Aalto-yliopisto
Perustieteiden
korkeakoulu

Probabilistic Modelling of Chess Outcome Distributions and Estimation of the Fair Frontier

Results Presentation

Ellis Saavalainen

30.4.2026

Ohjaaja: *Ahti Salo*

Valvoja: *Ahti Salo*

Aalto-yliopisto
Perustieteiden
korkeakoulu

Recap and Outline

Research questions revisited

RQ1. How do rating, engine evaluation, clock, and move number impact outcome probabilities?

RQ2. Which combinations of these parameters define the fair frontier, where White and Black win probabilities are equal?

Outline of this presentation

- Model recap and key symbols (θ, a_m, d_m)
- Training and evaluation setup (objective, optimiser, learning-rate α)
- Predictive performance and calibration on the test split
- Robustness to learning rate
- Fair-frontier geometry across feature pairs (RQ2)
- Time-trouble heatmaps and absolute-rating effect
- Conclusions, limitations, and future work

Model Recap

Trainable parameters θ. Collection of all learnable weights and biases of the network. The model is fully specified by $p_\theta(\cdot \mid x, m)$.

Structured WDL logits. Two interpretable scalar heads parameterise the outcome distribution:

- $a_m(x)$ — advantage score for White at input x with mask m.
- $d_m(x)$ — draw-tendency score at input x with mask m.

Logits = $(a_m, d_m, -a_m)$ for (White, Draw, Black). Antisymmetry guarantees P(White) = P(Black) iff $a_m(x) = 0$, irrespective of $d_m(x)$.

Partial monotonicity. Selected features (e.g. elo_diff, eval_cp_white_pov) enter through monotone subnetworks so a(x) is non-decreasing in those features.

Gate constraint. Clock effects enter as side-specific multiplicative gates in logit space, $\ell\_W = a + \log g\_W$, $\ell\_B = -a + \log g\_B$, with a hard zero at non-positive clocks and a neutral gate (g = 1) when the clock is masked.

Mask-aware learning. Stochastic masks during training (40% full mask, 60% over the 62 other non-empty masks) yield a single model that handles arbitrary observation subsets.

Training and Evaluation Setup

Objective. Multiclass cross-entropy $\mathcal{L}\_\mathrm{CE}(\theta)$ with masked-input augmentation [Goodfellow et al., 2016].

Optimiser. Adam with cosine-annealed learning rate. Training is intentionally unregularised (no weight decay, no dropout, no early stopping) to keep the learned probability surfaces flexible enough for geometric analysis [Kingma & Ba, 2015].

Learning rate (α). Step size used by Adam at each parameter update; controls how aggressively θ follows the gradient. Three values were swept as a minimal robustness check: $\alpha \in \{ 3\times10^{-4}, 10^{-3}, 3\times10^{-3} \}$.

Data split. By unique game_id, approximately 80 / 10 / 10 % for train / validation / test (≈3.999M / 0.500M / 0.499M rows).

Evaluation metrics. Headline metrics on the test split under full observability (m = 1): multiclass log loss, multiclass Brier score, and accuracy [Brier, 1950; Gneiting & Raftery, 2007]. Diagnostics include classwise reliability diagrams, feature-binned predicted-vs-empirical curves, and 2D fair-frontier and heatmap plots.

Predictive Performance (RQ1)

Selected model ($\alpha = 3\times10^{-3}$)

