Building Stratified Lichess Benchmark Curves for Elo+Chess

Elo+Chess Research Notes

May 31, 2026

Abstract

Elo+Chess benchmark reports are built from large stratified samples of Lichess.org games, with the benchmark sample used in this note containing 483,974 1+0 bullet games, 479,974 3+0 blitz games, 925,636 10+0 rapid games, and 660,540 longer rapid games. For each major game type, the pipeline expands every game into before-move and after-move board states, calculates hundreds of intermediate and player-game variables, aggregates those variables into coaching metrics, and then averages those metrics by Lichess Elo bucket. The sample is deliberately stratified rather than proportional to the raw player pool: the purpose is to estimate stable behavior curves at each rating level, including buckets that would otherwise be sparse, not to reproduce the natural rating distribution. The result is a family of empirical benchmark curves: for each metric, we can ask how often players at different Elo levels show a given habit or principle in real games. This note describes the method, the scale of the computation, and several opening-principle examples involving castling, minor-piece development, repeated opening piece moves, and early queen movement. The emphasis is methodological: Elo+Chess does not assume that chess principles are universally linear across all skill levels. Instead, benchmark curves are inspected by rating range, and the product focuses on the beginner-to-early-advanced region where the relationships are strongest and most interpretable.

1 Purpose

The goal of the benchmark system is to convert broad chess advice into measurable quantities. A statement such as "develop your minor pieces" is useful, but it becomes more actionable when we can say how many minor pieces a player usually develops by move 10, how that compares with peers in the same rating bucket, and how that habit changes across the rating spectrum.

Elo+Chess therefore builds a metric library from actual games. Each metric is computed for every player-game row, summarized by game type and Elo bucket, and used as a benchmark curve in the report. The report can then place a user on the same curve and show whether the user is above, below, or near the behavior observed among comparable players.

2 Source Data and Stratification

The benchmark source is the public Lichess.org game database. Elo+Chess uses large stratified samples by game type and rating bucket rather than a small convenience sample. The current production runtime artifacts cover four major game categories:

- 1+0 bullet,
- 3+0 blitz,

- 10+0 rapid,
- rapid games longer than 10 minutes.

The stratification goal is breadth across rating levels. A proportional random sample would be efficient for describing the overall player pool, but it would be inefficient for building benchmark curves. The most common rating buckets would dominate the sample, while low-support tail buckets would have too few games for stable metric averages. Elo+Chess therefore samples by average game Elo bucket. In the current clean-sample construction, raw games are first filtered to valid rated games with known players, known ratings, valid results, and known time controls. Games are then filtered to players with enough games in that speed and stable enough ratings to represent a coherent skill level: for each player, Elo+Chess computes the player's within-speed rating span ($\max(\text{Elo}) - \min(\text{Elo})$) across the source pool and keeps games only when both players have at least the configured minimum game count and no more than 300 Elo points of within-speed drift. The drift filter is not meant to remove normal improvement or variance; it reduces cases where a player's games would mix materially different skill levels, such as provisional accounts, returning players, or accounts that moved rapidly through several buckets during the sampled period. Finally, eligible games are ranked within each average-Elo bucket by a deterministic hash of the game id and seed, and a capped number of games is retained from each bucket.

This design makes the benchmark curves much more useful. Each Elo bucket contributes enough observations to estimate the average value of each metric, and rare high- or low-rating buckets are not drowned out by the natural shape of the active-player distribution. The benchmark curves used by Elo+Chess are then restricted to the range most relevant to the product, roughly beginner through early advanced play. In the Lichess-scale reports, that primary range is 600–1600. When a user views the report on the Chess.com scale, the corresponding rating buckets are mapped onto the same underlying Lichess benchmark buckets using the separate cross-platform rating mapping method described in the companion paper: Estimating Cross-Platform Chess Rating Mappings with Modal Regression.

3 Active-Player Rating Distributions

If we look at the distribution of Lichess.org player Elo ratings as of the end of March 2026, by major game type, we see the following counts and percentiles. For this distribution view, each player is counted once per exact game type, using that player's latest observed rating in the corresponding raw time-control rows, after requiring at least five games in that exact game type.[1]

Game type	Players	Median	75th pct.	80th pct.	% ≤1600	Exact 1500	1500 share
1+0 bullet	1,221,146	1,479	1,858	1,951	58.9%	4,756	0.4%
3+0 blitz	1,144,045	1,542	1,845	1,912	54.9%	2,315	0.2%
10+0 rapid	1,679,401	1,381	1,691	1,765	68.3%	3,472	0.2%
>10 rapid	668,255	1,400	1,686	1,753	68.2%	7,969	1.2%

Table 1: Latest observed active-player Lichess rating distribution by exact game type after requiring at least five games in that exact game type.

[1] We apply a five-game minimum because lightly active accounts disproportionately remain at the 1500 starting/provisional anchor, creating an artificial spike in all-player latest-rating distributions. Figures 1–4 show the distributions with and without this filter.

This distribution should not be confused with the stratified benchmark sample. The distribution describes the active Lichess player pool observed in the raw scan. The benchmark sample is then deliberately stratified so that Elo buckets have enough support for stable metric curves.

To map from a Lichess.org Elo rating for a particular game type to an equivalent Chess.com rating, Elo+Chess uses the game-type-specific modal regression equations estimated in the companion Elo+Chess rating-mapping paper:

$$\begin{aligned}
\widehat{R}_{\text{Chess.com,bullet}} &= -531.71 + 0.9871 R_{\text{Lichess}}, \\
\widehat{R}_{\text{Chess.com,blitz}} &= -551.54 + 1.0853 R_{\text{Lichess}}, \\
\widehat{R}_{\text{Chess.com,10minrapid}} &= -495.04 + 1.0741 R_{\text{Lichess}}, \\
\widehat{R}_{\text{Chess.com,>10rapid}} &= -369.02 + 0.9374 R_{\text{Lichess}}.
\end{aligned}$$

Those equations map rating values, not percentile ranks. Percentile ranks are defined inside a particular player pool, and the Chess.com and Lichess player pools have different shapes. In Table 2, for the 10-minute rapid game type, the first column gives a small sample of possible Chess.com Elo rating values. The second column gives the corresponding Chess.com percentile rank displayed on Chess.com as of May 31, 2026. A 71.7% rank on Lichess.org for 10-minute rapid corresponds to a Lichess Elo rating of 1,644[2] and that value is shown in the third column. If we apply the Chess.com–Lichess mapping equation, 1,644 on Lichess is equivalent to about 1,271 on Chess.com. Assuming the estimated mapping equation is reasonable, the +522 gap between the Chess.com rating in column 1 and the mapped rating in column 4 can be explained by a large number of casual players in the Chess.com player pool inflating the Chess.com percentile rank. In other words, being in the 72nd percentile on Lichess.org is a larger accomplishment than being in the 72nd percentile on Chess.com, because percentile ranks do not cleanly map from one system to the other. Actual Elo values can be mapped using equations estimated from same-player scores across the two systems.

Chess.com rating	Chess.com pct.	Same-pct. Lichess	Mapped equiv.	Gap
749	71.7%	1,644	1,271	+522
1,012	86.2%	1,871	1,515	+503
1,358	95.4%	2,095	1,755	+397
2,383	99.9%	2,557	2,251	-132

Table 2: Illustrative comparison between Chess.com 10-minute rapid percentile points and the same percentile points in the active Lichess 10+0 rapid distribution after requiring at least five 10+0 games. "Mapped Chess.com equiv." applies the current Elo+Chess 10+0 rapid rating-scale conversion to the same-percentile Lichess rating. The highest-tail rows are included to show the scale distinction, but the fitted rating line is most reliable in the main coaching range.

The reason for the gaps in Table 2 is that the mapping estimates equivalent playing-strength ratings, whereas a percentile describes position within one platform's own population. A 71.7th-percentile player on Chess.com is therefore not automatically a 71.7th-percentile player on Lichess, and vice versa. The observed numbers are consistent with Chess.com's public rapid pool having a much larger lower-rated or casual-player mass, while the active exact-control Lichess pool is more concentrated among experienced repeat players. The key point is not that one percentile scale is

[2]Calculated from the cumulative distribution function of the Lichess 10+0 rapid latest-rating sample after requiring at least five games in that exact game type; see Table 1 and Figure 3.

