Referee Report — Round 6

Tennis Match Simulator

Referee 2

2026-02-10

Data Leakage Fixed; Elo Underperformance Correctly Diagnosed

Verdict: Accept with Minor Revisions

- ✓ **Data leakage eliminated** — date alignment module correctly resolves `tourney_date` mismatch (89.6% actual dates, 10.4% inferred)
- ✓ **Corrected results are trustworthy** — Elo 60.8%, MC 56.0%, gap 4.8pp (not 9.9pp)
- ✓ **Root cause identified** — career trajectory lag explains systematic Elo errors
- △ **Calibration and K-factor tests still pending** (flagged Rounds 3–5)
- △ **Proposed fixes are untested** — five remedies listed, zero evaluated

Leakage Fix: Date Alignment Module Works Correctly

How it works: Join ATP matches with betting data on player names + tournament to inherit actual dates. Infer dates from round for unmatched events.

Strategy	N	%
Exact name match	1,222	71.6%
Name variants	307	18.0%
Inferred from round	177	10.4%
Total	**1,706**	**100%**

Defensive measures:

- ✓ Cache-first loading
- ✓ Runtime fallback
- ✓ Warning if no alignment
- ✓ Validation function
- ✓ 8 unit tests

Residual risk: Inferred dates (10.4%) are for United Cup, Davis Cup — no impact on predictions.

Leakage Inflated Elo by 7.8pp, MC by 2.7pp — As Predicted

Model	Old	Clean	Drop
Elo	68.6%	**60.8%**	–7.8pp
MC	58.7%	56.0%	–2.7pp
Gap	+9.9pp	+4.8pp	

Metric	Elo	MC
Accuracy	60.8%	56.0%
Brier	0.2331	0.2451
Log Loss	0.6585	0.6846

Key takeaways:

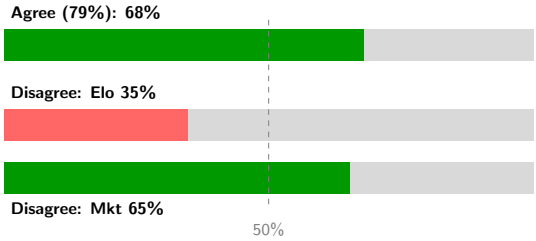
1. Elo still outperforms MC, but gap is 4.8pp not 9.9pp
2. Leakage affected Elo 3× more than MC (as referee predicted in Round 5)
3. **Neither model beats the market** (∼67% favorite accuracy)
4. ROI is negative at all thresholds

Why Elo was more affected:
K=32 per-match updates amplify leaked future results; MC averages stats over years, diluting leakage.

Agreement: 68% Accuracy; Disagreement: Market Wins

Scenario	N	Elo	Market
Agree	1,182	68.1%	68.1%
Disagree	317	34.7%	65.3%

Implication:
Elo provides zero incremental information. When it agrees, accuracy equals the market baseline. When it disagrees, it's an anti-signal.



Against Strong Favorites, Elo Is Wrong 84% of the Time

Favorite Odds	N	Elo Acc.	Market Acc.	Avg Gap
Heavy (<1.20)	6	16.7%	83.3%	50pp
Strong (1.20–1.40)	38	15.8%	84.2%	33pp
Moderate (1.40–1.60)	80	32.5%	67.5%	—
Slight (1.60+)	193	39.9%	60.1%	—

Worst case: Shelton vs Nishikori (French Open)
Market: 85% Shelton Elo: 23% Shelton Gap: **62pp** Result: Shelton won

Only **7 of 44** Elo contrarian picks against strong favorites were correct (16%).

Root Cause: Elo Cannot Track Rapid Career Trajectory Changes

Underrated (Rising):

Player	2022	2024	Shift
Etcheverry	18%	52%	+34pp
Struff	24%	59%	+35pp
Shelton	50%	62%	+12pp

These players improved sharply.
Elo still reflects their weaker past.
The market sees their current level.

Overrated (Declining):

Player	2023	2024	Shift
Mannarino	64%	33%	−31pp
Cachin	42%	12%	−30pp
Nishikori	67%	44%	−23pp

These players declined sharply.
Elo still reflects their stronger past.
The market prices in decline instantly.

Mechanism: Accumulated rating capital from hundreds of prior matches creates inertia that K=32 updates cannot overcome quickly enough.

Three Outstanding Items from Previous Rounds

Item	Rounds	Status	Impact
Calibration analysis	5, 6	Not done	Distinguishes directional vs probabilistic errors
K-factor sensitivity	3, 4, 5, 6	Not done	Directly tests trajectory lag fix
Statistical tests	5, 6	Not done	Standard practice, model comparison

K-factor sensitivity is the most actionable. The trajectory analysis shows Elo has too much inertia. Higher K directly addresses this by weighting recent matches more. Testing $K \in \{16, 24, 32, 48, 64\}$ on clean data requires a single backtest session with high diagnostic value.

Five Proposed Fixes Listed, Zero Tested

#	Fix	Mechanism	Effort	Status
1	Higher K-factor	More weight on recent results	Low	Untested
2	Time-decay	Older matches count less	Low	Untested
3	Form adjustment	Recent win rate multiplier	Medium	Untested
4	Refuse strong disagreements	Skip bets vs <1.40 favorites	Trivial	Untested
5	Hybrid model	Market prior + Elo adjustment	High	Untested

Recommendation: Test fixes 1 and 4 first.

- **Fix 1 (K-factor)**: Addresses root cause directly, one parameter change
- **Fix 4 (refuse disagreements)**: Trivial filter, eliminates 84%-loss-rate bets, immediate ROI improvement

Verdict: Accept with Minor Revisions

Resolved:

- ✓ Data leakage via `tourney_date` mismatch — properly fixed
- ✓ Corrected results reported with leakage impact quantified
- ✓ Root cause of Elo underperformance correctly identified
- ✓ Honest conclusion: neither model beats the market

Minor revisions requested:

1. Run calibration analysis using existing `calibration_summary()` function
2. Test K-factor sensitivity: $K \in \{16, 24, 32, 48, 64\}$
3. Add binomial test on disagreement accuracy, McNemar test for model comparison
4. Test "refuse strong disagreements" filter (skip bets vs <1.40 favorites)
5. Update `model_analysis.md` to distinguish pre- vs post-leakage numbers