Metric	Test value
Log loss ↓	0.7199
Brier score ↓	0.4337
Accuracy ↑	0.6608

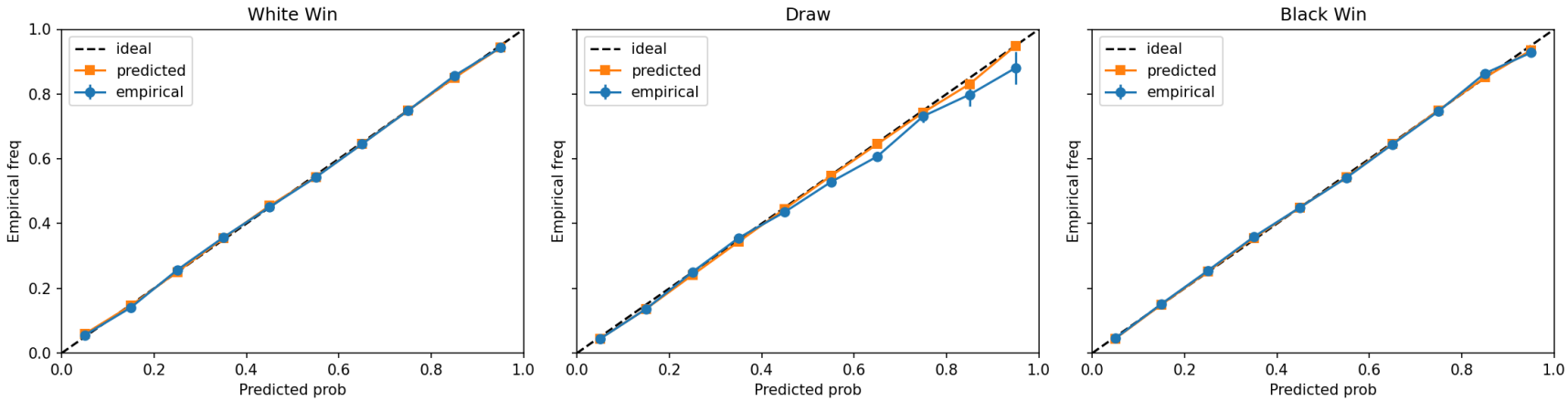
Takeaways

Probabilistic accuracy. Log loss 0.72 and Brier 0.43 indicate well-calibrated three-class probabilities; smaller is better for both proper scoring rules.

Hard accuracy. 66.1% argmax accuracy across {White win, Draw, Black win}, well above the ≈34% majority baseline.

Why scoring rules matter. The goal is probability prediction, not only classification. Log loss and Brier reward sharpness *and* calibration jointly.

Calibration: Classwise Reliability



Reliability diagrams (selected $\alpha = 3\times10^{-3}$). Per-class predicted vs. empirical frequencies. Points on the dashed diagonal indicate perfect calibration.

- **White Win and Black Win** are essentially on the diagonal across the full [0,1] range of predicted probabilities.
- **Draw** is well calibrated up to ≈0.7; for very high predicted draw probabilities the model is mildly over-confident, with empirical draw rate slightly below the predicted line.

Together with the Brier and log-loss values, this confirms the predicted distributions are usable as probabilities, not only for argmax decisions.

Robustness to Learning Rate α

Test-set headline metrics across the learning-rate sweep

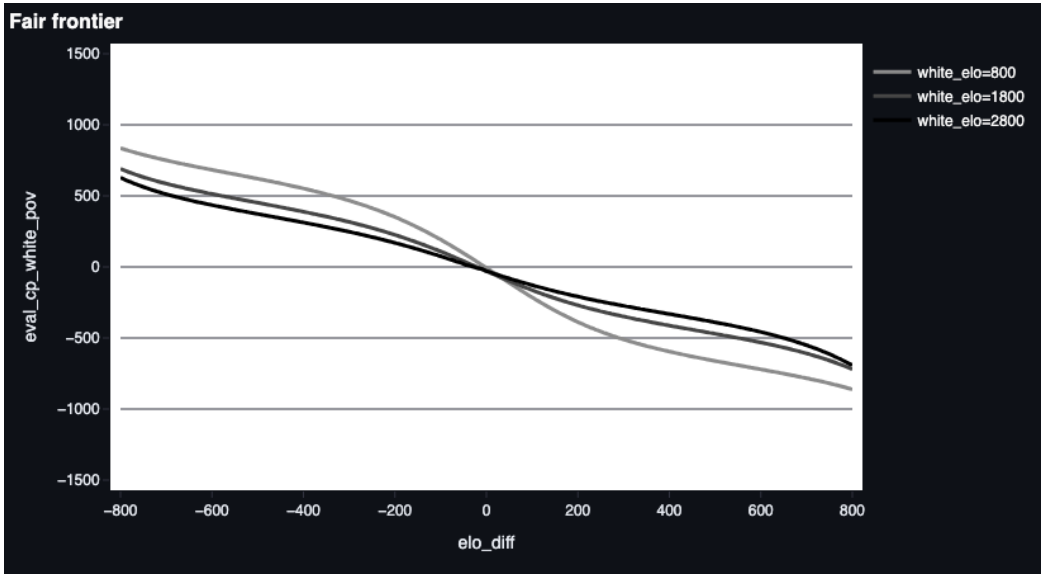
Learning rate α	Log loss ↓	Brier ↓	Accuracy ↑
3×10^{-3} (selected)	0.7199	0.4337	0.6608
10^{-3}	0.7255	0.4367	0.6584
3×10^{-4}	0.7335	0.4416	0.6560

Findings. Ordering is consistent across all three headline metrics: $3\times10^{-3} > 10^{-3} > 3\times10^{-4}$.