"right" and the other is "wrong"; it is that percentile rank is not portable across platforms unless the underlying player distributions have the same shape.

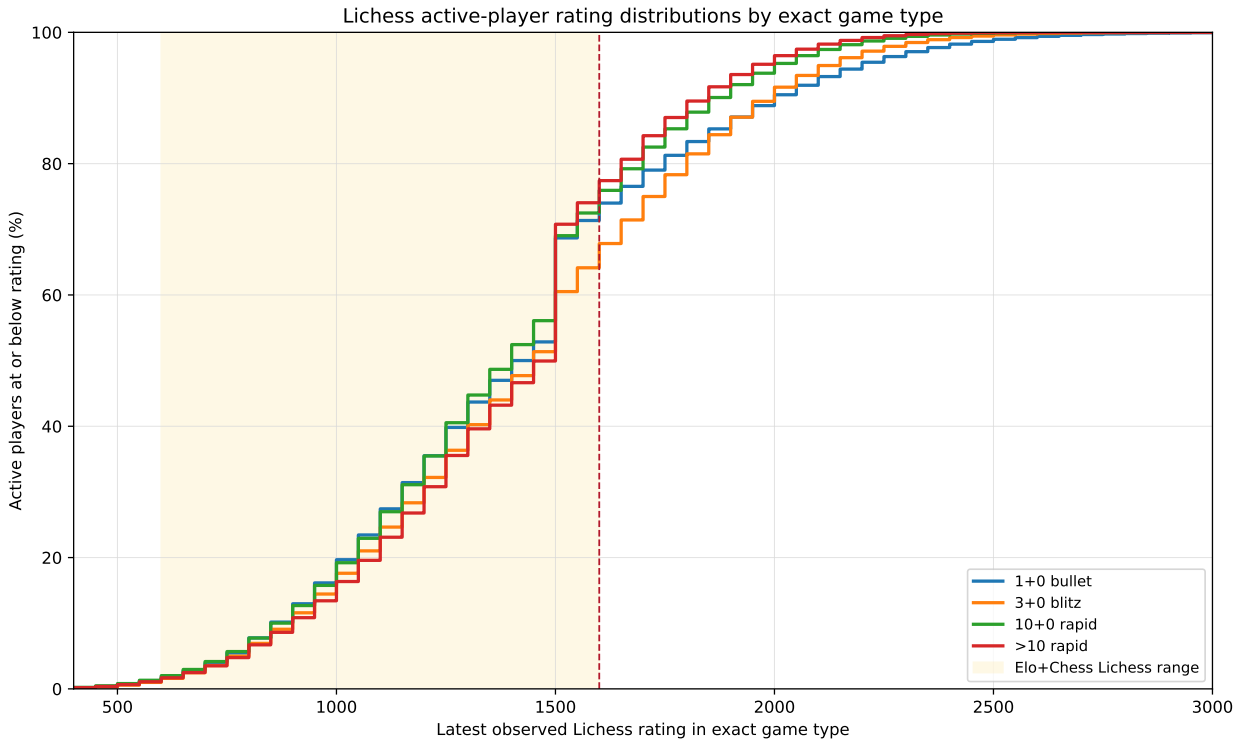


Figure 1: Cumulative active-player rating distributions by exact game type. The yellow band marks the 600–1600 Lichess range used by the main Elo+Chess coaching benchmarks.

Note on the 1500 spike. The visible jump near 1500 is present in the data rather than created by smoothing. In the latest observed rating distribution, exact 1500 ratings account for 13.1% of 1+0 bullet players, 5.6% of 3+0 blitz players, 9.4% of 10+0 rapid players, and 17.6% of longer rapid players. This is consistent with 1500 acting as a common starting or provisional anchor in the rating system.

The exact-1500 spike is largely an artifact of lightly active accounts. To check this, we recomputed the same latest-rating distributions after requiring each player to have at least five games in the exact game type. The spike nearly disappears under this filter: exact 1500 ratings fall from 5.6–17.6% of players in the all-player distribution to 0.2–1.2% in the five-or-more-games distribution reported in Table 1.

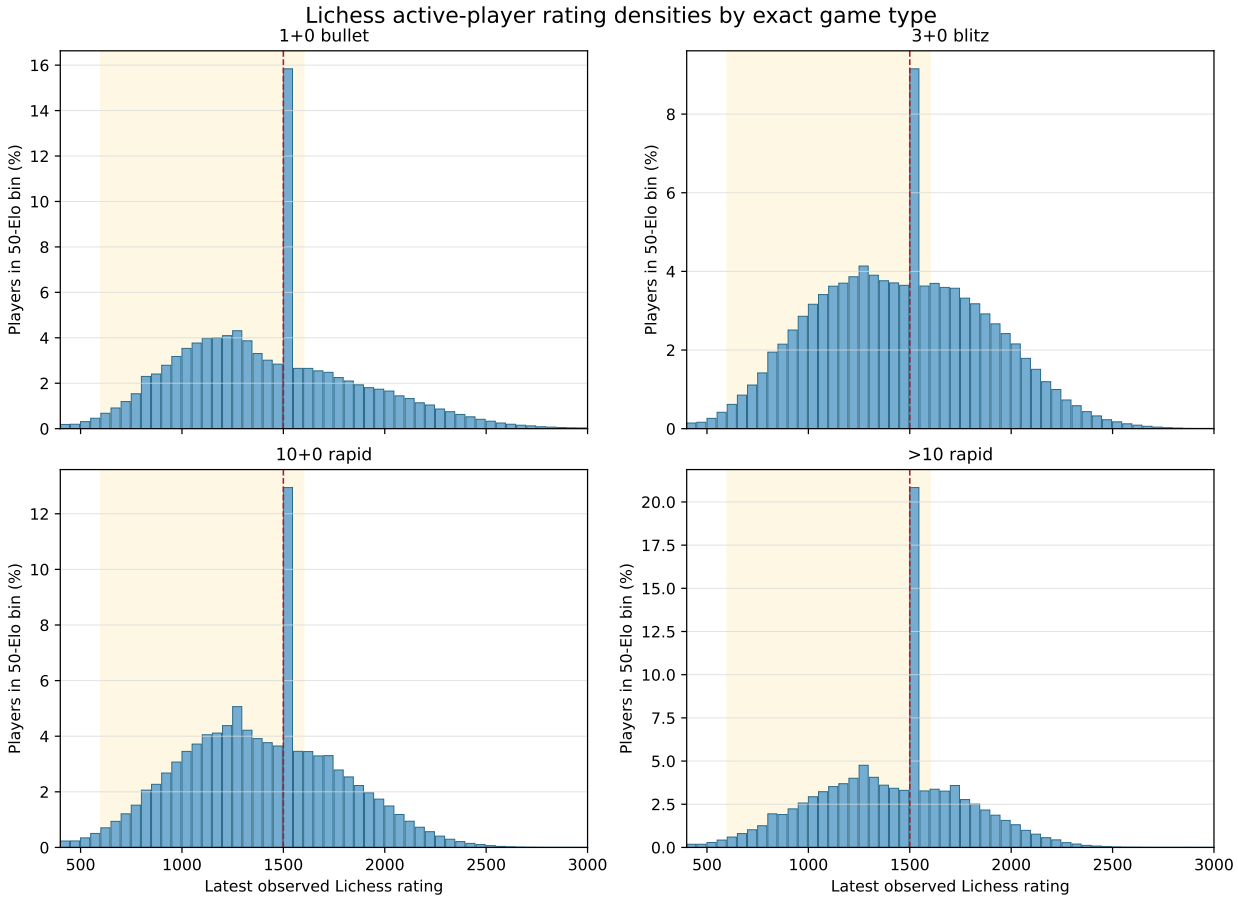


Figure 2: Density view of the same active-player rating distributions. The dashed line marks 1500 Lichess, where the raw latest-rating distributions have a large exact-rating mass point. The yellow band marks the 600–1600 Lichess range used by the main Elo+Chess metric reports. The spike is most visible in the all-player density because many lightly active accounts remain exactly at 1500; Figure 4 repeats the density plot after requiring at least five games in the exact game type, which largely removes this artifact.

