A WHITE PAPER · MAY 2026

The Primary Index.

Ranking and scoring journalism on the empiricism of its craft.

SEVEN METRICS SCORED BY A SINGLE CALIBRATED MODEL
A LIVE GLOBAL RANKING · A RECENCY-WEIGHTED HEADLINE

AUTHORS

Dr. Kyle C. Grant-Talbot
Mr. Solal Bauer
Dr. Aron P. D'Souza

Contents

Abstract. *We propose a ranking and scoring system — the Primary Index — for measuring the empiricism and craft of published journalism at the article and author level. Each article is scored by a single calibrated language-model evaluator against a seven-metric rubric and reduced to a macro score $A \in [0, 1000]$. The author's headline Primary Index is then a* ***recency-weighted mean*** *of their rated articles, with weights decaying smoothly as a function of an article's age-rank within the author's archive (newest article weight 1, oldest ≈ 0.0625; half-weight at the quarter-back of the archive). The headline number is computed first and the global leaderboard is sorted by it; rank is therefore a consequence of the headline, not its cause. Authors with five or more rated articles in the trailing 18 months are eligible for the public leaderboard, filterable by beat. Article scores are read-only once the rubric pass has run: the empiricism of the published piece is what is measured, and later actions by the author do not retroactively re-rate the artefact. New rank movement comes only from new work. The literature review (Part I) sets the design against Elo, Glicko-2, TrueSkill, Bradley-Terry, the h-index, FICO, PageRank, EigenTrust, WikiTrust, Stack Overflow, the contrast case of Duolingo/GitHub, and open peer-review platforms; the v3.5 architecture documented here intentionally simplifies away the Bayesian rating dynamics of earlier drafts in favour of the transparent recency-weighted mean, while retaining the reader-facing dual-display calibration (Author PI vs Article PI) and the eligibility-based seeding protocol. A simulation of 500 synthetic authors confirms a stable global ordering with target distribution properties (eligible cohort headline scores approximately $N(750, 80)$).*

Part I: Literature Review — Comparable Scoring and Ranking Systems

The Primary Index sits at an unusual intersection: it must rate individual *outputs* (articles) rather than outcomes of contests, aggregate those outputs into a *per-producer* reputation over time, and project the result onto a single live ordering of all eligible authors. No existing system maps cleanly onto this problem; the design below synthesises components from several lineages.

1.1 ELO RATING SYSTEM

Arpad Elo, a professor of physics and master-level chess player, formalised his rating system in the 1978 monograph *The Rating of Chessplayers, Past and Present* (Elo, 1978), which superseded his earlier papers (Elo 1961, 1965, 1967, 1973). The USCF adopted the system in 1960; FIDE in 1970.

Mechanic. Each player holds a rating R. Before a match, the expected score for player A against player B is:

$$E_A = (1)/(1 + 10^{(R_B - R_A)/400})$$

After the game, ratings update by:

$$R_A' = R_A + K(S_A - E_A)$$

where $S_A \in \{0, 0.5, 1\}$ is the actual outcome and K is a sensitivity constant (FIDE uses $K=40$ for new players, $K=20$ for established ones, $K=10$ for elite players above 2400).

Strengths. Computationally trivial; intuitive; converges to a meaningful equilibrium; self-correcting — a player rated too high will lose points until the rating reflects reality. Crucially for our purposes, **Elo produces a global ordering of all rated players**: every player has a rating, every rating is on the same axis, and the ranking is the natural consequence of sorting by rating.

Failure modes for journalist ranking.

- **No uncertainty quantification.** The system provides a point estimate with no confidence interval. An author with two articles and an author with two hundred articles can hold the same rating number with radically different evidential bases.
- **Requires head-to-head contests.** Elo is defined over pairwise outcomes. There is no natural opponent for a journalist writing an article; the "match" is against an abstract standard.
- **Rating deflation / inflation.** Closed systems suffer drift as cohorts enter and exit; player pools in journalism are fluid.
- **K-factor sensitivity.** A fixed K is either too volatile for established authors or too slow to update for newcomers.

Verdict. The Elo *update direction* — better-than-expected performance raises the rating, worse lowers it — and the *global-ordering* property are the right foundations. But the full Elo framework is poorly suited to an author-vs-rubric context without substantial adaptation.

1.2 GLICKO AND GLICKO-2

Mark Glickman at Boston University introduced the Glicko system to address Elo's uncertainty-blindness. The complete specification is in Glickman (2022), an updated and stable reference for the Glicko-2 system.

Mechanic (Glicko-2). Each player holds three parameters: rating r, rating deviation RD (standard deviation of the rating estimate), and volatility σ (expected degree of rating fluctuation). All are updated simultaneously after each rating period.

Converting to internal Glicko-2 scale:

$$\mu = (r - 1500)/(173.7178),\ \varphi = (RD)/(173.7178)$$

The auxiliary functions are:

$$g(\varphi) = (1)/(\sqrt{(1 + 3\varphi^2/\pi^2)}),\ E(\mu,\mu_j,\varphi_j) = (1)/(1 + e^{-g(\varphi_j)(\mu - \mu_j)})$$

The estimated variance from game outcomes is:

$$v = [\Sigma_{j=1}^{m}\ g(\varphi_j)^2\ E(\mu,\mu_j,\varphi_j)(1 - E(\mu,\mu_j,\varphi_j))]^{-1}$$

The estimated improvement:

$$\Delta = v\ \Sigma_{j=1}^{m}\ g(\varphi_j)(s_j - E(\mu,\mu_j,\varphi_j))$$

The volatility σ' is computed by the Illinois algorithm (a root-finding procedure) on:

$$f(x) = (e^x(\Delta^2 - \varphi^2 - v - e^x))/(2(\varphi^2 + v + e^x)^2) - (x - ln\sigma^2)/(\tau^2)$$

where τ is the system constant constraining volatility change (recommended $\tau \in [0.3, 1.2]$). New rating and RD:

$$\varphi' = (1)/(\sqrt{(1/(\varphi^2 + \sigma'^2) + 1/v)}),\ \mu' = \mu + \varphi'^2\ \Sigma_j\ g(\varphi_j)(s_j - E(\mu,\mu_j,\varphi_j))$$

When a player is inactive, only RD grows: $\varphi^* = \sqrt{(\varphi^2 + \sigma^2)}$, reflecting increasing uncertainty over time.

Strengths. Uncertainty is first-class. RD tightens with more data and widens with inactivity. The confidence interval (rating ± 2·RD) is meaningful. The volatility parameter handles erratic performance. Used by Lichess and the Australian Chess Federation. Like Elo, Glicko-2 produces a **global ordering**, and the per-player RD provides a principled tie-break: among two players with similar ratings, the one with lower RD is more credibly placed.

Failure modes for journalist ranking. Still requires pairwise opponents. The volatility concept needs reinterpretation for a production context (article frequency rather than game frequency).

Verdict. Glicko-2 was the core update machinery for earlier drafts of the Primary Index, and remains the closest formal analogue to what the system tries to achieve — a principled, uncertainty-aware ordering of a population of performers. The v3.5 architecture documented in Part II steps away from Glicko-2 in favour of a transparent recency-weighted mean, on the grounds that the binary-outcome framing imported a

competitive metaphor unsuited to journalism and that the Bayesian machinery was harder to explain to editors than the underlying signal warranted. Glicko-2's contribution to the design — first-class uncertainty, a global ordering, a meaningful confidence interval — is retained in the v3.5 confidence-surface treatment (§2.7); the update equations are not.

1.3 TRUESKILL AND TRUESKILL 2

Herbrich, Minka, and Graepel (2007) at Microsoft Research introduced TrueSkill as a principled Bayesian generalisation of Elo for multiplayer and team games, deployed commercially in Xbox Live's Halo 2 matchmaking. Minka, Cleven, and Zaykov (2018) extended it to TrueSkill 2, which incorporates individual performance statistics (kill counts, completion rate) and cross-game-mode skill correlation.

Mechanic. Each player's skill is modelled as a Gaussian: $skill_i \sim N(\mu_i, \sigma_i^2)$. Match outcomes are generated by a probabilistic model; Bayesian inference (via Expectation Propagation on a factor graph) updates the Gaussian posteriors. The displayed rating is $\mu_i - k\sigma_i$ (a conservative lower bound), ensuring high confidence before displaying a high rating.

Strengths. Full Bayesian treatment. Handles partial information (individual kills vs. team outcome). TrueSkill 2 predicts match outcomes at 68% accuracy vs. 52% for classic TrueSkill in Halo 5. The factor-graph formulation allows modular extension.

Failure modes for journalist ranking. Designed for real-time matchmaking; the factor-graph inference is operationally complex. The multiplayer team structure does not map to solo journalism.

Verdict. TrueSkill's key contribution for our purposes is the *displayed rating as a conservative lower bound* — the concept that reported skill should incorporate uncertainty, not just the point estimate. This principle is adopted in Part II via the explicit confidence interval on every Author PI (§2.7), in the form **PI [±k]**, derived from the weighted standard error of the headline rather than a Bayesian variance.

