Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quality-Guarded Multi-Attribute Foundation Model Recommendation: A Scoring Pipeline with Pareto Stratification

Surbhi Agrawal

surbhiagrwal.rvitm@rvei.edu.in

RV Institute Of Technology & Mangement

Divya Chouksey

RV Institute Of Technology & Mangement

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11472403/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Quality-Guarded Multi-Attribute Foundation Model Recommendation: A Scoring Pipeline with Pareto Stratification

Surbhi Agrawal[1*] and Divya Chouksey[1]

[1*]Department of CSE (AI & ML), RV Institute of Technology & Management, Bangalore, India.

*Corresponding author(s). E-mail(s): surbhiagrwal.rvitm@rvei.edu.in;
Contributing authors: divychouksey136@gmail.com;

Abstract

Selecting a foundation model requires balancing quality, cost, latency, context, privacy, and availability, yet public leaderboards rank on a single dimension and linear weighted sums let cheap or fast models offset inadequate quality. We present a recommendation pipeline combining hard feasibility constraints, Pareto stratification, constant-elasticity-of-substitution (CES) aggregation, soft-knee percentile normalization, and a continuous quality-deficit penalty in place of threshold cliffs. We prove monotonicity and, under stated factor bounds, a quality-conditional score ceiling and properties of a bounded display mapping; we also show that Pareto dominance over the efficiency triplet implies score order only under a stated condition, with a counterexample outside it. On a frozen 47-model Agent Arena cohort, regret is measured against the leaderboard's net-improvement outcome, which does not price cost or latency. Under capability-first intent, the pipeline has the lowest point-estimate mean regret among the original unpenalized integrative baselines (0.09 versus 0.49), although paired intervals include zero. Corrected ablation leaves regret unchanged when the deficit factor is removed, but it rises to 1.28 when the frontier reward is removed (paired difference interval 0 to 2.49). Adding the same deficit factor and frontier reward to WSM also yields 0.09; adding the deficit factor alone to WSM or TOPSIS leaves regret at 0.49. Replacing CES with a linear aggregator leaves mean regret unchanged and also retains top-three overlap of 2/3. Under budget and balanced intents the pipeline trails several baselines; these are limitations because the outcome cannot adjudicate cost-sensitive preferences. A secondary common-utility analysis applies shared cost caps and an explicit utility to all methods,

using weights that have not been validated against user choices. The results show outcome alignment on one cohort, not user-intent satisfaction.

Keywords: Foundation models, quality-deficit penalty, Constant elasticity of substitution, Multi-attribute utility theory, Pareto optimality, recommendation systems, Model selection

1 Introduction

The operational landscape of generative artificial intelligence is characterized by an expanding frontier of foundation models with divergent architectural specializations, parameter scales, deployment modalities, and cost structures [5, 16]. Enterprise architects and system developers face a high-dimensional multi-criteria decision dilemma: selecting an optimal model is no longer a matter of querying the top of a general-purpose leaderboard. A production workflow may mandate sub-second response times for real-time customer interaction, multi-hundred-kilotoken context windows for complex codebases, air-gapped on-premises privacy, or aggressive unit economics for high-throughput batch extraction [25].

Despite the multidimensional nature of foundation model selection, existing evaluation tools and routing frameworks suffer from four critical structural deficiencies:

1. **Single-Metric Myopia:** Public leaderboards typically rank models along a single aggregate dimension, such as Chatbot Arena Elo or standardized benchmark suites (e.g., MMLU, HumanEval, SWE-bench) [7, 12, 15]. These aggregate rankings ignore operational dimensions such as throughput (tokens per second), latency to first token, dollar cost per million tokens, context limits, and licensing compliance.
2. **Compensatory Collapse in Linear Additive Utility:** When multi-criteria ranking is attempted, existing tools almost universally employ naive weighted linear arithmetic sums: $U = \sum w_i x_i$. Under linear aggregation, attributes are treated as perfect substitutes. A model with unacceptable task proficiency can outscore an elite frontier model simply because its low cost or rapid generation inflates its composite score. In mission-critical workflows, task quality is non-compensatory: exceptional affordability cannot compensate for defective reasoning.
3. **Cliff Effects in Discontinuous Filtering:** While hard constraint filtering is essential, applying crude step penalties (e.g., penalizing a score by a flat 40% if quality drops below an arbitrary threshold)

creates severe instability. Models hovering marginally above or below the cutoff experience discontinuous ranking inversions.