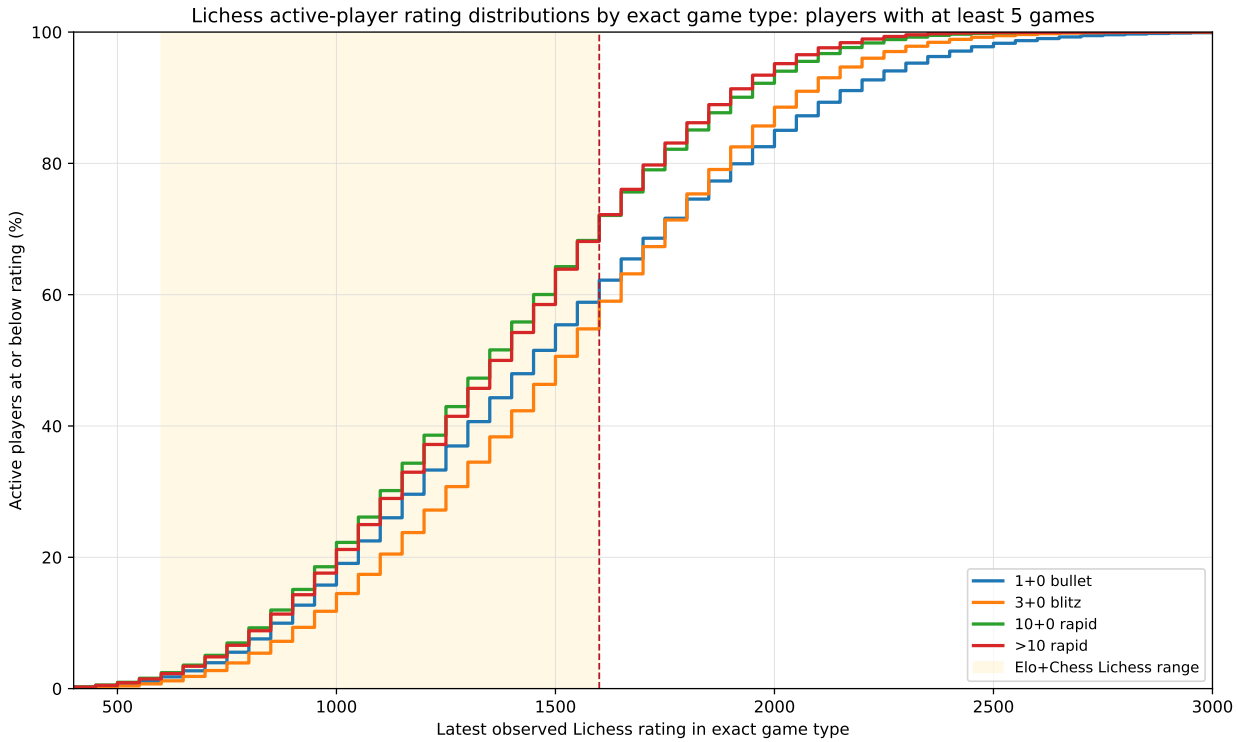


Figure 3: Cumulative active-player rating distributions after requiring at least five games in the exact game type. The 1500 jump is greatly reduced.

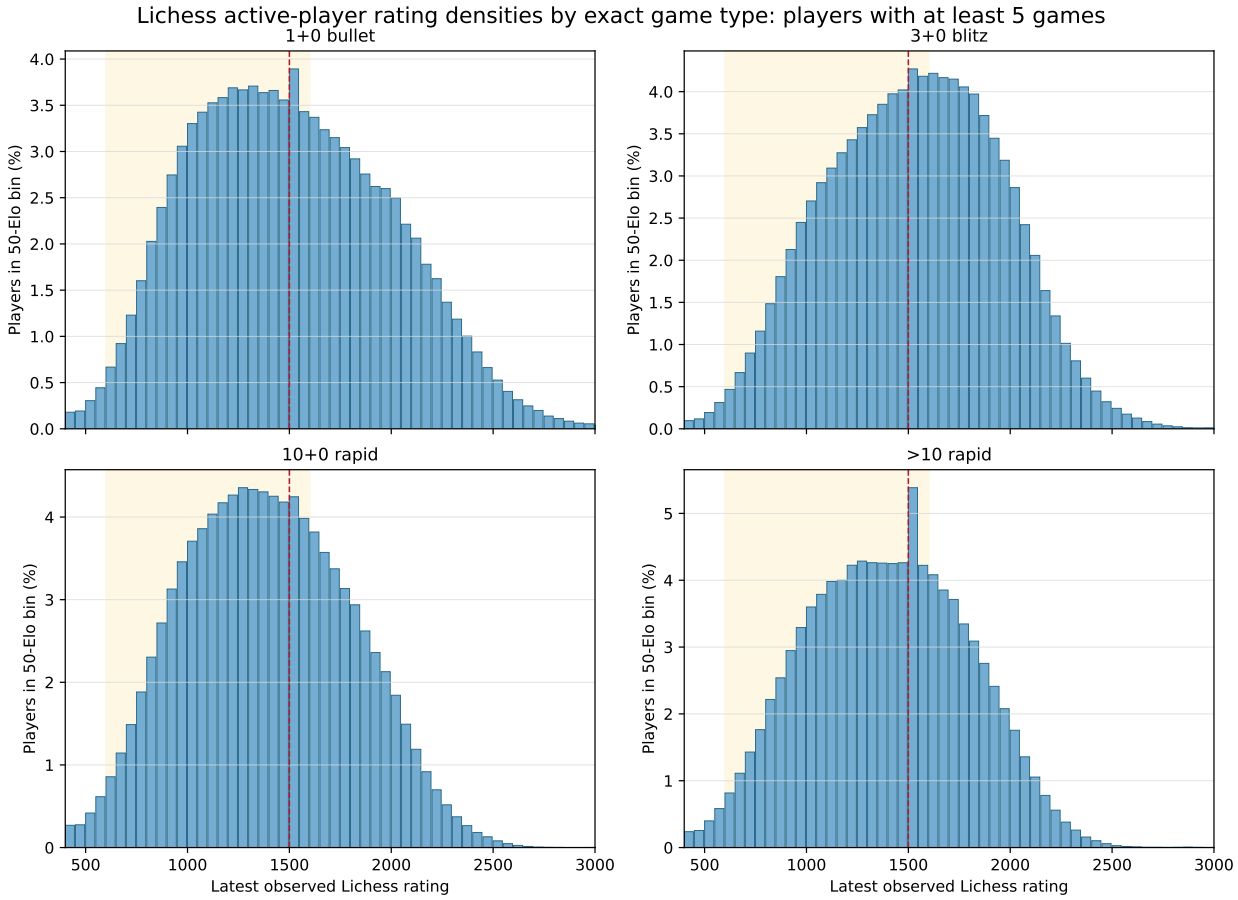


Figure 4: Density view after requiring at least five games in the exact game type. This view confirms that the large 1500 mass point in the all-player distribution is mostly driven by lightly active accounts. The yellow band marks the 600–1600 Lichess range used by the main Elo+Chess metric reports.

4 Scale of the Current Benchmark Artifacts

The computation is intentionally large. It is not a profile scrape, a small set of downloaded PGNs, or a hand-curated collection of instructive examples. The initial Lichess pass covered complete monthly files from January 2025 through March 2026. The unstratified scan table contains 1,382,178,087 games and 2,764,356,174 player-game rating rows. Across those rows, 6,623,438 unique player usernames appear at least once.