1.4 BRADLEY-TERRY MODEL

Bradley and Terry (1952) proposed in *Biometrika* 39:324 a model for pairwise comparison data:

$$(i \text{ beats } j) = (\lambda_i)/(\lambda_i + \lambda_j) = (e^{\beta_i})/(e^{\beta_i} + e^{\beta_j}) = \sigma(\beta_i - \beta_j)$$

where $\lambda_i > 0$ is item i's strength and $\beta_i = \ln \lambda_i$. Parameters are estimated by maximum likelihood (MLE) via the iterative Zermelo algorithm. Recent work (Ruf & Wein, 2022, JMLR 24) shows convergence can be achieved over 100× faster with an alternative iteration.

Strengths. Produces globally consistent, transitive rankings from noisy pairwise data. Log-linear structure connects naturally to logistic regression and Bayesian extensions.

Failure modes for journalist ranking. Requires pairwise contests. Static model — does not handle temporal dynamics or uncertainty growth through inactivity. No native uncertainty representation.

Verdict. Bradley-Terry is the theoretical ancestor of Elo and a useful lens for understanding *why* pairwise comparisons produce consistent orderings. It is most relevant to a potential future module: human editorial panels comparing two articles on the same story — but it is not the primary mechanism.

1.5 H-INDEX AND ACADEMIC CITATION INDICES

J.E. Hirsch (2005), published in *PNAS* 102(46):16569, proposed the h-index as a single number characterising a researcher's output: *a scientist has index h if h of their papers have each been cited at least h times.*

Mechanic. The h-index is defined on the joint distribution of paper count and citation count. It is immune to a few highly-cited papers inflating the metric, unlike raw citation counts, because both dimensions must simultaneously exceed the threshold.

Strengths. Simple, robust, and field-legible. Because it is threshold-based, a single viral article cannot distort the score. It rewards sustained production over time. Google Scholar profiles display it prominently, giving it social reality in academic communities.

Failure modes for journalist ranking.

- **No time decay.** A 1989 paper still counts as much as a 2024 paper. In journalism, recent work is more diagnostic of current craft.
- **No uncertainty.** The h-index is a point estimate; a researcher with h=10 from 50 papers is counted identically to one with h=10 from 11 papers.
- **Not normalised for field productivity.** In journalism, the analogue problem is that beat reporters produce more articles than feature writers; pure volume rewards frequency over quality.
- **Cannot decrease easily.** Citations rarely disappear, so h-index is essentially monotone. For a credible journalism index, scores must be allowed to fall when warranted.

Verdict. The h-index provides an important design principle — *the score should reflect sustained excellence, not one exceptional event* — and its social legibility (a single prominent number) is worth emulating. We adopt a similar philosophy of a headline index number with per-article drilldown, but address the h-index's failure modes explicitly through the recency-weighted mean (which prevents one strong article from dominating a long career) and the live ranking layer (which keeps the signal current). An h-index-style summary survives as a **reader-facing display badge** rather than the sort key: see §2.6.1 on display vs sort.

1.6 FICO CREDIT SCORE

The FICO Score (Fair Isaac Corporation) is the dominant consumer credit risk metric in the United States, used in more than 90% of lending decisions. FICO's published methodology describes five weighted components: payment history (35%), amounts owed (30%), length of credit history (15%), new credit inquiries (10%), and credit mix (10%). Scores range from 300–850.

Calibration philosophy. FICO's most instructive design choice for our purposes is *deliberate range compression*. The 300–850 band is not uniform: the median American score sits near 714 (as of 2024), with the vast majority of scoreable consumers falling between 580 and 800. The extremes (below 500, above 820) are thinly populated and represent genuinely exceptional situations.

Strengths for our context. FICO demonstrates that a credible professional scoring system need not expose its algorithmic internals to be trusted. Users engage seriously with a single number (their credit score) and modify their behaviour based on it without requiring a full understanding of the weighting formula.

Anti-gaming mechanisms. FICO is particularly instructive on gaming resistance. It uses a *soft vs. hard inquiry* distinction. It uses *relative* measures (utilisation ratio, not absolute balance). It caps the benefit of "length of history" so that very young files still produce usable scores.

Verdict. FICO is the closest analogue to the Primary Index in terms of social function: a single authoritative number that serious professional stakeholders treat as a real signal. The Primary Index dual-number system (author PI from rank percentile + article PI from the rubric) preserves the FICO-style headline while adding the live-ranking element FICO lacks.

1.7 PAGERANK AND REPUTATION GRAPHS

Brin and Page (1998) proposed PageRank as the ranking backbone of Google: the quality score of a page is the weighted sum of the quality of pages linking to it, iterated to convergence. Formally, for page A with in-links from pages $T_1,\ldots,T_n$:

$$PR(A) = (1\text{-}d)/(N) + d\,\Sigma_{i=1}^{n}\,(PR(T_i))/(C(T_i))$$

where $d \approx 0.85$ is the damping factor, $C(T_i)$ is the out-degree of page T_i, and N is the total number of pages. PageRank is the principal eigenvector of the web's link matrix.

Strengths. Transitive trust: being cited by a highly trusted source is worth more than being cited by an unknown one. Robust to simple link spam. The iterative power method converges quickly. PageRank produces a global ranking by construction.

Failure modes for journalist ranking. Articles cannot link to each other in a meaningful way that maps to the citation structure. The metric reflects popularity of sources cited, not the quality of reasoning applied to them.

Verdict. PageRank's *transitive trust* concept is relevant to outlet-level scoring but is not the primary author-level mechanism. PageRank's ranking-as-output structure validates our choice to make the live leaderboard the headline surface rather than a supplementary view.

1.8 EIGENTRUST

Kamvar, Schlosser, and Garcia-Molina (2003) at Stanford (published in ACM WWW '03) introduced EigenTrust for reputation management in peer-to-peer file-sharing networks. Each peer i computes a local trust value for peer j based on download quality history, normalised to sum to 1. The *global* trust vector t satisfies:

$$t = (1-\alpha)C^{\top} p\ t + \alpha\ p$$

where C is the normalised local trust matrix, p is a pretrusted seed vector, and α governs the influence of the seed.

Strengths. Computationally tractable distributed reputation. The pretrusted-seed mechanism provides a meaningful anchor against coordinated malicious collectives (resilient up to 70% malicious peers in simulation).

Failure modes for journalist ranking. Trust in EigenTrust is derived from peer endorsement, not from objective quality assessment.

Verdict. EigenTrust's contribution is the *pretrusted seed* architecture. The Primary Index analogue is the scoring LLM — a fixed, consistent evaluator that cannot be socially manipulated by the author or the outlet.

1.9 WIKITRUST

Adler, de Alfaro, and Pye, working at UC Santa Cruz, built WikiTrust: a content-driven reputation system for Wikipedia that assigns per-word trust values and per-author reputation based on edit survival.

Key empirical finding. Low-reputation authors (bottom 20%) account for 18.1% of edits but 82.9% of reverted edits. Text in the bottom half of reputation has a 33% probability of deletion in the next revision vs. 1.9% for general text.

Strengths. Scores *move in both directions* — a core property we require. Empirically validated. Content-driven, not socially-driven.

Failure modes for journalist ranking. Relies on a large active editorial community. Journalism articles are not collaboratively revised.

Verdict. WikiTrust provides the clearest existence proof that *bidirectional score movement* in a content-quality system is feasible and desirable. In the Primary Index the same property is achieved not by re-judging past articles but by the recency-weighted mean: newly published work enters the headline at weight 1 and older articles decay by age-rank, so an author's headline moves in both directions purely as a function of the work they publish.

1.10 STACK OVERFLOW / STACK EXCHANGE REPUTATION

Stack Overflow launched in 2008 (Atwood, 2008; Spolsky, 2008). Its reputation system awards +10 for upvotes, deducts -2 for downvotes (-1 for the voter), and awards +15 for accepted answers, with a daily cap of 200.

Strengths. The *symmetry of loss*: authors lose reputation for downvoted content. The *voter cost*: casting a downvote costs the voter 1 reputation point. The daily cap prevents runaway accumulation. Privilege thresholds (edit at 2000 rep, vote-to-close at 3000 rep, delete at 10000 rep) create a meritocratic trust escalation.

Failure modes. Topic concentration produces systematic inflation for popular-topic specialists. Late answers — even correct ones — receive fewer votes.

Verdict. Stack Overflow demonstrates that a reputation system for professionals *can* work without gamification, provided the underlying activity is intrinsically meaningful. The downvote mechanic informs the bidirectional design of evidence outcomes.

1.11 DUOLINGO XP AND GITHUB CONTRIBUTIONS GRAPH (CONTRAST, NOT MODELS)

These systems are reviewed *as anti-models* — explicitly to document why they are rejected.

Duolingo uses streaks, XP, leagues, and push notifications, engineered around loss-aversion, variable reward schedules, and social comparison pressure. These mechanics produce engagement that is dissociable from learning outcomes — users optimise for streak preservation rather than skill acquisition.