4. **Cohort Normalization Collapse and Tie Clumping:** Standard min-max normalization maps the empirical maximum to 1.0, leaving scores acutely vulnerable to outlier distortion. Conversely, clipping values at a fixed upper percentile P_{95} collapses all top-tier models onto identical 1.0 normalized scores, destroying discriminatory power among elite alternatives. Similarly, hard display caps (e.g., $\min(99, S)$) generate artificial plateaus among top-scoring candidates.

To overcome these structural limitations, this paper develops a principled, multi-attribute decision-analytic scoring methodology for foundation model recommendation. We draw theoretical inspiration from the *Cobb–Douglas production function* [10] and *Constant Elasticity of Substitution (CES)* utility models from microeconomic theory [1], integrating them with modern non-compensatory decision theory [24] and multi-objective Pareto optimization.

Our architecture ingests live catalog metadata from the OpenRouter API [20], open-weight repository telemetry from the Hugging Face API [13], and independent capability benchmarks from the Artificial Analysis API [2].

Rather than treating attributes as linearly substitutable, the method aggregates them with a constant-elasticity-of-substitution (CES) power mean at $\rho = 0.5$ ($\sigma = 2.0$). This setting penalizes dimensional imbalance relative to a linear sum while remaining more substitutable than the Cobb–Douglas limit ($\sigma = 1$). Protection against inadequate quality comes from a smooth deficit penalty and a pre-scoring constraint filter, not from the CES setting. The score couples utility multiplicatively with category fit, source trust, a quadratic confidence discount, and a specialist bonus, then maps it to a strictly bounded, non-saturating asymptotic scale.

Contributions. This paper makes six contributions:

1. An outlier-robust soft-knee percentile normalization that preserves strict rank order from the fifth cohort percentile upward and avoids the upper-tail tie-saturation of a hard P_{95} clip. It is continuous but has slope changes at P_5 and P_{95}; this regularity limitation is stated explicitly.
2. A pipeline separating non-compensatory hard-constraint filtering, Pareto stratification over the quality–affordability–speed triplet, and multi-attribute scalar scoring. Pareto membership is explanatory

metadata; a counterexample shows that triplet dominance and score order can disagree when other coordinates carry weight.

3. A CES weighted-power-mean formulation (U_ρ, default $\rho = 0.5$) with a floored Cobb–Douglas endpoint ($\varepsilon = 0.05$), combined multiplicatively with category fit, source trust, and a quadratic confidence discount.
4. A continuous quality-deficit penalty, C^1 for the frontier archetype and C^0 for balanced, budget, and speed archetypes; a convex frontier reward; and a C^1 asymptotic display ceiling that bounds scores to $[0, 99.4)$ while preserving rank monotonicity.
5. Formal results (Propositions 1–5), each with its scope condition. Proposition 2 is conditional: Pareto dominance implies weak score order only when every positively weighted coordinate lies in the efficiency triplet; a strict score gain is guaranteed when a strictly improved triplet coordinate has positive weight. Section 4.6.1 gives a counterexample outside the condition.
6. A frozen-cohort evaluation against Agent Arena's cost-blind headline outcome. Under capability-first intent, the pipeline has mean regret 0.09 versus 0.49 for the original unpenalized integrative baselines, although paired intervals include zero; WSM with both matched quality factors also attains 0.09. Corrected component ablation distinguishes the deficit factor from the frontier reward, and matched WSM/TOPSIS variants test the shared quality factor. Budget and balanced results are limitations because the outcome proxy cannot adjudicate cost-sensitive preferences. A shared-utility analysis reports cost, quality, utility regret, and Pareto dominance under common constraints; whole-catalogue coverage addresses the stratum unavailable to outcome-based panels.