Speed	Games in raw scan	Player-game rows	Unique players
Bullet	519,375,423	1,038,750,846	3,077,635
Blitz	839,823,863	1,679,647,726	5,987,634
Rapid	22,978,801	45,957,602	1,535,865
All scanned speeds	1,382,178,087	2,764,356,174	6,623,438

Table 3: Unstratified Lichess scan scale before benchmark sampling. The scan covers complete monthly game-history files from January 2025 through March 2026.

The production benchmark artifacts are much smaller than the raw scan because they are selected for metric construction. They are still large: the benchmark stage is a move-state expansion and aggregation pipeline over hundreds of thousands of games per game type.

Game type	Games	Player-game rows	Unique players	Move states	Feature columns
1+0 bullet	483,974	967,948	119,732	22,985,272	277
3+0 blitz	479,974	959,948	129,398	28,138,904	277
10+0 rapid	925,636	1,851,272	229,235	57,554,339	277
>10 rapid	660,540	1,321,080	133,661	38,580,214	277

Table 4: Scale of the stratified benchmark artifacts after alignment to the common 277-column runtime feature schema. A "player-game row" means one game from one player's perspective, so most games produce two player-game rows.

For 3+0 blitz alone, the current production benchmark uses 479,974 games involving 129,398 unique players and 28,138,904 distinct move-state rows. The aligned blitz player-game feature table contains 277 derived variables. The rapid feature-build artifacts are even wider at the move-state stage: the full rapid move-state table used during feature construction contains 168 move-state columns before aggregation into a 277-column player-game feature table. The frontend runtime benchmark artifacts for all four game types are then aligned to the same 277-column player-game feature schema, including the expanded tactic-taxonomy breakout variables. From these feature tables, the report selects and displays the coaching metrics that are most interpretable.

These counts are important for interpreting the results. The benchmark curves are not estimates from a few hundred manually reviewed games. Instead, they come from tens of millions of evaluated board states and more than four million player-game rows across the four current production game categories.

5 Pipeline Overview

The benchmark construction pipeline has six major steps:

1. Select rated Lichess games in the target game type and rating strata.
2. Parse each game move by move and store the before-move and after-move board state.
3. Calculate intermediate move-state variables from the board, the move, the player perspective, and the opponent perspective.
4. Aggregate move-state variables into player-game feature variables.
5. Convert player-game variables into report metrics with a clear direction of interpretation: higher is usually better, lower is usually better, or descriptive.
6. Average each metric by Elo bucket and manually inspect the resulting relationship.

This structure is important because many report metrics are not directly visible in a PGN header. They require reconstructing the position, determining which pieces moved from which starting squares, comparing before-move and after-move material and structure, and then aggregating the result from the move level to the player-game level.

6 Intermediate Variables

The intermediate variables are designed to describe what changed on the board and why it matters from the report player's perspective. Examples include:

- move number, player move number, side to move, mover color, and player color;
- from-square, to-square, moving piece type, captured piece type, and whether the move was a capture, check, castle, promotion, or mate;
- before-move and after-move material balance;
- whether a minor piece left its starting square by a given move number;
- whether the same piece moved on consecutive player moves in the opening;
- whether the queen moved before move 10;
- castling side, castling move number, and whether rooks became connected after castling;
- loose pieces, hanging pieces, doubled pawns, isolated pawns, passed pawns, and backward pawns;
- rook file status, open files, half-open files, and king-safety features;
- tactical opportunities, executed tactics, material-pressure tactics, guaranteed follow-up gains, and realized gains.

The exact intermediate set varies by build stage. Runtime report databases keep only the columns needed to serve reports quickly, while the full feature-build artifacts include wider move-state tables and feature patches used to construct the final player-game feature rows.

7 Opening-Principle Metric Examples

Opening principles are useful examples because they are easy for users to understand and easy to connect to move-state calculations.

7.1 Minor Pieces Developed by Move 10

This metric counts how many of a player's four knights and bishops have left their home squares by the player's 10th move. The intermediate calculation uses the mover color, moving piece type, starting square, and player move number. For White, the home squares are b1, g1, c1, and f1. For Black, they are b8, g8, c8, and f8. For each player-game, Elo+Chess counts the distinct home squares among knights and bishops that moved by player move 10.

This is a higher-is-usually-better metric. It does not claim that every minor piece must always move by move 10, but among developing players the empirical relationship is clear: players who reach higher Elo buckets tend to develop more minor pieces earlier.

7.2 Repeated Piece Moves in the Opening

This metric counts extra early moves spent moving the same piece again instead of developing something new. The intermediate calculation uses the previous destination square, the current origin square, and the player move number. If the same piece moved on the previous player move and is moved again within the first 10 player moves, the repeated-move count increases.

This is a lower-is-usually-better metric. It measures wasted opening tempi in a simple way. Some repeated moves are tactically justified, but at scale the bucket curve captures the general tendency: stronger players at the target levels spend fewer early moves moving the same piece repeatedly.

7.3 Queen Touched Before Move 10

This metric counts queen moves before player move 10. The intermediate calculation uses moving piece type and player move number. If the moving piece is the queen and the player move number is at most 10, the count increases.

This is also lower-is-usually-better for the Elo+Chess target range. Early queen moves can win material or force concessions, but many developing players bring the queen out too early, lose tempi to attacks on the queen, and delay minor-piece development and king safety.

7.4 Castling and King Safety

Castling metrics use castling flags, castling side, castling move number, and the board state after castling. Examples include whether the player castled, whether the player castled by move 10, whether the player castled kingside or queenside, whether rooks were connected after castling, and whether rook files were open or half-open after castling.

These metrics are not all interpreted the same way. Castling completion and castling by move 10 are generally higher-is-usually-better in the target range. Queenside castling or castling into open files can be more context-dependent. That is why Elo+Chess does not simply hard-code a moral rule; it builds bucket curves and inspects which relationships are stable enough for coaching.

8 Bucket Curves and Manual Inspection

After player-game metrics are calculated, Elo+Chess groups them by Lichess Elo bucket and computes bucket averages. The bucket average is the raw empirical benchmark point for that metric. A trend line is then used in the report to assign Elo-style grades to user values.

The curves are inspected manually before being used. A metric is more useful when the bucket means show a coherent relationship across the rating range served by the product. A metric is less useful when it is noisy, flat, highly non-monotonic, or driven by special cases that are hard to explain to a user.

This manual inspection step is especially important above the Elo+Chess upper cutoff. Stronger players know when to break opening principles and do so effectively. Above the beginner-to-early-advanced range, the relationship between simple habits and rating often flattens or changes shape. The product therefore emphasizes the range where the empirical relationships are strong and pedagogically useful.

9 Example Full-Range Opening Curves

Figure 5 shows four 3+0 blitz opening-principle curves. Gray points show the wider Lichess rating range. Blue points and the blue trend line show the 600–1600 Lichess range used by the main Elo+Chess coaching benchmarks. The dashed red line marks the upper cutoff.

The same pattern is visible in the full-range 10+0 rapid and longer-rapid artifacts. The central modeling decision is unchanged: use the empirically stable range for coaching claims and avoid pretending that a simple habit curve remains equally informative forever.

The same pattern appears numerically in Table 5. Slopes are reported as metric-value change per 100 Elo points. Castling, minor-piece development, repeated piece movement, and early queen movement all show clear movement across 600–1600. Above 1600, the slopes are much smaller in these examples, which is consistent with the idea that higher-rated players break basic principles more selectively and more successfully. A second mechanism is saturation. For some habits, the metric reaches a practical ceiling or sweet spot: strong players have already learned the habit, so there is no reason for the curve to keep moving sharply upward. More is also not always better after that point. The data therefore flatten both because stronger players can break rules in concrete positions and because many basic habits have already reached their useful operating range.

10 Why the Upper Cutoff Matters

The Elo+Chess target user is a beginner-to-early-advanced player. For these players, basic habits are strongly associated with rating: develop pieces, avoid repeated early piece moves, avoid unnecessary early queen moves, castle in time, connect rooks, manage material, and avoid leaving pieces loose or hanging.