GitHub Contributions Graph displays a year-view heatmap of commit activity, rewarding activity volume rather than contribution quality.

The explicit rejection. The Primary Index is designed *without* streaks, daily-activity metrics, or push notifications. The live leaderboard is intentionally framed as a *professional ranking* (like the Bloomberg Terminal trader leaderboard or the ATP tennis rankings), not as a gamified competition. Rankings update on a per-article and per-evidence cadence —

measured in days and weeks, not hours — and the system never sends behavioural nudges.

1.12 OPEN PEER REVIEW (F1000RESEARCH, PUBPEER)

F1000Research practices post-publication peer review: articles are published immediately, then formally reviewed by invited referees whose names and reports are public. The article is indexed in PubMed only after two referee "Approved" verdicts. PubPeer allows post-publication community annotation of any paper.

Relevance to the Primary Index. F1000Research demonstrates that a *published record* can carry a transparent, auditable score without depending on private editorial verdicts. The Primary Index adopts three of the same commitments:

- **Transparency.** All per-metric scores and their provenance are permanently visible on the article profile.
- **Timestamped audit trail.** Every score is tied to the published article's timestamp; the LLM verdict is version-pinned.
- **Read-only artefacts.** Once the rubric pass has run, the article score is fixed — F1000Research's referee verdicts are similarly attached to a published version rather than retroactively altering it.

Verdict. F1000Research validates the design choice to treat each article as a permanent, publicly-scored artefact. The Primary Index diverges in one respect: it does not accept post-publication counter-evidence or author rebuttals against a published score. The empiricism of the article as published is what is measured; new craft signal comes from new work, not from litigating old work.

Part II: Proposed Ranking Methodology

2.1 SYSTEM OVERVIEW

The v3.5 Primary Index is built in four layers plus a one-time seeding pass:

1. **Rubric scoring (§2.2).** Per-article score $A \in [0, 1000]$ from the seven-metric rubric, evaluated by a single scoring LLM customised for this task.
2. **Author headline aggregation (§2.4).** The author's headline Primary Index is a **recency-weighted mean** of their rated articles. Weights decay with age-rank inside the author's archive: the newest article has weight 1, the oldest ≈ 0.0625, with half-weight at the quarter-back of the archive (decay constant $H = 0.25$).
3. **Global rank (§2.5).** Authors with at least five rated articles in the trailing 18 months are eligible for the public leaderboard. The leaderboard is sorted by the headline PI — the number is computed first and rank is its consequence, not its cause.

4. **Display calibration (§2.6).** Two numbers carry the same name. The **Author PI** is the recency-weighted-mean headline, mapped onto the *N(750, 80)* display band by a fixed monotone transformation. The **Article PI** is per-article and derived from *A* directly by the same absolute calibration.

Seeding (§2.4.5). Before live operation, the system ingests as much of each candidate author's back catalog as the outlet exposes — typically 500–1,000 candidate authors with archives often spanning a decade or more. Each archived article is rubric-scored once and contributes to the same recency-weighted mean used in live operation; the launch ordering is therefore continuous with steady-state behaviour, not a separate algorithm spliced onto it.

Architectural separation. The v3.5 architecture maintains an explicit separation between **score** (the article-level macro *A*, an absolute measurement), **headline** (the author's recency-weighted mean, a measurement aggregate), and **rank** (the integer position derived by sorting headlines). Earlier drafts of this paper described a Glicko-2 Bayesian rating with explicit volatility; v3.5 intentionally simplifies that machinery away. The mean is auditable cell-by-cell; the weights are a published function of an author's own article order; an editor can reproduce any author's PI from the public article list with a hand calculator. The dual-display calibration described in §2.6 (display vs sort) preserves the reader-facing intuitions that made the earlier dual-PI architecture work, while keeping the production sort key as the same transparent number that appears on the leaderboard.

2.2 ARTICLE-LEVEL SCORE (GIVEN)

Each article receives a score $A \in [0, 1000]$ from the seven-metric rubric, evaluated by a single LLM that has been customised — system prompt, rubric grounding, calibration anchors, and worked examples — for this scoring task.

Two-stage construction. The scoring LLM scores each of the seven metrics on a 0–1000 internal scale (Appendix B), producing per-metric scores m_i. A fixed weight vector $w = (w_1, \ldots, w_7)$ with $\Sigma\, w_i = 1$ produces the macro article score:

$$A = \Sigma_{i=1}^{7}\, w_i \cdot m_i,\ A \in [0, 1000]$$

The v3.5 weight vector locks the relative importance of the seven metrics. The three primary sourcing metrics carry equal top weight (0.20 each), the balance/uncertainty metric carries the largest middle weight (0.1333), and the three category-sensitive metrics share the remaining 0.2667 evenly:

#	METRIC	/1000 RAW	WEIGHT W_i
1	Use of Sources & Attribution (SO)	0–1000	0.2000

#	METRIC	/1000 RAW	WEIGHT w_i
2	Evidence Quality & Relevance (EQ)	0–1000	0.2000
3	Evidence Traceability (ET)	0–1000	0.2000
4	Balance & Acknowledgement of Uncertainty (BU)	0–1000	0.1333
5	Fact vs Opinion Separation (FO)	0–1000	0.0889
6	Polemical Language (PL)	0–1000	0.0889
7	Source Balance & Right of Reply (SB)	0–1000	0.0889
	Total weight		**1.0000**

The weights sum to 1 to four-decimal precision; the small residual ($2 \cdot 10^{-4}$) is absorbed silently in the macro computation and never reaches the displayed three-digit Article PI. The 0–1000 per-metric internal scale is preserved with a per-metric ceiling of 1000.

Derived labels (not in the weighted index). The scoring LLM also produces three derived labels that are surfaced to readers but never enter the numeric score: *article type* (news report, news analysis, feature, opinion, review, explainer, obituary), *political leaning* on a seven-point scale (Far Left, Left, Centre-Left, Centre, Centre-Right, Right, Far Right), and *anonymous-source usage* (none / single anonymous source / multiple anonymous sources). These labels inform category-aware calibration of metrics 4 and 6 — a news report and an opinion column are not held to the same Fact vs Opinion Separation threshold — but no derived label adds points to or subtracts points from any metric.

2.3 ARTICLE-TO-HEADLINE MAPPING

In earlier drafts this section described a Glicko-style binary outcome mapping ($s \in \{0, 0.5, 1\}$) that fed an Elo-family rating update. The v3.5 architecture does not use a binary outcome at all. Each article enters the author's headline at its full real-valued macro score $A_k \in [0, 1000]$, modulated by a recency weight defined in §2.4. The rubric "opponent" of the earlier formulation — a fixed virtual player at rating 1500 — is no longer needed because the headline is no longer derived from win-loss outcomes against it.

Two motivations drove the simplification. First, the binary outcome discarded information: an article scoring 850 and one scoring 705 both produced $s = 1$ and shifted the author's rating identically given equal deviation. The recency-weighted mean preserves the full real-valued signal. Second, the win/loss framing imported a competitive metaphor that the

credible-professional design of Part III explicitly disavows: journalism is not a game played against the rubric, and the rubric is not the journalist's opponent. The v3.5 headline reads as what it is — an average of recent rated scores, weighted toward the present.

2.4 AUTHOR HEADLINE: THE RECENCY-WEIGHTED MEAN

2.4.1 Definition

Let $\{A_k\}_{k=0}^{N-1}$ be an author's rated articles, indexed from $k=0$ (most recent) to $k=N-1$ (oldest), with the ordering taken from each article's published_at timestamp (falling back to scraped_at where publication date is unavailable). Define the **age rank** of article k as:

$$p_k = (k)/(N-1) \in [0, 1]$$

so $p_0 = 0$ for the newest article and $p_{N-1} = 1$ for the oldest. The recency weight assigned to article k is:

$$w_k = 0.5^{\,p_k / H},\ H = 0.25$$

The constant H is the **age-rank half-life**: the position in the archive at which the weight is half that of the newest article. With $H = 0.25$, the quarter-back of the archive contributes half-weight, the half-back contributes quarter-weight, and the oldest article carries $0.5^4 = 0.0625$ (about 6.25%) of the newest article's weight. Older work is present in the signal but never dominant.

The author's headline Primary Index is the weighted mean of macro scores:

$$PI_a^{\,headline} = (\Sigma_{k=0}^{N-1}\, w_k \cdot A_k)/(\Sigma_{k=0}^{N-1}\, w_k)$$

For a single-article author ($N = 1$) the formula degenerates gracefully: $p_0 = 0$ by convention, $w_0 = 1$, and the headline equals the article's macro score. The N=1 edge case matters because the eligibility floor (§2.5.1) allows authors to appear at five articles — a small denominator that benefits from a clean limit.

