How the Win-Probability Model Works

From Elo basics to the 8-bucket softmax model. The methodology is region-agnostic; the calibration dataset used here is 13,739 EUW solo/duo games across 120 players spanning the full Emerald-to-Challenger ladder.

1. The problem

Given a 10-player ranked lobby (5 allies, 5 enemies, each with a known op.gg peak rank), output one number: the probability the ally team wins. No champion picks, no role context, no team comp logic — just rank vs rank.

This document walks the whole journey: how the first version was built, what was wrong with it, and how it was calibrated against real match outcomes to land on the current numbers.

2. First swing — the intuition-based model

The model I'll call the **"OG" version** was built by hand. Eight discrete buckets (six for Master-and-above split by LP, one for Diamond, one catch-all for Emerald-and-below), each assigned an MMR value by gut feel:

Bucket	Peak rank range	OG MMR
HC	Master 2500+ LP	+1275
LC	Master 2000–2500 LP	+1000
GM	Master 1500–2000 LP	+750
HM	Master 1000–1500 LP	+500
MM	Master 500–1000 LP	+250
LM	Master 0–500 LP	0
DM	Diamond (any division)	–150
EM	Emerald or below	–350

Plus two model-wide knobs: **T = 400** (the softmax "temperature" — more on this later) and **S = 400** (the Elo scale — same). The values were plausible-looking guesses. None of them were tested against real outcomes.

The OG model is going to be the reference point throughout the rest of this doc. Every time we say "calibrated" or "new" version, the implicit comparison is to this hand-tuned starting point.

3. The math under the hood — Elo (chess)

The model is an **Elo system** — the same family of skill ratings developed by Arpad Elo around 1960 for ranking chess players (adopted by the US Chess Federation that year, by FIDE in 1970).

Trivia: Elo was a real person — a Hungarian-American physics professor. "ELO" is **not** an acronym, despite a persistent backronym ("Expected Level of Opposition"). It's just his surname. The same rating family is used in everything from chess to League to competitive Scrabble to (until 2019) Tinder's matchmaking algorithm.

In the chess (1v1) case, each player has a numeric rating R. Higher is better. If player A (rating R_A) plays player B (rating R_B), the predicted probability A wins is:

$$P(A \text{ wins}) = 1 / (1 + 10^{(R_B - R_A) / S})$$

S is the **scale** parameter — chess uses S = 400, with a specific interpretation: a 400-point gap means the higher-rated player wins ~91% of the time, an 800-point gap means ~99%. **Each S-point gap multiplies the favourite's odds by 10×**.

The base-10 isn't fundamental — you could rewrite the same model in natural log with a rescaled S. Elo picked base 10 for the "decade of odds per S-point" interpretability. It's a convention, not a deep mathematical truth.

For perspective: Magnus Carlsen's peak FIDE Elo is 2882. Bobby Fischer's was 2785 — a 97-point gap. Plugging into the formula: prime Magnus would beat prime Fischer ~64% of the time. The gap between the two greatest chess players ever is, by Elo's own scale, a slight favourite — not a layup.

After each game, ratings update toward the actual result. In chess:

$$R_{A,new} = R_{A,old} + K \cdot (actual - expected)$$

where K (typically 10–40) controls how fast ratings move. League's actual MMR system is proprietary and not pure Elo (it adds decay, smurf-queue heuristics, and uncertainty bands à la Microsoft's TrueSkill). But the **core mechanic — a numeric skill rating, a logistic prediction curve, update steps** — is Elo. Op.gg's "Top Tier" (peak rank in the current season) is the closest visible proxy, so that's what the model reads as input.

4. Adapting Elo to a 5v5 team game

The chess formula expects one rating per side. League has 5. How do you turn five player MMRs into one team rating?

Two obvious extremes:

- **Mean:** R_{team} = average of the 5 player MMRs. Assumes every role contributes equally and the weakest links count as much as the carries.
- **Max:** R_{team} = highest of the 5. Assumes the strongest player carries everything and weaker teammates barely matter.

Both are wrong in opposite directions. The truth lives between them, and the model needs a knob to dial it.

5. The softmax-Elo trick (Dehpanah-style aggregation)

The clean way to interpolate between mean and max is the **log-sum-exp (softmax)** aggregation. This is the approach Dehpanah et al. and others use for multiplayer Elo extensions in MOBA / team-shooter research:

$$R_{team} = T \cdot \log(\Sigma \exp(R_i / T))$$

where R_i are the 5 player MMRs and T is a **temperature** parameter:

- **T → 0:** the $\exp(\cdot/T)$ of the largest R_i dominates the sum, so $R_{team} \approx \max(R_i)$. The strongest player carries everything.
- **T → ∞:** all exp terms collapse to ~1, and R_{team} approaches the mean. Every player contributes equally.