At higher ratings, those same surface habits become less diagnostic. A strong player may delay castling because the center is closed, move the queen early because the position contains a concrete tactic, or move the same piece twice because it wins a tempo or forces a structural concession. The raw habit still has meaning, but the relationship between the habit and rating becomes more conditional. This is why Elo+Chess uses an upper cutoff for its main coaching claims and why users above that range are warned that the tight relationship between principled habits and rating breaks down.

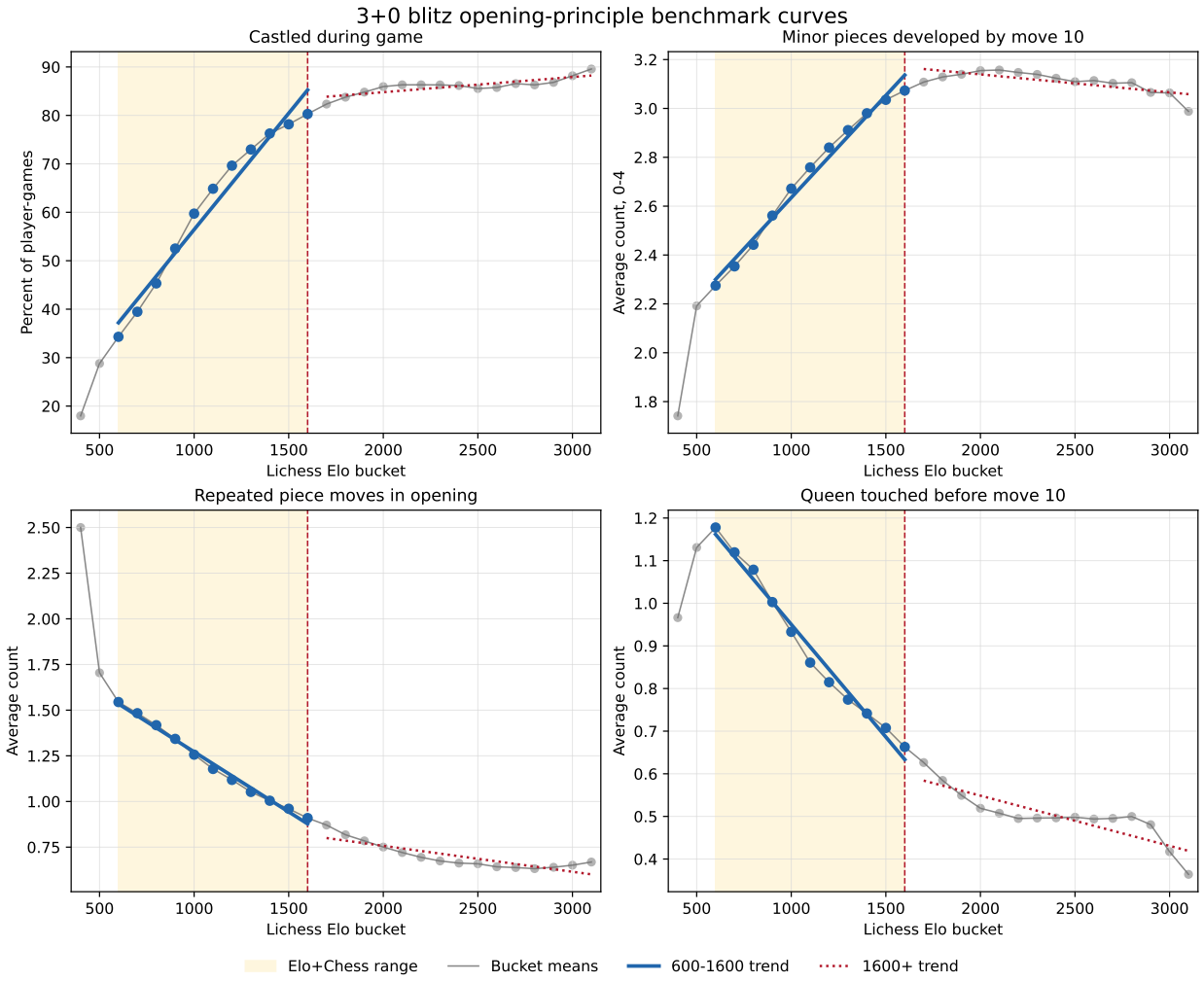


Figure 5: Selected opening-principle benchmark curves for 3+0 blitz. The relationships are strongest and most useful in the Elo+Chess coaching range. Above the upper cutoff, several relationships flatten substantially.

11 Two-Sided Tactical Metrics and Percentile Targets

Some metrics have the opposite shape from the opening-principle examples. Forks are a useful example because the event is inherently two-sided. A player who finds and converts more useful forks than peers is doing something good, but the population rate of converted forks falls with rating because stronger opponents allow fewer fork opportunities and defend against tactical threats more actively.

Figure 8 shows this effect. Both converted forks by the report player and converted forks by the opponent become less frequent as ratings rise. That does not mean executing forks is bad. It means the raw event rate is partly controlled by opponent error rate and game texture. For metrics like this, Elo+Chess therefore uses peer-history percentiles inside the player's own Elo bucket rather than relying only on a global bucket-mean trend.

To build these percentile targets, Elo+Chess samples players by game type and Elo bucket, then uses recent qualifying games for each sampled player. The current target set uses 11 buckets

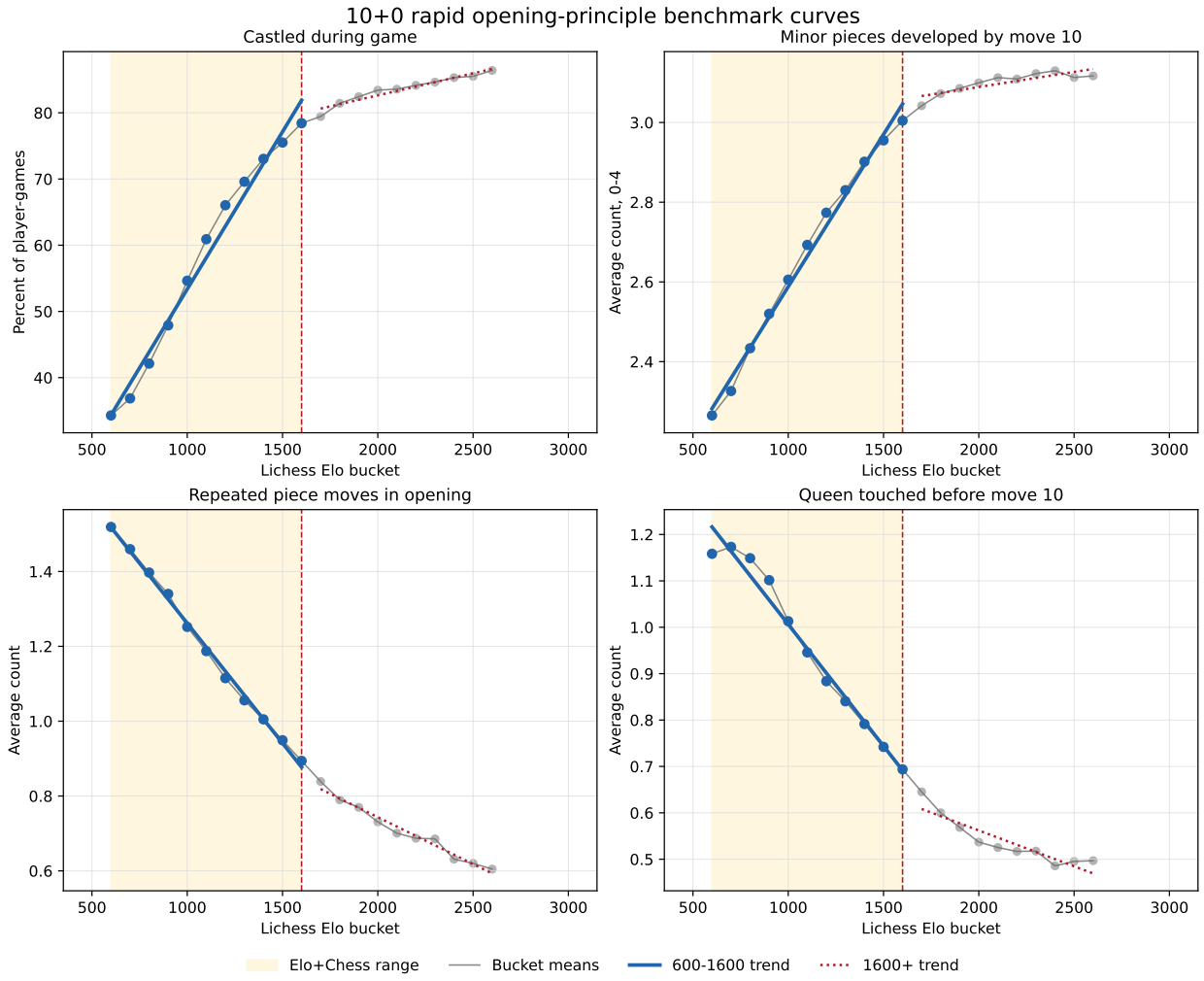


Figure 6: Selected opening-principle benchmark curves for 10+0 rapid games. The relationships are strongest and most useful in the Elo+Chess coaching range. Above the upper cutoff, several relationships flatten substantially.