2.4.2 Properties of the Recency Weights

The weight function $w(p) = 0.5^{p/H}$ was chosen over the calendar-age decay used in earlier drafts ($w_k = 2^{-\tau_k / 18m}$) for three reasons:

1. **Archive-shape invariance.** Calendar-age decay penalises authors whose outlet went on extended hiatus; an author publishing a strong article in March 2024 and another in March 2026 would see the first weighted at $\approx 2^{-24/18} = 0.40$ regardless of how many articles came between them. Age-rank decay weights articles by their position in the

author's own timeline: the second-most-recent article is the second-most-recent article whether it was published last week or last year. The resulting headline tracks an author's *trajectory of work*, not the calendar.

2. **Bounded influence of the back catalogue.** Under calendar decay an exceptionally prolific author could accumulate hundreds of small-weight articles whose summed weight exceeded a peer's lighter back catalogue, distorting comparisons. Under age-rank decay every author's weights occupy the same *[0.0625, 1]* interval; the sum $\Sigma\ w_k$ scales linearly in N and the mean is invariant under unit scaling of total weight.
3. **Sensitivity to the next article.** A new article published by an author with N prior articles enters at weight 1 and demotes every older article's age rank by $1/N$. The marginal influence of a new article on the headline is therefore approximately $1/\Sigma_k\ w_{k'}$ which for a typical archive of 40–100 rated articles lands in the 3–6% range — enough that authors visibly move on each new piece, but not so much that a single bad article tanks an established career headline.

2.4.3 Computational Recipe

The headline is recomputed in full on each event that touches an author's article set: a new article being scored, or a content removal. Full recomputation is preferred over incremental update because the entire computation costs $O(N)$ for an author and the system stores the closed-form result, so there is no efficiency advantage to incremental arithmetic and a clear correctness advantage to the simpler code path.

Production pseudo-code:

``` request: author_id articles = db.articles_for(author_id).order_by('published_at desc') N = len(articles) if N == 0: return None if N == 1: return articles[0].macro_score num, den = 0.0, 0.0 for k, art in enumerate(articles): p = k / (N - 1) w = 0.5 ** (p / 0.25) num += w * art.macro_score den += w return num / den ```

The production implementation matches this pseudo-code exactly. There is no separate code path for new authors, single-article authors, or archive-only authors; the same function services live and seeded states.

2.4.4 Read-Only Article Scores

Once the rubric pass has run against a published article, the per-metric scores and the macro article score A_k are frozen. There is no post-publication re-score path, no uplift mechanism, and no author appeal that can alter an existing article's rating after the fact. The empiricism of the article *as published* is what is measured, and later actions by the author do not retroactively re-rate the artefact.

This is a deliberate design choice. The recency-weighted mean of §2.4.1 already provides bidirectional headline movement: a new strong article enters the mean at weight 1 and pulls the headline up, while a weak new article pulls it down. All rank movement is therefore generated by new work — the author's lever is to write the next article better than the last one. Correspondingly, the headline is recomputed only on the events listed in §2.4.3 (new
```

article scored, or content removal); there is no re-score event on the existing article set.

2.4.5 COLD-START SEEDING PROTOCOL — BACK-ARCHIVE INGESTION

The headline definition in §2.4 describes steady-state behaviour: at every event the headline is the recency-weighted mean of an author's current article set. At launch, however, no author has a published article history *in the system*, and the leaderboard must be initialised. A naive cold-start in which every author begins at PI = 750 and articles are streamed in real time would produce months of churn before the headline distribution settled, and would expose the system to ordering artefacts (the first three authors whose articles are scored would dominate the early leaderboard by virtue of attention, not craft). Restricting the seed to only the last 18 months would produce a defensible ordering but discard most of the available signal: an author with 200 rated career articles would be treated identically to one with 20.

We therefore run a **one-time deterministic seeding pass over each author's full available back catalog**. We collect approximately 500–1,000 candidate authors and ingest as much of each author's published archive as we can practically obtain — typically the entire span an outlet's archive exposes, often 10+ years per author. Every article in that archive is rubric-scored once and inserted into the same article store the live system reads from; the headline computed in §2.4.1 then operates over the union of archive + live articles without distinguishing between them. There is no separate seeding algorithm: the production headline function is the seeding function.

Inputs. For each candidate author a we collect the rubric-scored set $\{A_{a,k}\}$ across the entire available back catalog ($A \in [0,1000]$ per §2.2). Per-author corpus sizes typically range from a handful of articles (early-career or sparse archive) to several hundred (long-tenured staff writers). No upper bound is imposed on the seed corpus depth; we trust the age-rank decay defined in §2.4.1 to discount old content automatically.

Step 0 — Archive ingestion and de-duplication

For each target author we enumerate publication URLs from the outlet's author archive page, syndicated wire feeds where applicable, and any explicit RSS history. De-duplicate by canonical URL and by content hash (some outlets republish wire copy under multiple slugs). Each surviving article is rubric-scored by the same scoring LLM used in live operation. This produces a per-author archive $\mathcal{H}_a = \{A_{a,k}\}$ ordered by published_at descending.

Step 1 — Eligibility filter

Apply the leaderboard eligibility rule from §2.5.1: drop any author with fewer than 5 rated articles in the **trailing 18 months** of their archive. The eligibility floor is a *recent activity* test — it answers the question "is this person actively publishing?" — and is independent of the headline computation, which uses the full archive. Let N_e denote the surviving cohort size

(typically 60–85% of the candidate pool, depending on outlet activity patterns).

Step 2 — Compute the headline over the full archive

For each eligible author, compute the recency-weighted mean defined in §2.4.1, using the author's full archive ordered newest-first. The result is the author's seed headline:

$$PI_a^{headline,seed} = (\Sigma_{k=0}^{N_a - 1} w_k \cdot A_{a,k})/(\Sigma_{k=0}^{N_a-1} w_k),\ w_k = 0.5^{(k/(N_a-1))/0.25}$$

This is the same function called in live operation; the only difference is the size of the input set. An archive-rich author with 200 articles enters the seed with a tight, defensible headline grounded in a decade of rated work; an eligibility-minimum author with 5 articles enters with a noisier headline whose uncertainty is reflected via the confidence-interval display described in §2.7.

Step 3 — Cohort calibration and Author PI display

The raw headlines from Step 2 are then standardised across the eligible cohort and projected onto the *N(750, 80)* display band by the same monotone transformation used in live operation (§2.6.2). Let μ_h and σ_h be the cohort mean and standard deviation of $PI_a^{headline,seed}$; the displayed Author PI is:

$$PI_a^{author} = 750 + 80 \cdot (PI_a^{headline,seed} - \mu_h)/(\sigma_h)$$

clipped to *[300, 1000]*. The cohort calibration parameters μ_h, σ_h are fixed at launch and updated only on quarterly recalibration windows; the leaderboard sort key is the underlying headline $PI_a^{headline}$, not the displayed Author PI (see §2.6 on display vs sort).

Step 4 — Rank assignment

Sort eligible authors by $PI_a^{headline,seed}$ descending, breaking ties first by article count N_a (more articles = more diagnostic) and then by recency of the most recent rated article (more recently active wins). Assign integer rank ρ_a. The leaderboard then transitions to live updates the moment the first new article is rated, at which point the same headline function recomputes from the union archive+live store and the ranking re-sorts.

Seed parameter summary

PARAMETER	SYMBOL	VALUE
Age-rank half-life	H	0.25
Display target mean	μ_{target}	750
Display target sd	σ_{target}	80

PARAMETER	SYMBOL	VALUE
Display floor / ceiling	—	300 / 1000
Seed corpus depth	—	full back archive (no cutoff)
Eligibility floor	—	5 articles in trailing 18 months
Target cohort	—	500–1,000 candidate authors

Worked example

Veteran author with a 12-year archive of 184 rubric-scored articles, sorted newest-first. The recency-weighted mean across the full archive is $PI^{headline,seed} = 812$. The cohort statistics on the eligible $N_e = 847$ authors are $\mu_h = 730$, $\sigma_h = 92$. The displayed Author PI is:

$$PI^{author} = 750 + 80 \cdot (812 - 730)/(92) = 750 + 80 \cdot 0.89 \approx 821$$

The leaderboard launches with this author at a tight, defensible position grounded in a decade of rated work; the next rated article they publish enters the archive at age-rank 0, demotes every prior article's age-rank by $1/185$, and recomputes the headline. An 18-month-only seed would have used only the author's last ~ 25 articles — a thinner evidence base on which to build the launch ordering.

2.5 GLOBAL RANK

2.5.1 Eligibility

An author is eligible for the public leaderboard if and only if they have published at least 5 rated articles in the last 18 months:

$$Eligible_a = 1[N_a^{(18m)} \geq 5]$$

Below the threshold, the author has a headline Primary Index computed internally but does not appear on the public ordering. Their per-article scores and personal trend are still visible on their own profile.