Why softmax is the natural choice: it's the smooth, differentiable interpolation between min/mean/max that comes out of statistical mechanics (it's the Boltzmann sum). For MOBA games specifically, the Dehpanah-style research argues that skill in team games is asymmetrically transferable — a stronger player elevates teammates more than weaker players drag them down. The softmax with a moderate T captures this: the strongest player's MMR gets the heaviest exponential weight, but everyone still contributes.

The OG model picked T = 400 — a guess. We'll see how the data feels about that.

6. Why discrete buckets instead of continuous LP

In principle each player could be assigned a continuous MMR from their exact peak LP. Two reasons we stuck with 8 discrete buckets:

- **Display.** A bucketed model fits cleanly into a counts grid: "3 LMs, 1 DM, 1 HM on each side". Easy to read at a glance, easy to input by hand in the web tool.
- **Statistical power.** 8 discrete bucket parameters is way fewer than fitting a continuous curve, which helps with our ~13k-game dataset. A flexible 50-parameter curve would overfit.

Gauge choice: the model is anchored so **Master 0 LP = MMR 0** exactly. This is a free choice — adding a constant to all 8 MMRs doesn't change any prediction (predictions only see *differences*) — so we have to fix one. Master 0 LP being the zero point is a natural pick.

7. Time to calibrate

The OG model's parameters (8 bucket MMRs + T + S) were 10 numbers, all guesses. The plan: collect a large set of (lobby ranks, actual outcome) pairs and let the data tell us what they should be.

Data collection

120 hand-picked EUW players, 15 per bucket × 8 buckets, with at least 50 ranked solo/duo games since the 2026-04-29 season reset. For each player we pulled their first up-to-200 games chronologically, then scraped every unique lobby participant's op.gg "Top Tier" peak rank — about 44,000 unique players, scraped via headless browser at ~30/min over many hours.

Final dataset: **13,739 games**, each with 10 known peak ranks and a known winner. Bucketed by peak LP into the 8 buckets.

Fitting method

Maximum likelihood. For each game, the model predicts P(ally wins) given the lobby's bucket counts and the current parameters. Each game contributes log P(observed outcome) to the total log-likelihood. We minimise the negative log-likelihood (i.e., maximise the probability the model assigns to the actual outcomes) using L-BFGS-B (a standard quasi-Newton optimiser from scipy).

Overfitting controls

- **Player-level holdout:** 30 of the 120 players (25%) were set aside from the start, never used for fitting. They're the sealed test set. Player-level (not game-level) holdout matters because games from the same player are correlated — splitting at the game level would leak per-player skill into the holdout.
- **5-fold cross-validation** on the 90 training players to pick the winning model level by held-out Brier score (mean squared error of predicted vs observed win/loss).
- **AIC/BIC** for parameter-count penalties as a sanity check.

The model ladder

We didn't just fit "the model" — we fit five nested versions, each adding flexibility on top of the previous one, and picked the best by cross-validation. This guards against overfitting: a more complex model only wins if it actually helps held-out predictions.

Level	What's free	# params
L0	nothing (OG baseline, all values fixed)	0
L1	8 bucket MMRs (anchored to Master 0 LP = 0)	7
L2	L1 + T + S (global values)	9
L3	L2 + rank-dependent T: T(R) = T■ + a·R	11
L4	L3 + rank-dependent S	13

8. What came out

Brier scores (lower = better calibration)

Model	5-fold CV Brier	Sealed holdout Brier	Holdout Δ vs OG
L0 (OG)	0.2425	0.2481	—
L1	0.2265	0.2309	–0.0172
L2	0.2260	0.2303	–0.0178
L3 (winner)	0.2260	0.2303	–0.0178
L4	0.2260	0.2303	–0.0178

L3 wins on the strict CV tiebreak, but L2, L3 and L4 are statistically identical — rank-dependence of T and S didn't help. **L2 is the practically simplest equivalent** and what the website ships: same 8 bucket MMRs, one global T, one global S.

Final calibrated parameters

Parameter	Calibrated	OG (intuition)	Δ
HC MMR	+1125	+1275	–150
LC MMR	+930	+1000	–70
GM MMR	+793	+750	+43
HM MMR	+654	+500	+155
MM MMR	+398	+250	+148
LM MMR	+49	0	+49
DM MMR	–34	–150	+116
EM MMR	**–312**	**–350**	**+38**
T (temperature)	**1317**	**400**	**+917 (≈3.3×)**
S (Elo scale)	510	400	+110

Side note — the HC vindication: earlier exploratory work on a much smaller apex-only dataset suggested HC ≈ 1109 (via a joint MLE on apex games). The new fit on the full 120-player dataset lands at **HC = +1125** — within 16 MMR of that prior. Two independent datasets, same answer. Good sign.