from 600 through 1600, 200 players per bucket, and up to 50 recent qualifying games per player for each production game type. This gives 2,200 player samples per game type and approximately 110,000 player-game rows per game type for within-bucket percentile distributions.

Figure 9 illustrates the resulting target distribution for sampled-player converted fork rate in 3+0 blitz. Each box is a within-bucket distribution across sampled players, not a bucket mean across individual games. In the report UI, a user's value is compared to the relevant box for the user's mapped Elo bucket. This is the basis for statements such as whether a player's converted fork rate is high or low relative to peers at the same level, even when the overall population event rate declines with rating.

12 Replication Cost

The method is reproducible in principle, but it is not lightweight. Rebuilding the benchmark curves requires:

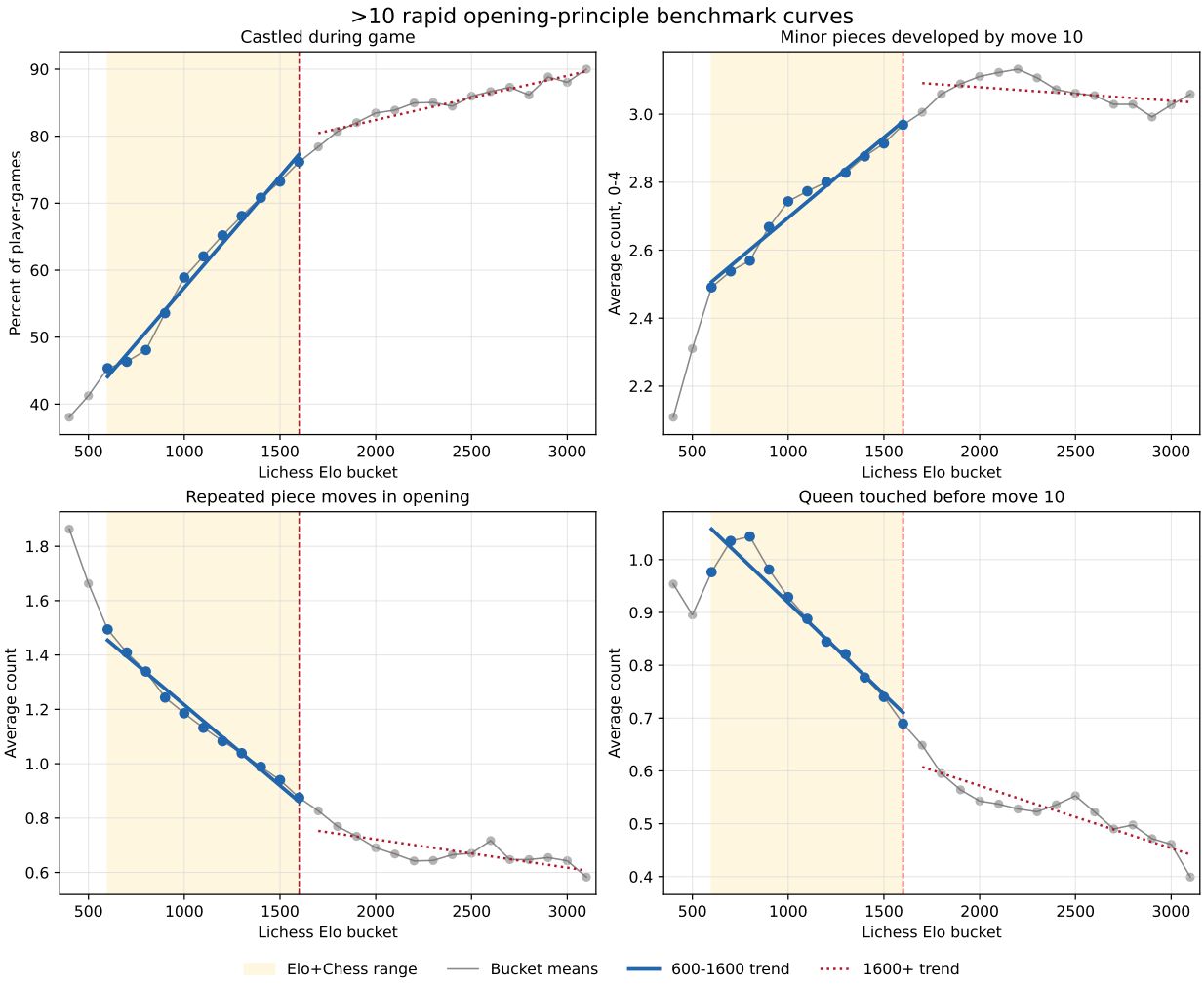


Figure 7: Selected opening-principle benchmark curves for longer rapid games.

- access to large Lichess monthly game-history files;
- stratified filtering by game type, rating bucket, and eligibility rules;
- a legal move parser and board-state evaluator capable of expanding tens of millions of move states;
- feature logic for opening, material, pawn structure, king safety, rook activity, tactics, time management, and game outcomes;
- aggregation from move state to player-game features;
- bucket-level summaries for every report metric;
- manual review of every candidate benchmark curve before it is used in a coaching report.

For 3+0 blitz, this means nearly half a million games, almost one million player-game rows, and more than 28 million move-state rows. Across the four benchmark game types, the current

Game type	Metric	600–1600 slope per 100 Elo	1600+ slope per 100 Elo
1+0 bullet	Castled during game	4.137	0.550
	Minor pieces developed by move 10	0.063	0.009
	Repeated piece moves in opening	-0.051	-0.026
	Queen touched before move 10	-0.040	-0.011
3+0 blitz	Castled during game	4.804	0.313
	Minor pieces developed by move 10	0.084	-0.007
	Repeated piece moves in opening	-0.066	-0.014
	Queen touched before move 10	-0.053	-0.012
10+0 rapid	Castled during game	4.753	0.665
	Minor pieces developed by move 10	0.076	0.008
	Repeated piece moves in opening	-0.064	-0.025
	Queen touched before move 10	-0.052	-0.015
¿10 minute rapid	Castled during game	3.320	0.659
	Minor pieces developed by move 10	0.047	-0.004
	Repeated piece moves in opening	-0.059	-0.010
	Queen touched before move 10	-0.035	-0.012

Table 5: Approximate bucket-curve slopes for the four example opening-principle metrics across production game types.

Game type	Elo buckets	Players per bucket	Games per player	Player samples
1+0 bullet	11	200	49–50	2200
3+0 blitz	11	200	27–50	2200
10+0 rapid	11	200	37–50	2200
¿10 minute rapid	11	200	47–50	2200

Table 6: Current peer-history percentile target sets. These are player-history samples used to estimate within-bucket distributions for percentile-style metrics.

benchmark artifacts contain more than 2.55 million games, more than 5.10 million player-game rows, and more than 147 million move-state rows. The scale is central to the value of the benchmark: it lets the report show users what actually happens in games played by players near their level rather than relying only on general chess slogans.