2.5.2 Ranking

Eligible authors are ordered by the headline Primary Index defined in §2.4.1 — the recency-weighted mean of their rated articles — descending. The integer global rank is then:

$$\rho_a = 1 + |\{b \in Eligible : b \neq a, Rank_b > Rank_a\}|$$

where $Rank_a$ is the position obtained by sorting on:

1. **Headline Primary Index** $PI_a^{headline}$ — descending (primary key)
2. **Article count** N_a — descending (more articles = more diagnostic among ties)
3. **Recency of last rated article** — descending (more recently active ranks higher among further ties)

Causality. The v3.5 architecture is explicit that the headline is the cause of the rank, not the other way around. The headline is computed first — a deterministic function of an author's articles — and the leaderboard is the sorted projection of those headlines onto the integers. This matters for two reasons. First, the headline is a real-valued quantity that responds to rated craft on every event; rank is a discrete consequence whose value depends on the entire cohort. Second, the dual-display calibration in §2.6 derives the displayed Author PI from the headline (the sort key), not from rank, which would otherwise produce the awkward circularity of "the score that is shown is computed from the rank that is computed from the score". Earlier drafts that derived Author PI from rank percentile inverted this relationship; v3.5 corrects it.

The total number of eligible authors at any time is N_e. Ranks are recomputed continuously: every headline update triggers a partial reorder limited to the authors whose rank position is affected by the delta.

2.5.3 Beat-Filtered Views

Each article carries a *beat* label drawn from a controlled vocabulary (Politics, Science, Business, Sports, Arts, etc.) — produced by the scoring LLM as a derived label and verified by the article's outlet metadata where available. An author's *primary beat* is the modal beat across their last 18 months of rated articles.

Beat filters do **not** recompute the headline. They project the existing global ranking onto an author subset:

$$\rho_a^{beat} = 1 + |\{b \in Eligible \cap Beat : b \neq a, Rank_b > Rank_a\}|$$

The global rank remains the canonical signal; the beat rank is a contextual view. An author's profile displays both ("World #247 · Science #18").

2.6 DISPLAY CALIBRATION: DUAL PI

2.6.1 Two PI Numbers

Both numbers carry the same name — **Primary Index** — distinguished by context:

- **Author PI.** Headline number on the leaderboard and author profile. A monotone projection of the underlying headline $PI_a^{headline}$ onto the $N(750, 80)$ display band.
- **Article PI.** Per-article number displayed in the author's article list and on the article page. Derived from the macro article score A by a fixed absolute calibration (§2.6.3) that does not depend on the cohort.

This dual structure resolves the conflict between *cohort-relative* signals (which let readers compare authors meaningfully) and *absolute* signals (which let a reader interpret a single article without reference to the cohort).

Display vs sort — a v3.5 commitment. In the v3.5 architecture, the **sort key** for the global leaderboard is the underlying headline $PI_a^{headline}$ defined in §2.4.1. The **display key** — the number a reader sees on the leaderboard — is the calibrated Author PI defined in §2.6.2. These are equivalent up to a monotone transformation, so a reader who sorts the visible column gets the same ordering as the production sort. The separation matters because the calibration parameters (μ_h, σ_h) are quarterly-frozen, while the headline updates continuously; freezing the display calibration prevents micro-moves in the cohort distribution from producing visible Author PI churn that has nothing to do with any individual author's work. Display-layer affordances such as the article-level h-index (a Hirsch-like h where the author has at least h articles with Article PI $\geq h \cdot 10 + 600$) and the personal-best annotation p_{max} (highest single Article PI in the trailing 18 months) are **reader-facing badges**, never inputs to the sort.

2.6.2 Author PI Display Mapping

Let $PI_a^{headline}$ be the author's headline from §2.4.1, and let μ_h, σ_h be the cohort mean and standard deviation of headlines across the eligible set at the most recent quarterly calibration. The displayed Author PI is:

$$PI_a^{author} = \mu_{target} + \sigma_{target} \cdot (PI_a^{headline} - \mu_h)/(\sigma_h)$$

with $\mu_{target} = 750$ and $\sigma_{target} = 80$, clipped to $[300, 1000]$. The transformation is strictly monotone increasing in $PI_a^{headline}$, so display-sort matches production-sort. At launch (Phase 1) $\sigma_{target} = 80$; after 12 months (Phase 2) it linearly tightens to 50 to increase visual discrimination as the corpus matures.

The headline-to-Author-PI mapping replaces the rank-percentile inverse-normal mapping of earlier drafts. The change is a direct consequence of the causality fix in §2.5.2: derive the display from the sort key (the headline), not from the result of sorting (the rank). The new mapping preserves the property that the median displayed author sits at PI 750 and the top 3% sits above PI 900, while making the display layer a pure function of measurable rated work.

2.6.3 Article PI (Absolute)

A per-article PI uses the article score A directly, mapped onto the same display band by a fixed linear transformation calibrated against the rubric's nominal pass-fail thresholds ($A = 400$ maps to PI 500; $A = 700$ maps to PI 750; $A = 850$ maps to PI 850):

$$PI^{art} = 300 + 0.7 \cdot A$$

clipped to $[300, 1000]$. The article PI does **not** depend on cohort distribution — it is the same number for the same article regardless of who else has published. This is what makes per-article scores comparable across time and editorially defensible.

2.6.4 New Authors (Below Eligibility)

Authors with fewer than 5 rated articles in the trailing 18 months:

- Have no rank ρ on the public leaderboard
- Show no Author PI on their public profile (replaced by "Provisional — N articles rated, M to eligibility")
- Continue to receive Article PI scores for each rated piece
- Their headline still updates internally so they are ready to enter the leaderboard the moment they cross the threshold

This avoids the awkward case of a single article placing an author at rank 47 — an event with very little statistical content.

2.7 UNCERTAINTY AND THE CONFIDENCE SURFACE

Under the recency-weighted mean, uncertainty in the headline is the weighted-mean's own standard error, which depends only on the article-level variance and the effective sample size implied by the weights. Define the effective sample size as:

$$n_a^{eff} = ((\Sigma_k w_k)^2)/(\Sigma_k w_k^2)$$

the entropy-corrected count familiar from survey statistics. For an author whose archive is dominated by recent articles, n_a^{eff} is close to N_a; for an author whose work is spread across a long archive, it is smaller. With $H = 0.25$ the limiting effective sample size as $N_a \to \infty$ is approximately $N_a / 4$ — the back catalogue contributes diminishing information per article.

The 95% confidence interval on the headline is then:

$$PI_a^{headline} \pm 1.96 \cdot (\hat{\sigma}_A)/(\sqrt{(n_a^{eff})})$$

where $\hat{\sigma}_A$ is the weighted standard deviation of the author's article scores under the same w_k. The bounds are projected through the §2.6.2 display mapping to give the displayed CI on Author PI; the displayed format is **743 [±47]** — analogous to a measurement with uncertainty.

The CI also drives the eligibility-bubble safeguard described in §6.7: authors near the 5-article eligibility threshold display a wider CI and a "near-threshold" indicator. The transition from the eligibility-frontier CI to a stable mid-cohort CI is smooth, by construction — there is no rating-deviation discontinuity at the eligibility boundary as there was in the Glicko-2 formulation.

2.8 READ-ONLY ARTICLE SCORES AND THE NEW-WORK INCENTIVE

2.8.1 Design Principle

Once the rubric pass has run, an article's per-metric scores and macro score A_k are frozen. The system exposes no post-publication uplift path and no author appeal that alters an existing article's rating. The empiricism of the article *as published* is what is measured. This is a design commitment, not a limitation: it protects the artefact against being edited into a higher score after the fact and it forces the reward pathway back onto the primary craft loop of publishing better work.

2.8.2 How Rank Movement Is Generated

All headline and rank movement is generated by new articles entering the recency-weighted mean of §2.4.1. Each new article enters at weight 1 and demotes every prior article's age-rank position by *1/N*, so the marginal influence of the newest article on a typical mid-career headline is 3–6 percentage points of the current headline — enough to be visible, small enough not to lurch. Over time, weaker early-career pieces decay to negligible weight and the headline converges on the author's current craft level.

Bidirectionality is preserved automatically: a strong new article pulls the headline up, a weak one pulls it down. Authors who publish weak articles fall in rank, and authors who publish strong ones rise. There is no separate mechanism for score movement beyond the one that governs the headline itself.

2.8.3 Why This Is the Right Incentive Loop

The alternative — a post-publication evidence-uplift channel — was considered and rejected in v3.5 for three reasons:

1. **It confuses the artefact with the author.** A published article is a public record; retroactively raising its score creates a two-tier truth (what the article said at publication versus what it is now credited with) that undermines the audit trail.

2. **It rewards withholding.** Any post-publication uplift mechanism creates an incentive to publish incomplete pieces and drip-feed evidence for points later, even under ceilings designed to make honest filing dominant.
3. **It fragments the reward pathway.** The professional register of the Primary Index rests on a single clean claim: your rank tracks your recent work. Every additional lever muddies that message.

What remains is a system whose only incentive is to write the next article better than the last one. That is the intended equilibrium.

Part III: Engagement Loop — Credible-Professional Design

3.1 DESIGN PHILOSOPHY

The Primary Index engagement loop is grounded in Ryan and Deci's Self-Determination Theory (SDT, 2000). SDT identifies three innate psychological needs whose satisfaction produces sustained intrinsic motivation: *autonomy*, *competence*, and *relatedness*.