Magnitudes are modest: relative to 10^{-3}, the selected model improved log loss by 0.0057 (≈0.78% relative), Brier by 0.0031, and accuracy by 0.25 pp. Relative to 3×10^{-4}: −0.0136 / −0.0079 / +0.49 pp.

Calibration & feature-binned diagnostics tell the same qualitative story: the highest α yields the best agreement with empirical frequencies, with the clearest gain in draw-class reliability.

Fair Frontier — elo_diff × eval (RQ2)



All three curves cross near the origin (elo_diff = 0, eval = 0): when ratings and position are equal, the model assigns equal win probabilities to both colours regardless of skill level.

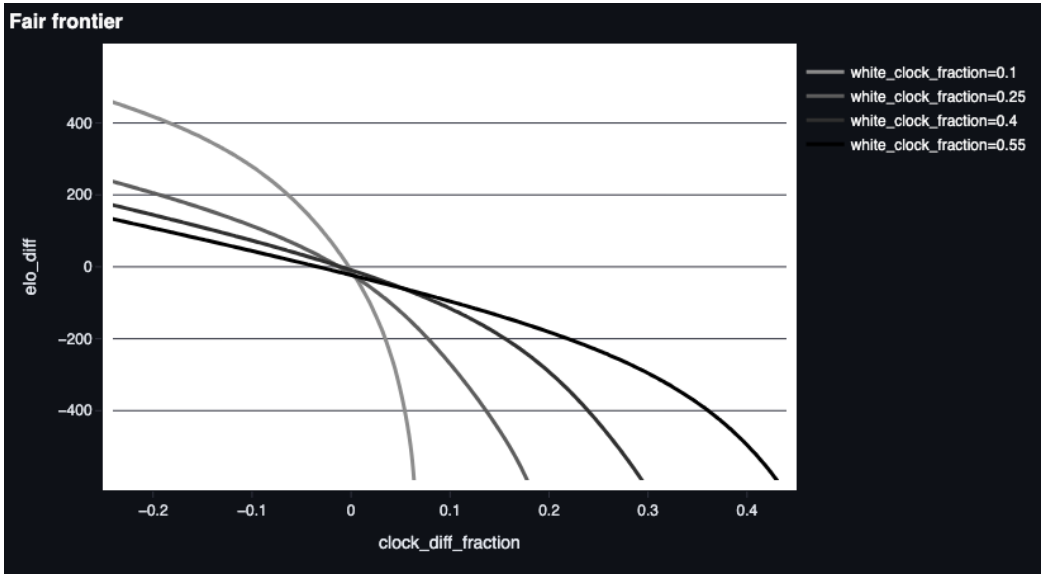
What is plotted

Frontier curves $\mathcal{F}$ in the elo_diff × eval_cp_white_pov plane for three absolute white_elo levels (800, 1800, 2800).

Findings

- Frontier is approximately linear with mild S-shaped curvature.
- Local exchange rate ≈ 8 Elo points per 7 centipawns along the boundary.
- Lower white_elo ⇒ steeper curves: stronger players convert positional advantage more reliably, so less Elo compensation is needed.

