs pass: figures captioned; equations referenced; todos burned down or deferred.