Framing Note. The core aggregator is a weighted power mean, i.e., a member of the CES family. Cobb–Douglas represents its $\rho \to 0$ limiting member, while the multiplicative coupling with category fit and trust factors is Cobb–Douglas in functional form. We evaluate both the Cobb–Douglas endpoint (with regularization floor $\varepsilon = 0.05$, since the pure unfloored product vanishes when any metric is missing) and the default $\rho = 0.5$. Crucially, the non-compensatory protection is established by the smooth deficit penalty (Proposition 3), whereas absolute exclusions are enforced strictly by the pre-scoring constraint filter.

Evidential scope. Three classes of claim have different status. The formal properties (Propositions 1–5) are proved under their stated conditions and describe the method, not any cohort. The capability-first

regret differences are point estimates whose paired intervals include zero. The cost-sensitive Arena results are negative: under the budget prompt the pipeline has greater mean regret than WSM, TOPSIS, Cobb–Douglas, and CES; under the balanced prompt it has greater regret than WSM, floored WSM, TOPSIS, and CES. Because the outcome omits cost and latency, these panels do not establish user-intent satisfaction. The separate shared-utility analysis reports cost-sensitive outcomes under explicit weights, but those weights are not validated against user choices. Section 7.2 collects these boundaries.

2 Related Work and Theoretical Foundations

2.1 Multi-Attribute Utility and Non-Compensatory Decision Rules

Multi-Criteria Decision Making (MCDM) and Multi-Attribute Utility Theory (MAUT), formalized by Keeney and Raiffa [17], provide the axiomatic foundation for evaluating alternatives characterized by conflicting objectives. Standard additive utility models rely on the assumption of mutual preferential independence:

$$U(\mathbf{x}) = \sum_{k=1}^{K} w_k u_k(x_k), \quad \sum_{k=1}^{K} w_k = 1. \tag{1}$$

When individual utilities $u_k(x_k)$ are linear, the Marginal Rate of Substitution (MRS) is constant:

$$\mathrm{MRS}_{i,j} = -\frac{\partial U/\partial x_i}{\partial U/\partial x_j} = -\frac{w_i}{w_j}. \tag{2}$$

This implies that a deficit in one attribute can be offset by an equivalent linear surplus in another, regardless of how deficient attribute i becomes. An additive rule therefore cannot itself express a requirement to reject candidates below an essential threshold; such a requirement must be represented as a hard constraint or an explicit deficit penalty. Tversky's elimination-by-aspects model is related non-compensatory choice theory, but it does not establish the particular penalty function used here [24]. Alternative MCDM methods such as AHP [23], TOPSIS [14],

PROMETHEE [6], ELECTRE [22], and VIKOR [21] address outranking and compromise solutions. Below, we benchmark our method against linear sums, hard-floor linear sums, TOPSIS, and Cobb–Douglas.

2.2 Cobb–Douglas and Constant Elasticity of Substitution (CES)

In production economics, the Cobb–Douglas function models output as a multiplicative interaction of factor inputs [10]:

$$Y = A \prod_{k=1}^{K} x_k^{\alpha_k}, \quad \sum_{k=1}^{K} \alpha_k = 1. \tag{3}$$

The Cobb–Douglas formulation possesses an elasticity of substitution $\sigma = 1.0$. It inherently penalizes dimensional imbalances: as any input $x_k \to 0$, output $Y \to 0$. However, applying a pure multiplicative Cobb–Douglas product directly to foundation model selection presents a critical practical challenge: *sparse metric annihilation*. In live model registries, optional secondary telemetry attributes (such as specialized agent evaluations or third-party context benchmarks) may be unavailable for newly released models. A pure multiplicative product would annihilate the entire utility score of an elite model if a single attribute is missing.

To resolve this limitation while retaining balance preference, Arrow, Chenery, Minhas, and Solow introduced the Constant Elasticity of Substitution (CES) generalized mean [1]:

$$U_\rho(\mathbf{x}) = \left(\sum_{k=1}^{K} \bar{w}_k x_k^\rho \right)^{1/\rho}, \quad \bar{w}_k = \frac{w_k}{\sum_j w_j}, \quad \sigma = \frac{1}{1-\rho}, \tag{4}$$

where $\rho \in (-\infty, 1]$ governs the elasticity of substitution [11]:

- As $\rho \to 1$, U_ρ converges to the linear weighted arithmetic mean ($\sigma \to \infty$).
- As $\rho \to 0$, U_ρ converges asymptotically to the Cobb–Douglas geometric mean $\prod x_k^{\bar{w}_k}$ ($\sigma = 1$).
- As $\rho \to -\infty$, U_ρ converges to the Leontief minimum function $\min_k\{x_k\}$ ($\sigma = 0$).