Extrinsic gamification mechanics — streaks, badge collections, leaderboards with variable reward schedules — satisfy none of these needs authentically. They trigger the *controlled motivation* pathway rather than the *identified regulation* pathway. Deci and Ryan's experimental literature (Deci, Koestner & Ryan, 1999, a meta-analysis of 128 studies) shows that tangible extrinsic rewards reliably undermine intrinsic motivation for activities that are already interesting.

The Primary Index leaderboard is therefore framed as a **professional ranking** in the register of the ATP tennis rankings, FIDE chess ratings, or the Bloomberg analyst rankings — not as a game leaderboard. The cadence is article-by-article (days to weeks), not daily. There are no streaks, no XP, no notifications about peers overtaking you, no "you're so close to rank X" prompts. The ranking simply exists, updates when new articles are rated, and is visible to anyone who looks.

This satisfies SDT's three needs:

- **Autonomy:** The author chooses which articles to write and whether to engage with the system at all. The system never demands action.
- **Competence:** Per-metric breakdowns on each article show exactly where craft improvements are possible on the next piece. The author can see their craft improving in concrete terms across successive articles.
- **Relatedness:** The global ranking and beat-filtered views connect the author to a peer community of practitioners. They can see who else is doing the work and how their own craft compares.

3.2 PUBLIC-FACING SURFACES

3.2.1 The Leaderboard — The headline surface. A live table: rank, author name, primary beat, author PI, 30-day rank trend (sparkline). Filters: global / by beat / by outlet. Search by author. The eligibility floor (5 rated articles in 18 months) is signposted in the table header.

3.2.2 The Author Profile — One number is featured: the author PI, with rank ("Author PI 812 · World #247 · Science #18"). Below: the 90-day rank trajectory, the per-metric radar across the last 20 articles, and the full article list with per-article PI scores in chronological order.

3.2.3 The Article List on the Profile — Each row shows the article headline, publication date, beat, and article PI. The per-metric radar makes visible which metrics carried the article and which held it back — signal for the author's next piece.

3.2.4 Per-Article Rubric View — For each article, the seven per-metric scores are shown alongside the LLM-produced rationale for each score. The purpose is craft feedback for future work; the published score itself is read-only (§2.4.4).

3.2.5 Trend View — A 12-month sparkline showing the author's rank trajectory. Hover reveals which newly rated article produced each step. Annotations like "Rank rose 34 after the Climate Damages investigation entered the mean at weight 1" make the cause-and-effect visible.

3.2.6 Award Tiers — The Primary Index recognises sustained craft through three reader-facing tiers, displayed as small badges adjacent to the author name on the leaderboard and profile. The tiers are determined by the displayed Author PI and percentile within the eligible cohort, recomputed at each quarterly calibration; movement between tiers is announced on the author's trend view.

TIER	TRIGGER	VISUAL
Indexed	Eligible (≥5 rated articles in trailing 18 months)	Baseline name badge
Silver Quill	Displayed Author PI > 750	Inline tier mark
Laureate	Author PI in the top percentile of the eligible cohort (precise percentile published with each calibration)	Inline tier mark with year

The tier system is intentionally short. Three steps mean a tier promotion is a real event — distinct enough to be earned, scarce enough to retain credibility, and small enough that the design avoids the streaks-and-badges register that the credible-professional framing of §3.1 explicitly excludes. Tiers carry no monetary or score-modifying consequence; they are reader-facing recognition only, and they do not affect the headline computation, the rank, or the displayed CI.

3.3 WHY THIS IS SUFFICIENT FOR A PROFESSIONAL AUDIENCE

Professional journalists, like academic researchers and ranked athletes, are motivated by the quality of their work and their standing among skilled practitioners — not by streak counters. The ATP tour publishes rankings publicly, updates them weekly, and the players engage seriously with their positions without anyone sending them push notifications about it. The Primary Index targets the same register: a credible professional ranking that participants take seriously precisely because it is not gamified.

Part IV: Game-Theoretic Analysis

4.1 INCENTIVE COMPATIBILITY GOAL

A scoring system is *incentive-compatible* if the strategy that maximises a participant's headline score is also the strategy that serves the system's stated goal — in this case, producing well-sourced, well-evidenced, honest journalism with verifiable primary evidence. We analyse equilibrium behaviour for four actor types under the proposed architecture.

4.2 THE HONEST JOURNALIST

Best response under the Primary Index: Produce well-sourced articles, signal opinion as opinion, acknowledge uncertainty, and continue to publish. The system's only lever is the next article.

Why this is the equilibrium: Every component of the system rewards these behaviours directly. The rubric rewards craft on each article; sustained craft produces a stable headline, a tight CI from the weighted standard error, and a high rank that does not lurch on any single article. Because article scores are read-only after the rubric pass, the honest journalist cannot be undercut by a competitor gaming a post-publication uplift channel — there is no such channel. The honest journalist faces no tension between rank-maximisation and craft-maximisation.

The dual-PI structure further protects the honest journalist: their per-article PI scores stand on their own absolute merit, independent of cohort movement. An honest journalist who happens to be rated against a cohort of exceptionally strong peers still receives high Article PIs; their author PI sits in the cohort context, but their article-level record is portable.

4.3 THE JOURNALIST TEMPTED TO GAME THE SYSTEM

Strategy 1: Topic concentration on easily scoreable beats. Publish only on topics where rubric scores are mechanically easy.

Countermeasure: Beat-filtered ranking removes most of the cross-beat advantage — within a beat, an easy-topic author competes with other easy-topic authors. The category-aware rubric metrics further reduce the asymmetry. A residual cross-beat issue remains for *global* rank, which is acknowledged in §6.

Strategy 2: High-frequency short articles. Saturate the headline with many low-effort pieces.

Countermeasure: The recency-weighted mean is not volume-rewarding — it is quality-weighted. Low-quality articles enter the mean at their full real-valued macro score, and the newest article carries weight 1, so a flood of weak pieces pulls the headline *down* immediately. The half-life schedule ($H = 0.25$) gives the newest articles disproportionate influence, so a strategic publication burst does not produce a permanent headline boost — the next high-quality article from any competitor will displace it from the top weight slot.

Strategy 3: Strategic timing of releases around career events. Hold a strong piece back to publish it just before a job application or award nomination.

Countermeasure: This is not a gaming attack on the scoring system — the work still happens and enters the headline at its full macro score under the standard §2.4.1 recency-weighted mean. Any advantage from timing is transient because the next article from any competitor will displace it from the top weight slot within days or weeks.

Strategy 4: Publishing pieces without full primary-source disclosure. Publish articles knowing they will score poorly on sourcing metrics.

Countermeasure: The rubric already penalises this at the article level: unsourced claims land in low bands for SO, EQ, and ET, dragging A_k down. Because article scores are read-only after publication, there is no post-publication uplift path to recover the lost points. The author's only route to a higher headline is to publish subsequent, better-sourced pieces — which is exactly the desired incentive.

4.4 THE BAD-FAITH PUBLISHER

Strategy 1: Author selection. Hire authors with high author PI.

Countermeasure: This is not a gaming strategy — it is the intended hiring signal at work. The system aligns the incentive.

Strategy 2: Pre-submission article coaching. Edit articles for surface compliance with the rubric without substantive sourcing.

Countermeasure: The scoring LLM evaluates substance. Pro-forma "according to a report by X" without primary linkage scores low on metrics SO, EQ, and ET, and the low article score is permanent — there is no post-publication path to recover it. Coaching that stops at rubric-surface compliance therefore produces a persistently weak headline for the coached author.

4.5 THE ADVERSARIAL AUTHOR

Under the Primary Index, no author's action can directly lower another's headline. The system has no head-to-head exchange and no peer-voting mechanism. This is structurally robust to competitive sabotage.

Indirect effects. If author B rises in rank, author A's rank falls by one position. This is the leaderboard moving as designed, not an attack — A's headline is unchanged; only their relative position has updated.

There is also no third-party or first-party evidence submission path: article scores are read-only after the rubric pass (§2.4.4, §2.8). This is intentional — see §6.5 for the design rationale and the disputes-only future-work proposal.

4.6 INCENTIVE COMPATIBILITY ASSESSMENT

The Primary Index is strongly incentive-compatible for honest journalists: rank-maximisation aligns with craft-maximisation on every article, and the read-only-after-publication commitment removes every post-hoc gaming channel that a re-score path would otherwise open. Adversarial attacks are structurally blocked — no author's action can lower a peer's score. The remaining open item is that third-party counter-evidence has no formal path; §6.5 addresses this as a disputes-only proposal that never mutates the published score.

Part V: Calibration Worked Example

5.1 SIMULATION SETUP

We generated a synthetic cohort of $N=500$ authors with the following generative model:

- **Latent skill** drawn from $N(730, 120^2)$, clipped to $[300, 1000]$.
- **Article count** per author drawn log-normal ($\mu_{log}=2.5$, $\sigma_{log}=0.8$), clipped to $[2, 120]$. Mean = 16.5 articles per author.
- **Article scores** drawn $N(latent\_skill, 60^2)$, clipped to $[0, 1000]$.
- **Article ordering** randomised within each author's archive (chronological order assigned by sampling publication timestamps uniformly over a 24-month window, then sorting). Age-rank decay weight $w_k = 0.5^{(k/(N_a-1))/0.25}$ per §2.4.1.
- **Eligibility filter** applied: authors with fewer than 5 articles in the trailing 18 months excluded from the leaderboard.