13 Limitations

The benchmark curves are empirical descriptions, not causal proof. A player does not gain rating merely by matching a bucket mean. Metrics are also imperfect summaries of complex positions: an early queen move can be excellent in one position and reckless in another. The purpose of the curves is to identify recurring habits and compare them against peer behavior, not to replace calculation or coaching judgment.

The curves are also Lichess-scale curves. When users bring Chess.com game histories, Elo+Chess maps the relevant rating bucket onto the Lichess benchmark scale using the separate cross-platform mapping model. That mapping layer is described in the companion Elo+Chess research note on Lichess-to-Chess.com rating conversion: Estimating Cross-Platform Chess Rating Mappings with Modal Regression. This mapping should not be read as preserving percentile rank. It preserves an estimated rating equivalence between the two rating systems; percentile rank remains platform-specific because the active player distributions are different.

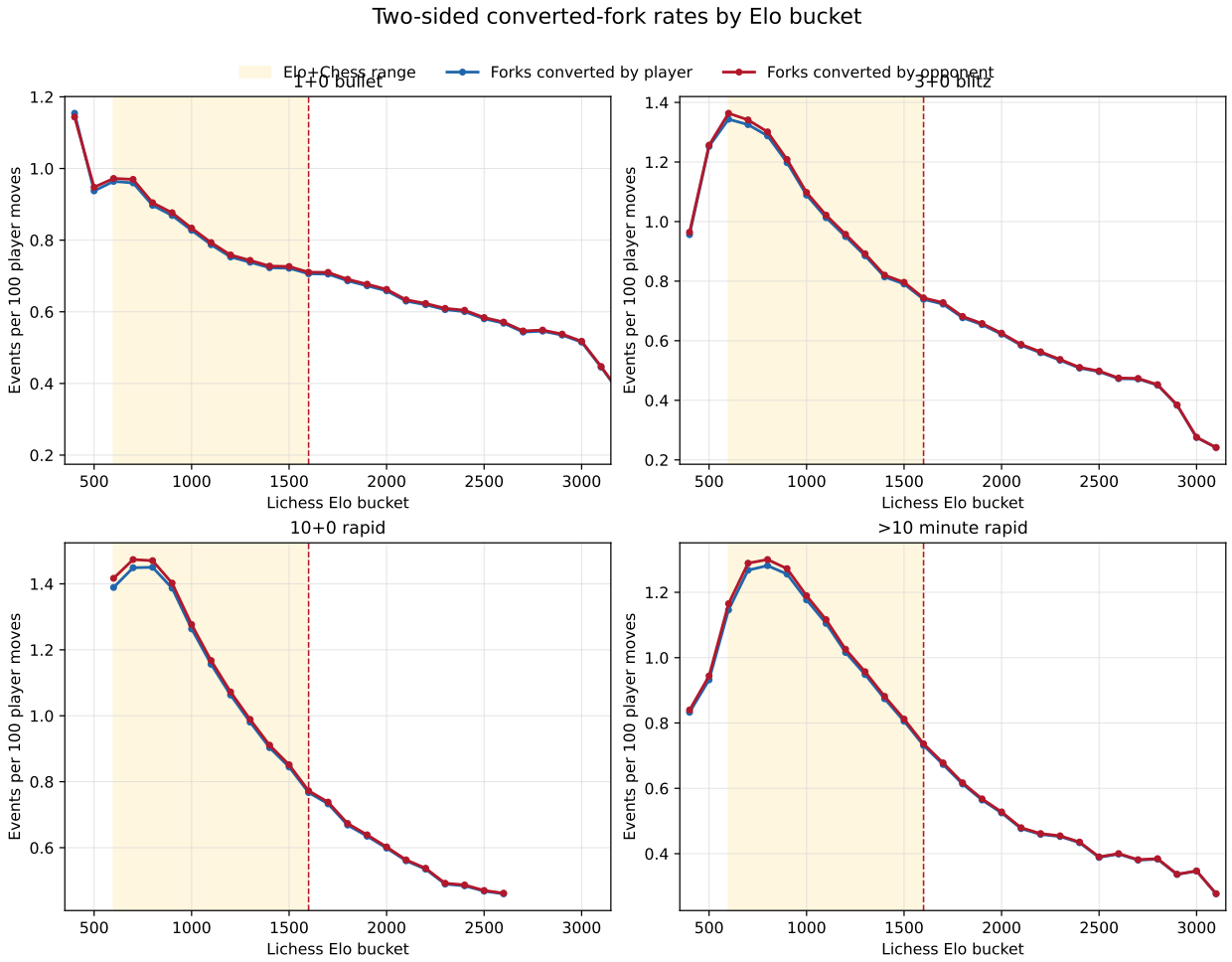


Figure 8: Two-sided converted-fork rates by Elo bucket. Converted fork events fall at higher ratings because both sides defend better and fewer clean fork conversions are available.

14 Production Report Graphics

The production report turns the benchmark artifacts into metric-level panels that can be read without knowing the full build pipeline. Figures 10 and 11 show three examples from the live Elo+Chess report design.

The opening-principle panels in Figure 10 use the ordinary benchmark-curve presentation. Each card starts with the metric family and metric name, then gives a short interpretation rule. The large grade in the upper-right corner is intentionally separated from the raw metric value: it answers the user-facing question "how does this habit grade against the benchmark?" while the rows below preserve the measured values. The four summary rows show Lifetime, Last 10, Wins, and Losses. This makes the panel diagnostic rather than merely decorative. If Lifetime and Last 10 agree, the habit is probably stable. If wins and losses diverge, the metric may be result-dependent or tied to game state rather than a simple habit.

The chart area uses a restrained dark background, gold trend line, and colored markers for the four slices. The trend annotation explicitly states whether the benchmark relationship rises or falls with Elo. This is important because some metrics are higher-is-better, such as minor-

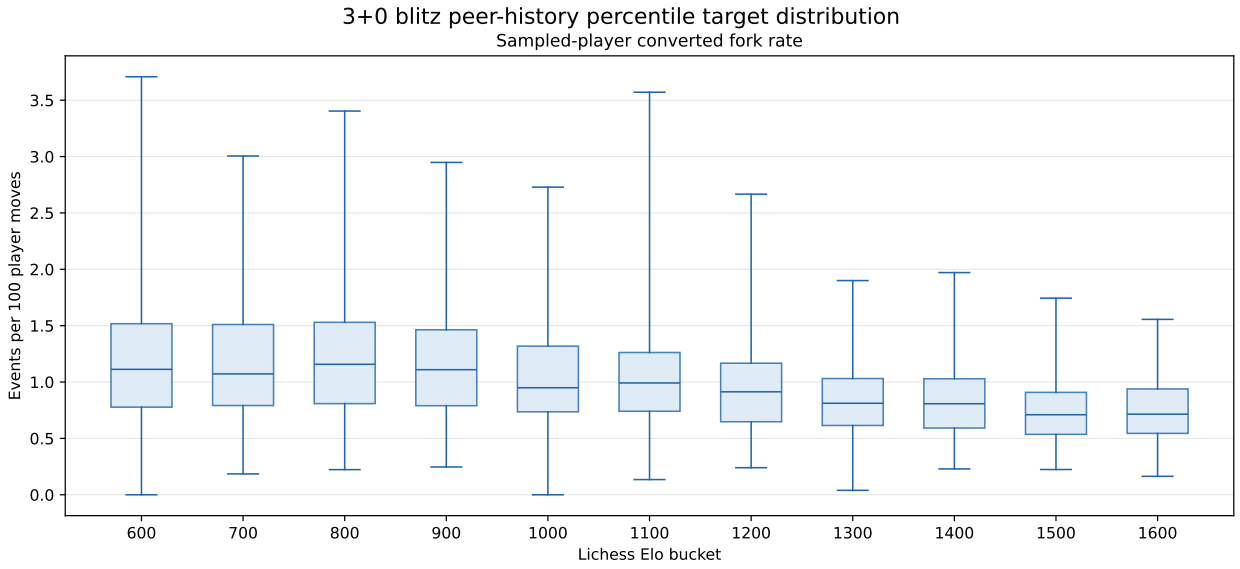


Figure 9: Example 3+0 blitz peer-history target distribution for sampled-player converted fork rate. Percentiles are computed within Elo bucket so the report compares a player with peers facing broadly similar tactical opportunity rates.