For $\rho \in (0, 1)$, the CES aggregator is concave and penalizes dimensional imbalances relative to a weighted arithmetic mean. Unlike the unfloored

Cobb–Douglas formulation, it remains well-defined when an attribute is zero and retains positive utility provided at least one positively weighted attribute is nonzero. By adopting $\rho = 0.5$ (the square-root power mean, $\sigma = 2.0$) as our default, we obtain a concave aggregator that permits greater substitutability than Cobb–Douglas ($\sigma = 1$), but less than the linear weighted mean. Stronger protection against sub-threshold task quality is provided by the separate continuous quality-deficit penalty, while absolute feasibility requirements are enforced through pre-scoring hard constraint filtering.

2.3 Foundation Model Telemetry, Leaderboards, and Dynamic Routers

The foundation model ecosystem encompasses empirical benchmark suites, crowdsourced arenas, and automated inference gateways. MMLU [12], GSM8K [9], HumanEval [7], and SWE-bench [15] quantify domain-specific execution fidelity. Crowdsourced arenas such as LMSYS Chatbot Arena [25] and Agent Arena evaluate relative human and agent preference, and their per-model records are distributed as dated datasets so that a cohort can be reanalysed independently [4]. Model routing architectures such as FrugalGPT [8] and RouteLLM [19] train classifier heads to cascade simple queries to smaller models and difficult queries to frontier APIs. However, dynamic routers do not solve multidimensional procurement under user-specified operational constraints.

To provide reliable inputs, our framework ingests structured metadata and benchmarks from three primary online sources:

- **OpenRouter API** [20]: Provides dynamic model catalog metadata, endpoint availability, context window limits, supported parameters (tool calling, structured JSON output), and token pricing ($P^{\text{in}}, P^{\text{out}}$).
- **Hugging Face API** [13]: Supplies open-weight model architectures, repository download statistics, community likes, and open-source license validations.
- **Artificial Analysis API** [2]: Delivers independent, verified performance measurements, including generalized intelligence indices, coding benchmark indices, agentic tool scores, and empirical generation throughput (tokens per second).

3 Problem Formulation

Let $\mathcal{M} = \{m_1, m_2, \ldots, m_N\}$ denote a universe of candidate foundation models. Each model $m \in \mathcal{M}$ is characterized by an architectural attribute profile:

$$\mathbf{a}_m = \langle \mathrm{id}_m, \mathrm{name}_m, \mathrm{provider}_m, \mathcal{C}_m, \mathcal{T}_m, \mathrm{ctx}_m, P_m^{\mathrm{in}}, P_m^{\mathrm{out}}, \mathrm{Th}_m, \mathcal{E}_m \rangle, \quad (5)$$

where $\mathcal{C}_m$ represents the primary category (e.g., general, image, video, voice, music, document), $\mathcal{T}_m$ is the set of supported modalities and tools, $\mathrm{ctx}_m \in \mathbb{N}^+$ is the context window length in tokens, $P_m^{\mathrm{in}}, P_m^{\mathrm{out}} \in \mathbb{R}_+$ denote input and output token pricing per million tokens, $\mathrm{Th}_m \in \mathbb{R}_+$ is the empirical throughput in tokens per second, and $\mathcal{E}_m$ represents verified benchmark telemetry.

A recommendation request is defined by a tuple:

$$\mathcal{R} = \langle \mathcal{C}^*, \mathbf{w}, \Omega, \theta \rangle, \quad (6)$$

where $\mathcal{C}^*$ is the target task category, $\mathbf{w} \in \mathbb{R}_+^K$ is a priority weight vector across the six evaluation dimensions (quality, affordability, speed, context, privacy, availability), Ω is a set of hard operational constraints, and $\theta \in \{\mathrm{frontier}, \mathrm{balanced}, \mathrm{budget}, \mathrm{speed}\}$ designates the operational archetype.