9. Worked example — putting the calibrated model to work

Five Low Masters (each with MMR +49) on the ally team. Four Low Masters plus one High Challenger (MMR +1125) on the enemy team. Using the calibrated T = 1317 and S = 510:

$$\text{Ally rating} = 1317 \cdot \log(5 \cdot \exp(49/1317)) = 1317 \cdot \log(5.190) = 2169$$
$$\text{Enemy rating} = 1317 \cdot \log(4 \cdot \exp(49/1317) + \exp(1125/1317)) = 1317 \cdot \log(6.501) = 2466$$

Difference: enemy team is +297 ahead. Plug into the Elo formula:

$$P(\text{ally wins}) = 1 / (1 + 10^{297 / 510}) = 1 / (1 + 3.83) = 0.21$$

So one HC on the enemy side drops your win rate from 50% (mirrored lobbies) to 21%. **The same scenario under the OG model** (T = 400, S = 400, HC = +1275, LM = 0) would have given you ~**2%**

win rate. The OG T = 400 made carries far too dominant — a single Challenger looked like an automatic loss. The calibrated T = 1317 reflects what the data actually shows: solo carries matter, but less than the intuition-based model assumed.

10. What this tells us — human intuition

Diamond and Low Master are basically the same tier

DM = –34, LM = +49 — just **83 MMR apart**. Translated into the units you actually see on the client: the median Low Master in our data peaks around **Master 300 LP**, while the median Diamond peaks around **Diamond 2**. That's roughly **4 divisions of climbing distance** on the ladder.

By the model: in a hypothetical 1v1 between two players at their bucket medians, the LM wins ~59% of the time — barely above coin-flip. ($P = 1 / (1 + 10^{-83/510}) = 0.593$.) Four divisions of LP between Diamond 2 and Master 300 buys the higher one almost nothing in true skill. **The Master crest at that boundary is mostly cosmetic.**

The Emerald cliff

EM = –312, DM = –34 — a **278 MMR gap**, 3.3× bigger than the DM↔LM gap above. The median EM peaks around **Emerald 2** and the median Diamond around **Diamond 2** — about the **same 4-division LP distance** as DM↔LM, but more than 3× the skill jump.

Same LP distance, vastly different skill jump: P(higher wins 1v1) = 0.78 here, vs only 0.59 for DM↔LM. **The real cliff in solo Q lives at the Emerald/Diamond boundary, not the Diamond/Master one.**

And the EM bucket lumps everyone Emerald-and-below together — Platinum, Gold, Silver, Bronze, Iron, Unranked. Inside that bucket, the effective skill depends on the player mix; see the caveat in §11.

The compressed apex

HC dropped from the OG +1275 to **+1125**. The calibrated apex spread is **narrower** than originally assumed — High Challenger is only ~150 MMR above Low Challenger, not 275. Long-standing intuition about how spread out the apex really is was too wide.

What T = 1317 means in plain terms

T controls the carry-vs-team-play balance. In the softmax sum, each player's weight is exp(R/T). The intuitive question: how much does a strong player actually carry the team rating? Some concrete examples, all with 4 Low Master teammates (Master ~300 LP):

- **Mid Master carry** (Master ~700 LP, +349 MMR over teammates): weight ratio exp(349/1317) ≈ **1.30×**. The MM contributes 30% more to team rating than each LM. Barely carrying.
- **High Master carry** (Master ~1200 LP, +605 MMR over LMs): weight ratio **1.58×**. Noticeable but the LMs still dominate the sum.
- **High Challenger carry** (Master 2800+ LP, +1076 MMR over LMs): weight ratio **2.27×**. The HC counts as ~2.3 LMs. They carry meaningfully — but 4 LMs at 1× each still outweigh the one HC at 2.3×.
- **Truly dominating** a team of 4 LMs would require the carry to weight more than the sum of the rest — $\exp(\Delta/1317) > 4$, i.e., $\Delta > 1825$ MMR. That is **above HC** — no real player exists at that gap against LMs in a solo Q lobby.

Translation: in this model, no realistic carry in a Master+ lobby genuinely "1v9s". The strongest player tilts the sum, but 4-vs-1 weight-by-tier always matters more than any single player's ceiling.

The OG T = 400 set the same thresholds at **3.3× tighter**: a +349 MMR carry got exp(349/400) ≈ 2.4× weight (vs the new model's 1.3×). **That was way too max-heavy.** The data says solo carries are far rarer than the intuition believed. Average team strength matters more than your best player's ceiling.