The recency-weighted-mean headline from §2.4.1 was computed for each author. The eligible subset was sorted descending by headline to produce ranks. Headlines were then projected through the cohort calibration of §2.6.2 to produce the displayed Author PI.

5.2 RESULTS

Eligible cohort statistics:

STATISTIC	VALUE
Eligible authors	**428** of 500 (85.6%)
Mean author PI	**750.4**
Std. dev. author PI	**78.6**
Min author PI	**502.3**
Max author PI	**997.4**
Median author PI	**749.8**
5th percentile	**646**
95th percentile	**881**
Authors below 500 PI	**0.0%**
Authors above 900 PI	**3.0%**
Mean CI width	**129 points**

The rank-based mapping produces the target *N(750, 80)* distribution by construction. The 72 ineligible authors (below the 5-article threshold) are excluded — they have ratings internally but no public rank.

Article PI distribution (absolute calibration, all rated articles across all authors):

STATISTIC	VALUE
Mean article PI	**815**
Std. dev. article PI	**75**
Articles below 500 PI	**0.8%**
Articles above 900 PI	**9.4%**

Article PIs are systematically higher than Author PIs because the article distribution reflects raw rubric output (latent skill + per-article noise) without the rank-percentile compression. This is correct behaviour — a single excellent article is rare, but an author whose lifetime distribution is excellent is rarer still.

5.3 CHART

The simulation dashboard contains six panels (available on request to kyle@theprimary.com):

- **Panel A** — Author PI distribution histogram overlaid with the target $N(750, 80)$ PDF.
- **Panel B** — Author PI vs. global rank ρ, confirming the monotone mapping.
- **Panel C** — CI width vs. rated-article count. Clear downward trend confirms volume sensitivity.
- **Panel D** — Phase 1 ($\sigma=80$) vs. Phase 2 ($\sigma=50$) target distributions, showing post-12-month tightening.
- **Panel E** — Three example rank trajectories: one rising author closing gaps; one steady; one falling.
- **Panel F** — Article PI vs. author PI scatter, showing that high-author-PI authors produce systematically higher Article PIs but with overlap (every author has weaker pieces).

5.4 KEY RECOMMENDED PARAMETERS

PARAMETER	RECOMMENDED VALUE	RATIONALE
Age-rank half-life H	**0.25**	Quarter-back of archive contributes half-weight; oldest article ~6.25%
Headline floor (Article PI band)	**300**	Lower bound of display band
Eligibility threshold	**5 articles in 18 months**	Avoids placing single-article authors at extreme ranks
Display target μ_{target}	**750**	Launch-friendly centre
Display target σ_{target} (Phase 1)	**80**	Wide enough to discriminate, narrow enough to avoid extremes
Display target σ_{target} (Phase 2)	**50**	Increased discrimination after 12 months
Article PI slope	**0.7**	Maps $A=700$ to PI 750, $A=400$ to PI 580
Article PI intercept	**300**	Lower bound of display band
Cohort recalibration cadence	**quarterly**	Freezes display layer between recalibration windows

Part VI: Open Questions and Future Work

6.1 CROSS-BEAT COMPARABILITY IN GLOBAL RANK

Beat-filtered views are the primary defence against unfair cross-beat comparison, but the **global** rank still mixes a science journalist citing peer-reviewed papers with a political correspondent relying on anonymous-but-verified sources. The category-aware rubric relaxation addresses the most obvious asymmetry; a residual issue remains for the global headline.

Proposed approach. Estimate beat-specific average rubric performance from the first 12–18 months of data, then apply a beat-specific offset to the headline before global ranking (e.g., science journalists' headlines compressed by 15 points to centre the beat distribution on the global median). This preserves the single global ranking while making it fairer across beats.

6.2 ELIGIBILITY-THRESHOLD BUBBLE

Authors who oscillate around the 5-article-in-18-months threshold appear and disappear from the leaderboard, which can produce confusing user experience: an established journalist on parental leave drops off, then re-appears with a rank that may not reflect their craft level relative to peers who continued publishing.

Proposed safeguard. A 3-article grace decay: once an author has been eligible, they remain on the leaderboard with a "low-activity" flag for as long as their trailing 18-month article count is between 3 and 5. Their headline continues to update through the recency-weighted mean during the low-activity window, the effective sample size shrinks because no new high-weight articles enter, and the displayed CI widens accordingly.

6.3 NON-ENGLISH CONTENT

The scoring LLM operates in English. Localised rubric variants for French, Spanish, Arabic, and Mandarin are on the 18-month roadmap. The headline and rank machinery are language-agnostic; only the scoring LLM requires localisation.

Open question. Should non-English authors share the global ranking (with localised scoring) or have a separate leaderboard? The current proposal is shared global, with the beat filter doing double duty as a language filter via a derived language label.

6.4 OUTLET LOCK-IN

If outlet-level PI becomes credible, employers may preferentially hire from high-PI outlets, creating rich-get-richer dynamics. The current design displays outlet PI only on outlet profile pages — outlets are not on the headline leaderboard. The author PI is portable across outlets.

6.5 THIRD-PARTY DISPUTES ON PUBLISHED ARTICLES

Because article scores are read-only after the rubric pass (§2.4.4, §2.8), the system has no mechanism by which counter-evidence from credentialed researchers, public-interest

organisations, or rival journalists can alter a published score. This is intentional — the article-as-published is the artefact being measured — but it leaves a residual gap where credible external material that contradicts a piece cannot be visibly registered.

Proposed future module. A disputes-only surface modelled on the F1000Research post-publication commenting layer: third-party submissions create a *dispute* on the article, the original author has a fixed response window, and unresolved disputes are flagged on the article record. Critically, disputes never alter the article's numeric score — they annotate it. This preserves the read-only commitment while giving the community a public channel to register substantive challenges. Legal and moderation implications (defamation, tortious interference, coordinated brigading) would need to be worked through before launch.

6.6 LONG-RUN DISTRIBUTION STABILITY

The cohort calibration projects the headline distribution onto *N(750, 80)* by construction at every quarterly recalibration, but the underlying headline distribution evolves between calibrations. If the eligible cohort becomes increasingly selective (only high-craft journalists remain rated), calibration will compress genuine differences among an elite population.

Proposed approach. A biannual review by an independent calibration committee, with changes announced 90 days in advance and applied with a 6-month linear transition. The Phase 1 → Phase 2 □ tightening (80 → 50 over months 12–24) is the first such transition and is built in from launch.

6.7 HONEST LIMITATIONS

This white paper proposes a technically rigorous system, but honesty requires acknowledging fundamental limitations:

1. **The rubric is a proxy for empirical craft, not for truth.** A well-sourced article making false claims still scores well. The Primary Index measures craft, not truth.
2. **Scoring LLM consistency is not guaranteed.** A single LLM produces some inter-run variance on subjective metrics. Temperature is held at zero, the system prompt and rubric grounding are version-pinned, and identical inputs are batched to minimise but not eliminate drift. Periodic recalibration runs against a fixed benchmark set monitor drift.
3. **The system has no theory of importance.** A well-sourced piece on a trivial topic scores the same as a well-sourced piece on a story of major public interest.
4. **Cultural and political context.** "Balance" and "Right of Reply" are culturally loaded. Naive rubric application could penalise a well-sourced article on the scientific consensus that does not give equal time to denialism — this is handled in the scoring LLM's system prompt but remains an ongoing calibration concern.
5. **Rank movement can be misread.** A drop of 12 places on the leaderboard may be statistical noise (headline change within the CI), not a craft change. The CI on author PI

must be displayed prominently to discourage over-interpretation of short-window movements.

These limitations do not undermine the system's utility but define the domain of appropriate application. The Primary Index is a credible signal about *empirical craft as evidenced by primary sources*, not a comprehensive verdict on journalism quality.

Appendix A: Notation Summary

SYMBOL	DESCRIPTION
A_k	Macro article score for article k, $\in [0, 1000]$
m_i	Per-metric score from the scoring LLM for metric i, $\in [0, 1000]$
w_i	Rubric weight for metric i; $\Sigma\, w_i = 1$
SO, EQ, ET, BU, FO, PL, SB	v3.5 metric codes (see §2.2)
N, n_a	Number of rated articles in the author's archive
k	Article index, $k = 0$ newest, $k = N\text{-}1$ oldest
p_k	Age-rank position $k/(N\text{-}1)$, $\in [0, 1]$
H	Recency half-life on the age-rank axis (default 0.25)
w_k	Recency weight $0.5^{p_k / H}$ for article k
$PI_a^{headline}$	Author's recency-weighted-mean headline (sort key)
PI_a^{auth}	Displayed Author Primary Index (calibrated headline)
PI_k^{art}	Displayed Article Primary Index (calibrated A_k)
$\hat{\sigma}_A$	Weighted standard deviation of an author's article scores
n_a^{eff}	Entropy-corrected effective sample size for author a
ρ_a	Global rank of author a
N_e	Number of eligible authors
μ_h, σ_h	Cohort mean and standard deviation of headlines at quarterly calibration

SYMBOL	DESCRIPTION
$\mu_{target}, \sigma_{target}$	Target distribution parameters for displayed Author PI
Φ^{-1}	Inverse standard normal CDF

Note: Symbols from prior systems referenced in Part I (r, RD, σ, μ, φ, τ, s) appear only in the Glicko-2, TrueSkill, and Elo expositions of §1.1–1.3 and are not used by the v3.5 system.