Natural-language prompts are parsed by a hosted instruction-tuned model (`command-r7b-12-2024`) that returns the intent as structured data: the target category, an ordered list of prioritized evaluation dimensions, numerical limits (budget ceilings, minimum context), and provider or model exclusions. The parsed intent selects one of four operational archetypes (frontier, balanced, budget, speed), and each archetype maps to a fixed weight profile over the six dimensions; the ordered priority list determines the weight vector when no archetype keyword is present. The objective is to produce an admissible candidate subset $\mathcal{M}^* \subseteq \mathcal{M}$ and a scalar ranking score $S(m) \in [0, 99.4)$ such that:

$$S(\pi(1)) \geq S(\pi(2)) \geq \cdots \geq S(\pi(|\mathcal{M}^*|)). \quad (7)$$

4 The Decision-Analytic Scoring Methodology

4.1 Overview of the Proposed Recommendation Pipeline

Figure 1 illustrates the architecture of the proposed quality-guarded multi-attribute foundation model recommendation pipeline. The process

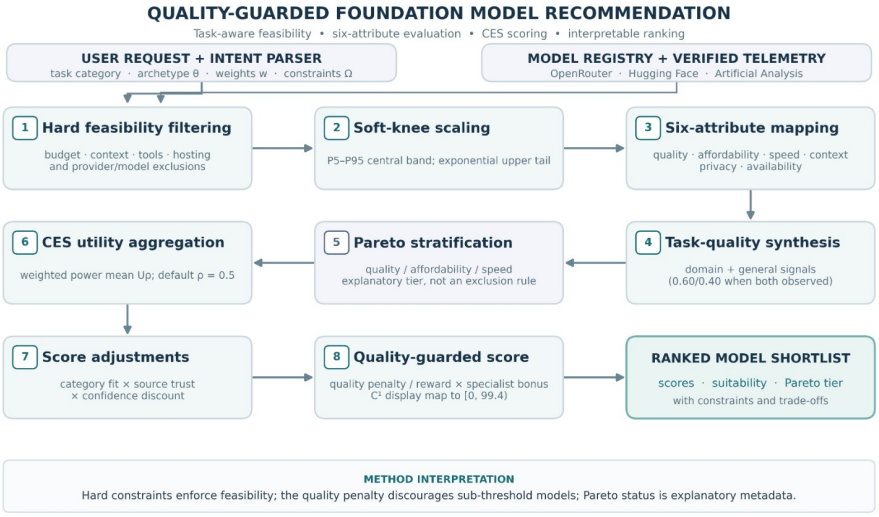

Fig. 1 Architecture of the proposed quality-guarded multi-attribute foundation model recommendation pipeline. Natural-language intent parsing identifies task priorities and operational constraints, while an integrated model registry supplies candidate attributes. Hard feasibility filtering, soft-knee normalization, and task-quality synthesis precede Pareto stratification over quality, affordability, and inference speed. The six-dimensional CES utility is combined with multiplicative scoring adjustments, archetype-dependent quality factors, and an asymptotic display ceiling to produce ranked recommendations. Pareto membership is retained as explanatory metadata and does not directly determine score order.

begins with a natural-language user request, which is interpreted to identify the target task category, operational constraints, and attribute priorities. The intent-parsing component determines the corresponding operational archetype—frontier, balanced, budget, or speed—and generates a weight vector for six evaluation dimensions: task quality, affordability, inference speed, context capacity, privacy, and availability.

Candidate models are retrieved from the integrated model registry, which incorporates metadata and benchmark observations from OpenRouter, Hugging Face, and Artificial Analysis. Hard feasibility constraints are applied to exclude candidates that violate explicitly specified requirements. The remaining models are evaluated using normalized telemetry attributes, where soft-knee percentile normalization reduces sensitivity to outliers while preserving discrimination among high-performing models. Task-specific quality is synthesized from available domain-specific and general capability measurements.