What S = 510 means in plain terms

S is the Elo scale: every S-point MMR gap multiplies the favourite's odds by 10×. With S = 510, anchored to concrete bucket pairs:

- **LM (Master ~300 LP) vs DM (Diamond 2)**, gap 83 MMR (~4 divisions of LP): 59% favourite.
- **HM (Master ~1200 LP) vs LM (Master ~300 LP)**, gap 605 MMR (~9 divisions of climbing in Master): 92% favourite.
- **HC (Master 2800+ LP) vs LM (Master ~300 LP)**, gap 1076 MMR (~25 divisions of climbing): 99.2% favourite.
- **HC vs EM (Emerald 2)**, gap 1437 MMR (~34 divisions across the Em → Diamond → Master span): 99.8% favourite.

S = 400 (the OG value) would have made every one of those gaps feel sharper — HM vs LM goes from 92% under S = 510 to 96% under S = 400. The calibrated S = 510 says rank gaps in solo Q convert to win rate **less steeply** than classical chess Elo predicts. That tracks — solo Q has more randomness (4 randos per side) than chess does.

11. Caveats and known limitations

The EM bucket is not uniform

The EM bucket sweeps in anyone with a peak below Diamond — Emerald, Plat, Gold, Silver, Bronze, Iron, Unranked. The calibration data was ~83% Emerald inside the EM bucket. But in lobbies where the EM-bucket share is heavier on Plat / Gold / lower (e.g., smurf accounts queuing in low MMR), the bucket's effective skill is lower than the calibrated –312, and the model will over-credit the higher-ranked side for beating them.

For analysing such cases (e.g., a Master smurf grinding through Emerald MMR) we extrapolate sub-Diamond tier MMRs from the EM→DM slope of the fitted curve. That gives:

```
Emerald: −312 | Plat: −591 | Gold: −859 | Silver: −1126 | Bronze: −1395
| Iron: −1660
```

These aren't part of the calibrated 8-bucket model that ships on the website — they're an analysis tool for the cases where the EM bucket's real composition diverges sharply from the calibration mix.

S could be rank-dependent

Earlier work hinted S might grow with rank — fits at the Low-Master end suggested S ≈ 400, fits on Challenger data suggested S ≈ 1325. The L4 model tried fitting S(R) = S■ + a·R but the slope came out essentially zero, because we don't have enough Challenger games in the dataset to drive the rank-dependence. The shipping value S = 510 is reasonable globally but possibly underfits the very top. Open follow-up.

Bucketing loses information

A Master 5 LP player and a Master 499 LP player both bucket as LM and get the same MMR. They're not the same. A continuous-LP fit (using exact peak LP through a monotone spline) is the natural next step and would improve sharpness, especially near bucket boundaries.

Op.gg's "Top Tier" is the input, with all its quirks

We use op.gg's Top Tier (peak rank in the current season). This is lagging for players who just hit a new high mid-season and is missing entirely for renamed accounts (~1.4% of slots). For unknowns we use a **lobby-mean fallback**: assign the unknown player the bucket whose MMR is closest to the average of the 9 known peaks in their game.

12. TL;DR

The model is an Elo system like chess, extended from 1v1 to 5v5 by replacing each side's single rating with a softmax aggregate of the 5 players' MMRs. Two knobs: T (temperature, how much the strongest player carries) and S (Elo scale, how rating gaps convert to win probability).

Started from a hand-tuned baseline (the OG model), fit 8 bucket MMRs + T + S on 13,739 EUW games across 120 players spanning Emerald to Challenger, with 25% player-level holdout and 5-fold CV. Result: **7.2% improvement in holdout Brier vs the OG**, well-calibrated reliability across the

prediction curve.

Three takeaways the numbers force you to accept:

- **Diamond and Low Master are essentially the same skill tier.** Median Diamond peaks at **Diamond 2**, median LM peaks at **Master 300 LP** — ~4 divisions of climbing distance, only 83 MMR of skill difference. LM wins 59-41 in a hypothetical 1v1. The Master crest at that boundary is mostly visual.
- **The real skill cliff is between Emerald and Diamond.** Median Emerald peaks at **Emerald 2**, median Diamond at **Diamond 2** — same 4-division LP distance as DM↔LM, but **3.3× the skill jump**. The Diamond wins 78-22 in 1v1. Only meaningful tier boundary in the Emerald-to-Master range.
- **Solo carries matter less than the OG model assumed.** Under the calibrated T = 1317: a High Master playing with 4 Low Masters (*+9 divisions above teammates*) contributes only 1.6× the weight of each teammate in the team rating. A Challenger playing with 4 Low Masters (*+25 divisions above*) contributes 2.3×. No real-world carry in a Master+ lobby is enough to outweigh the average of their team. Solo carries exist, they're just rarer than the intuition believed.