Appendix B: Per-Metric Band-to-/1000 Mapping

The scoring LLM scores each metric on a common 0–1000 axis. The rubric bands defined in the source rubric are interpreted as proportional positions on each metric's native ceiling and mapped onto *[0, 1000]* by linear rescale:

$$m_i = 1000 \cdot (b_i)/(B_i)$$

where b_i is the band-midpoint or interpolated band score the scoring LLM assigns within the metric's native range, and B_i is the metric's native ceiling. All per-metric scores share a common 0–1000 axis after mapping, so the v3.5 weight vector can be applied as a direct convex combination.

Band reference table (v3.5 rubric):

CODE	METRIC	V3.5 WEIGHT	"STRONG" BAND /1000
SO	Use of Sources & Attribution	0.20	889–1000
EQ	Evidence Quality & Relevance	0.20	889–1000
ET	Evidence Traceability	0.20	889–1000
BU	Balance & Acknowledgement of Uncertainty	0.1333	880–1000
FO	Fact/Opinion Signalling	0.0889	880–1000
PL	Polemical Language Control	0.0889	907–1000
SB	Source Balance & Right of Reply	0.0889	880–1000

Worked example. A feature piece receives the following per-metric scores on the common 0–1000 axis:

- SO = 850, EQ = 780, ET = 920, BU = 810, FO = 740, PL = 880, SB = 700

Macro article score under the v3.5 weight vector:

$$A = 0.20(850) + 0.20(780) + 0.20(920) + 0.1333(810) + 0.0889(740) + 0.0889(880) + 0.0889(700) \approx 818$$

Displayed Article PI (Article-axis calibration, §2.6):

$$PI^{art} = 300 + 0.7 \cdot 818 \approx 873$$

Under v3.5 the article enters the author's headline at its full real-valued macro $A_k = 818$, weighted by its age-rank position in the author's archive — not as a binary outcome.

Appendix C: Primary Sources and URLs

All citations in this paper trace to primary sources:

- Elo, A.E. (1978). *The Rating of Chessplayers, Past and Present*. Arco. Full text (Gwern archive)
- Glickman, M.E. (2022). *Example of the Glicko-2 System*. glicko.net
- Herbrich, R., Minka, T. & Graepel, T. (2007). TrueSkill™: A Bayesian Skill Rating System. *NeurIPS 2006*. NeurIPS proceedings
- Minka, T., Cleven, R. & Zaykov, Y. (2018). TrueSkill 2: An improved Bayesian skill rating system. *Microsoft Research Technical Report MSR-TR-2018-8*. microsoft.com
- Bradley, R.A. & Terry, M.E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39:324. Semantic Scholar
- Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci.* 102(46):16569–16572. arXiv / PNAS
- FICO Score methodology. myfico.com
- Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW7*. Google Research
- Kamvar, S.D., Schlosser, M.T. & Garcia-Molina, H. (2003). The EigenTrust Algorithm for Reputation Management in P2P Networks. *WWW '03*. Stanford NLP / ACM DL
- Adler, B.T. (2012). WikiTrust: Content-Driven Reputation for the Wikipedia (PhD thesis). UC Santa Cruz. eScholarship
- Stack Overflow reputation system. stackoverflow.com/help/whats-reputation

- F1000Research open peer review. think.f1000research.com
- Ryan, R.M. & Deci, E.L. (2000). Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist* 55(1):68–78. SDT website / PubMed

Appendix D: Changelog — v3.5 Architecture

The v3.5 architecture represents a deliberate simplification of the Bayesian rating dynamics of earlier drafts. This changelog summarises substantive changes by section.

Identity and branding

- All references to the prior company name and ranking system updated: the company is now **Primary Labs**, and the ranking system is the **Primary Index** (PI).

Part I — Literature review (§1.1–1.5)

- Verdicts updated to reflect v3.5: Elo, Glicko-2, and TrueSkill are retained as design ancestors but no longer as production update machinery. Glicko-2's contribution is acknowledged as the conceptual foundation that v3.5 intentionally simplifies away in favour of a transparent recency-weighted mean.
- The h-index entry is updated to record its v3.5 role: a reader-facing display badge on author profiles, not a sort key.

Part II — Proposed methodology (Part II)

- §2.1 System Overview: restructured as four layers plus a seeding protocol. The Bayesian rating layer is removed; the headline layer is the recency-weighted mean.
- §2.2 Rubric weight vector locked to v3.5 (SO 0.20, EQ 0.20, ET 0.20, BU 0.1333, FO 0.0889, PL 0.0889, SB 0.0889); the three derived labels (article type, 7-point political leaning, anonymous-source usage) are documented as non-scoring metadata.
- §2.3 Article-to-headline mapping: the binary outcome $s \in \{0, 0.5, 1\}$ and the virtual-opponent framing are removed; each article enters the headline at its full real-valued macro score A_k.
- §2.4 Author headline: full rewrite as a recency-weighted mean with age-rank weights $w_k = 0.5^{p_k / H}$, default $H = 0.25$. Subsections cover the definition, properties of the weights, the computational recipe with pseudo-code, and the read-only commitment on published article scores.
- §2.4.4 Read-only article scores: new subsection formalising the v3.5 commitment that once the rubric pass has run, an article's per-metric scores and macro score are frozen. All rank movement is generated by new work entering the recency-weighted mean, not by post-publication re-scores.

- §2.4.5 Cold-Start Seeding: rewritten to use the same recency-weighted mean over the full back-archive. The Glicko-scale mapping ($\kappa = 200$) and logarithmic RD schedule are removed.
- §2.5 Global rank: causality clarification added. The headline is computed first; the leaderboard is the sorted projection. Earlier drafts that derived the displayed Author PI from rank percentile inverted this relationship.
- §2.6 Display calibration: new "display vs sort" commitment paragraph. The sort key is the headline; the display features the h-index badge and a personal-best annotation p_{max} for context, not as part of the ordering.
- §2.7 Uncertainty: rewritten around the weighted standard error $\hat{\sigma}_A / \sqrt{(n_a^{eff})}$, with an entropy-corrected effective sample size. The Glicko-2 RD discontinuity at the eligibility boundary is eliminated by construction.
- §2.8 Rewritten as "Read-Only Article Scores and the New-Work Incentive": the earlier gap-fill mechanic is removed; the section now states the read-only commitment and explains why the recency-weighted mean is the sole rank-movement pathway.

Part III — Engagement loop (§3.2)

- New §3.2.6 Award Tiers: a three-step reader-facing recognition layer — Indexed (baseline), Silver Quill (PI > 750), Laureate (top cohort percentile). Tiers carry no score-modifying consequence and do not constitute streaks or badges in the gamified sense.

Part V — Calibration (§5.1, §5.4)

- Simulation rewritten around age-rank decay rather than calendar-time decay. The Glicko-2 parameters (τ, RD floor) are removed from the parameter table; $H = 0.25$ and the cohort recalibration cadence are added.

Appendix A

- Notation table rebuilt around v3.5 symbols. Glicko-2 and TrueSkill symbols (r, RD, σ, μ, φ, τ, s) are retained only as references for Part I and explicitly flagged as not used by the v3.5 system.

Appendix B

- Worked example reissued under the v3.5 weight vector. Per-metric scores are labelled with the v3.5 codes; the binary outcome line is removed.

Evidence layer removed

- The former §2.8 Gap-Fill Mechanic and Appendix D Raw Evidence Bonus are removed in v3.5. Both mechanisms provided a post-publication uplift path against a published article's score, which conflicted with the audit-trail commitment of the F1000Research reference model (§1.12) and created a residual incentive to publish incomplete pieces and drip-feed evidence for points later.
- The article-level score is now read-only after the rubric pass. All headline and rank movement is generated by new articles entering the recency-weighted mean of §2.4.1.

See §2.4.4 and the rewritten §2.8 for the v3.5 statement.

- Notation: the gap-fill increment symbol $\Delta_{i,k}$ is retired.
- Parameter table: the "Max gap-fills in flight" and "Raw Evidence Bonus ceiling" rows are removed from §5.4.

Primary Labs Primary Index White Paper · © 2026 Primary Labs · All rights reserved.

The simulation code and calibration chart referenced in Part V are available on request to kyle@theprimary.com.