Fair Frontier — clock_diff × elo_diff



Interpretation. *When White is near flagging, time advantage dominates the white-vs-black balance more strongly: small clock differences can move the same Elo gap from balanced to decisive.*

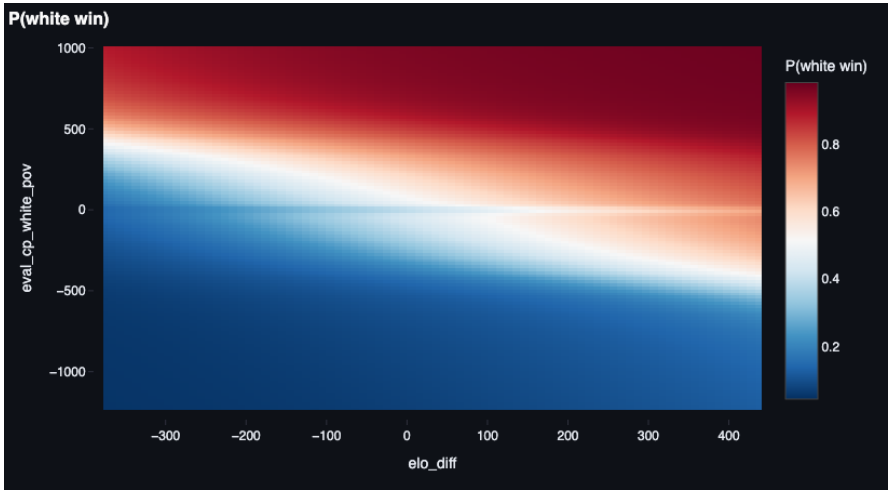
What is plotted

Frontier curves $\mathcal{F}$ in the clock_diff_fraction × elo_diff plane for four white_clock_fraction levels (0.10, 0.25, 0.40, 0.55).

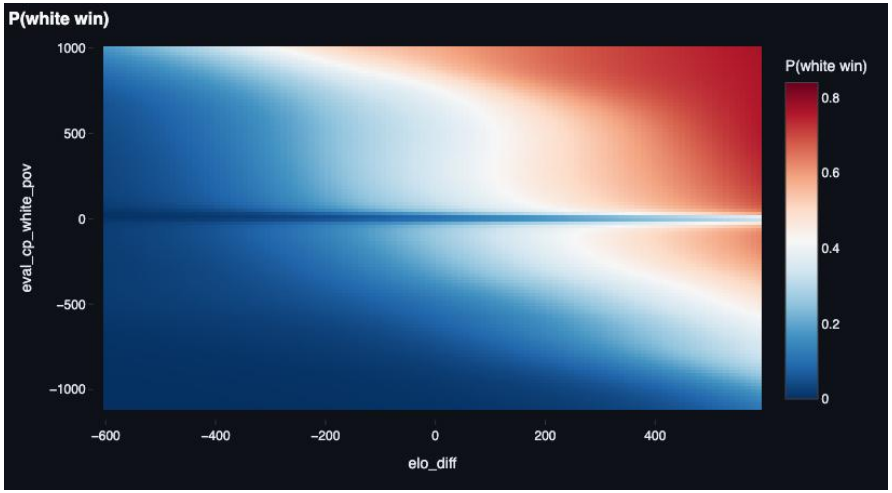
Findings

- Curves intersect close to the origin and rotate across white_clock_fraction.
- Frontier becomes visibly steeper as white_clock_fraction ↓ (less time on White's clock).
- Strong steepening near the boundary where clock_diff_fraction approaches the available white clock — consistent with a clock-gate mechanism.

Time Trouble Reduces Determinism



Baseline regime (no time pressure)