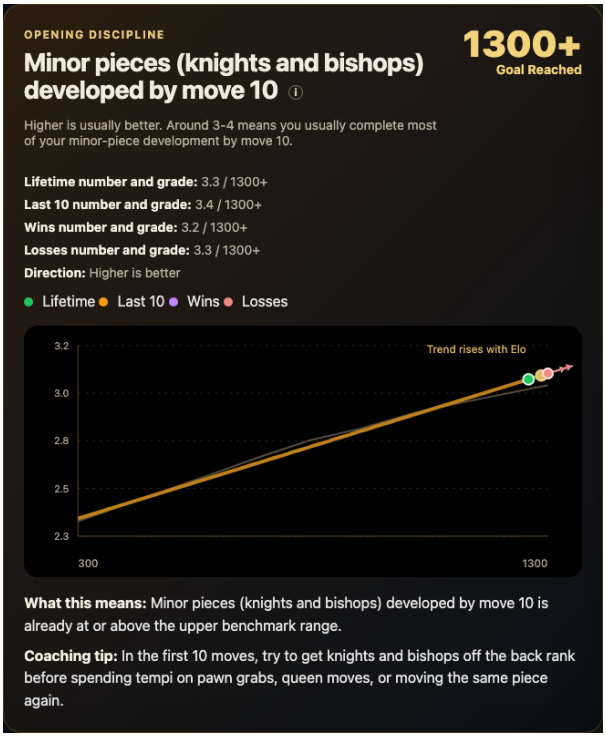
piece development, while others are lower-is-better, such as early queen movement. The card therefore does not ask the user to infer direction from the slope alone. The final "What this means" and "Coaching tip" lines translate the benchmark result into a concrete chess instruction while preserving the empirical context.

Figure 11 shows the percentile-style presentation used for metrics whose raw rate is not well represented by a single global trend. Winning forks are a tactical example: the population rate of converted forks changes with both player skill and opponent error rate. The card therefore compares the user's value with peer player histories in the mapped Elo bucket. The boxplot displays the peer distribution, including the interquartile range, median, and whiskers. The user's marker is plotted on the same scale and labelled with both raw value and percentile rank. Adjacent comparison buckets remain visible in muted form so the user can see that the peer bucket is local, not a universal scale.

This design makes percentile metrics auditable. The user sees the raw value, the percentile, the peer bucket used for comparison, and the translated Chess.com-to-Lichess bucket mapping. The same Lifetime, Last 10, Wins, and Losses rows are retained so tactical percentile scores can still be checked for recency effects and result dependence.

15 Conclusion

Elo+Chess benchmark reports are built from a large move-state calculation over stratified Lichess samples. Each metric begins as a set of concrete board-state and move-state variables, becomes a player-game feature, and is then averaged by Elo bucket to form an empirical benchmark curve. The opening-principle examples show why this approach is useful: simple chess habits have strong, visible relationships with rating in the target range, while those relationships often flatten or become more conditional above the upper cutoff. That is the central reason Elo+Chess focuses its coaching claims on the rating range where the data support them most clearly.



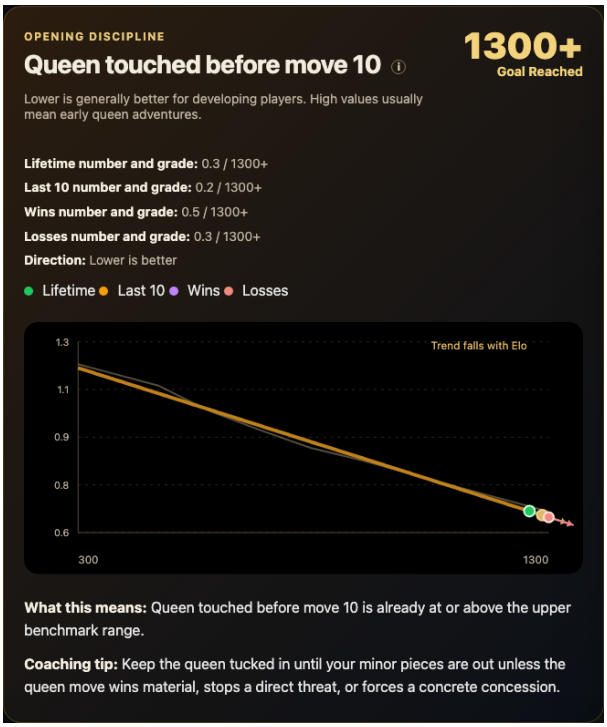


Figure 10: Production opening-principle panels. The same visual grammar handles a higher-is-better metric, minor-piece development, and a lower-is-better metric, early queen movement. Both panels expose Lifetime, Last 10, Wins, and Losses so the report can separate stable habits from recent or result-dependent patterns.

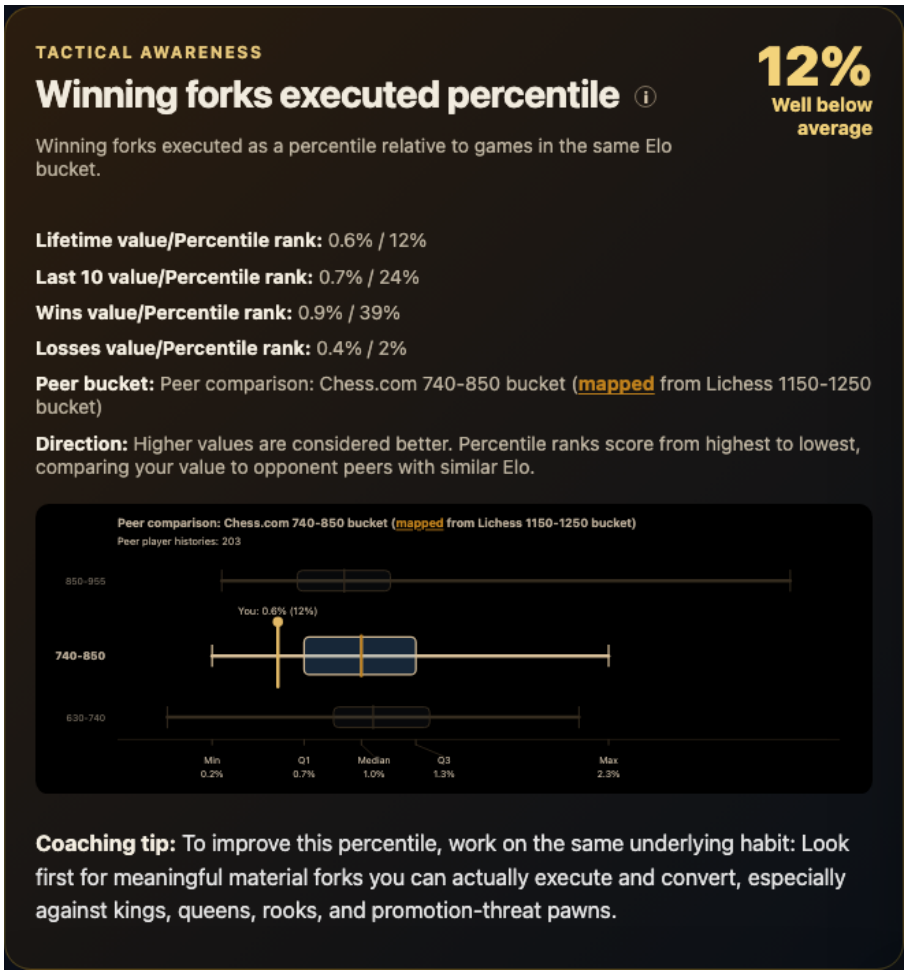


Figure 11: Production percentile panel for winning forks executed. The boxplot shows the same-bucket peer distribution used by the report, while the gold marker shows the user's raw value and percentile rank.