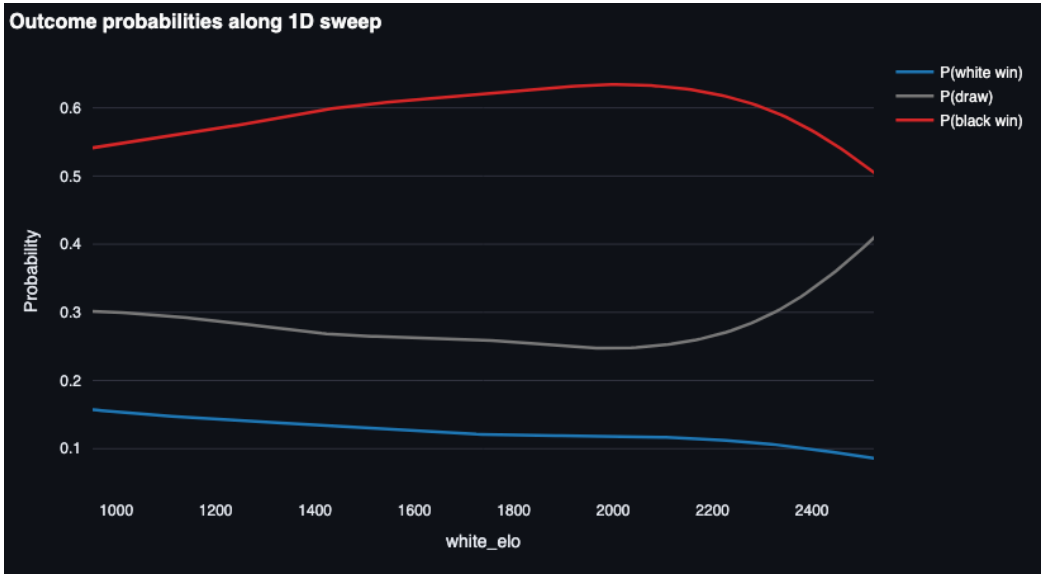


Time-trouble regime

Both heatmaps show $p\theta(\text{White win} \mid x, m)$ in the elo_diff × eval_cp_white_pov plane. Global geometry is similar, but in time trouble the boundary near eval = 0 is darker (lower decisiveness), high-probability white-win regions are less crisply defined, and overall colour intensity is weaker.

Practical interpretation. *Time pressure makes outcomes less deterministic and increases uncertainty in the conversion of positional and rating edges.*

Absolute Rating Matters Beyond ΔElo



Setup

Sweep over absolute white_elo with elo_diff fixed at 0 (equal-ratings stressed state: black slightly better, black ≈30 s, white ≈1 min).

Findings

- P(white win) decreases monotonically as absolute rating rises.
- P(black win) rises until ≈2000 Elo, then drops rapidly.
- P(draw) is flat-to-slightly-decreasing up to ≈2000 Elo and then climbs sharply at higher ratings.

Implication. *There is a non-trivial absolute-skill effect that is not reducible to the Elo difference alone: stronger players convert and defend differently in stressed positions, increasing draw propensity rather than the lower-rated player's win rate.*

Conclusions

RQ1 — Drivers of outcome probability. Player strength, position quality, and clock state all contribute substantially to WDL outcomes, and they interact non-additively. Time pressure flattens the probability surface; positional and rating edges become less determinative.

RQ2 — Geometry of the fair frontier. $\mathcal{F}$ is a structured, plausible object: nearly linear with mild curvature in elo_diff × eval, and steepening sharply as the absolute clock shrinks. Clock effects amplify near flagging.

Predictive headline. Selected model on the test split: log loss 0.7199, Brier 0.4337, accuracy 0.6608. Calibration is essentially diagonal for White-Win and Black-Win; mild over-confidence remains in the high draw-probability tail.

Main contribution. An interpretable, structured neural model that recovers a domain-meaningful trade-off between skill, position, and time — not just a predictive score.

Limitations and Future Work

Limitations

- One platform, one time control (Lichess 5+0 blitz); sampled positions, not full game trajectories.
- Observational data → associational, not causal interpretation. Clock state is entangled with strength, position, and dynamics.
- Low-dimensional feature set (six tabular variables): material, king safety, tactical volatility, opening family etc. enter only via the engine eval.
- Stockfish eval at fixed 20 ms with ±1000 cp clipping introduces noise.
- Unregularised training favoured geometric richness over conservatism → some local curvature may reflect flexibility rather than data support.
- Small α sweep; no benchmark against logistic regression, GBMs, GAMs, or other deep tabular baselines.

Future work

- Broader empirical validation across months, rating bands, and time controls.
- Richer features: material imbalance, phase indicators, opening info, tactical volatility, deeper engine signals.
- Engine WDL outputs as direct features; uncertainty-aware engine evaluation.
- Methodological benchmarking: multinomial LR, GAMs, GBMs, deep tabular models.
- Calibration & uncertainty: temperature scaling, isotonic, conformal prediction, bootstrap bands for frontier plots.
- Sequential modelling of evaluation trajectories and clock dynamics, particularly relevant in blitz.
- Frontier as a quantitative object: estimate slope, curvature, and uncertainty across game phase and rating.

Questions?

Thank you.

Probabilistic Modelling of Chess Outcome Distributions and Estimation of the Fair Frontier

Ellis Saavalainen — 30.4.2026