Pareto stratification identifies non-dominated models using the quality–affordability–speed efficiency triplet. Importantly, Pareto membership serves as explanatory metadata rather than as a mandatory ranking or elimination criterion. Subsequently, the six normalized attributes are aggregated through a weighted constant-elasticity-of-substitution (CES) utility function, with the default parameter set to $\rho = 0.5$. The resulting utility is adjusted using category fit, source trust, confidence discounting, and task-specific specialization factors.

A continuous quality-deficit penalty discourages models with insufficient task proficiency, while the frontier archetype additionally incorporates a quality-dependent reward. Finally, a continuously differentiable asymptotic ceiling maps the adjusted score to a bounded scale below 99.4 without introducing upper-end ranking ties. The pipeline produces an ordered list of feasible models accompanied by their evaluation scores, attribute information, and Pareto membership, thereby supporting transparent, task-aware model selection.

The recommendation methodology operates as an eight-stage pipeline:

1. Hard Operational Constraint Filtering ($\mathcal{M} \to \mathcal{M}^*$).
2. Outlier-Robust Soft-Knee Percentile Cohort Normalization.
3. Dimensional Telemetry Mapping.
4. Dual-Signal Task Quality Synthesis.
5. Multi-Objective Pareto Frontier Stratification.
6. Concave CES Power-Mean Utility Aggregation (U_ρ).
7. Cobb–Douglas Multiplicative Coupling and Confidence Discounting.
8. Continuous Deficit Penalties, Frontier Rewards, and C^1 Asymptotic Display Ceiling.

4.2 Hard Operational Constraint Filtering

Prior to scalar scoring, candidates must satisfy the non-compensatory constraint set Ω. The admissible space $\mathcal{M}^*$ is defined as:

$$\mathcal{M}^* = \{m \in \mathcal{M} \mid \forall c_j \in \Omega : \phi(m, c_j) = 1\}, \tag{8}$$

where the indicator predicate $\phi(m, c_j)$ enforces satisfaction across:

- **Context Capacity:** $\mathrm{ctx}_m \geq \mathrm{ctx}_{\min}$.
- **Budget Ceiling:** If a price ceiling is specified it is applied to the input price alone, $P_m^{\mathrm{in}} \leq P_{\max}^{\mathrm{in}}$, because the ceiling is stated per million input tokens and the input price is the term present for every priced model. A model whose input price is unlisted is retained: the predicate

removes only models *known* to breach the ceiling, absence of a price not being evidence of a high one. Two consequences are carried into the evaluation below: a workload dominated by output tokens can admit a model whose blended cost exceeds the stated ceiling, and an unpriced model is never removed by this predicate.

- **Governance and Self-Hosting:** Mandates for on-premises or air-gapped deployment strictly admit only open-source or open-weight models ($\text{access}_m \in \{\text{Open source}, \text{open-weights}\}$).
- **Modality and Tool Support:** Requires verified parameter support for function calling or structured outputs ($\mathcal{T}_{\text{req}} \subseteq \mathcal{T}_m$).
- **Provider/Model Exclusions:** $\text{provider}_m \notin \mathcal{P}_{\text{exclude}} \wedge \text{id}_m \notin \mathcal{M}_{\text{exclude}}$.

Constraint filtering represents the only absolute non-compensatory mechanism in the system.

4.3 Outlier-Robust Soft-Knee Percentile Normalization

Let $v_{m,k}$ denote the raw observed telemetry value of model m for metric dimension k (e.g., benchmark indices from the Artificial Analysis API, tokens per second, or context lengths). Standard min-max scaling is vulnerable to single-sample extremes, and hard-clipping at a fixed upper percentile P_{95} collapses all top-tier models onto an identical 1.0 score, causing *tie-saturation* among precisely the candidates a procurement decision must discriminate. We therefore scale the central cohort linearly against its own percentile band and attach a soft exponential tail above the upper knee, giving the *soft-knee percentile normalization* $\mathcal{N}(v; P_5, P_{95})$:

$$\mathcal{N}(v) = \begin{cases} \kappa \left(\dfrac{v - P_5}{r} \right), & P_5 \le v \le P_{95}, \\ \kappa + (1 - \kappa) \left(1 - \exp\left(-\dfrac{v - P_{95}}{\tau_{\text{tail}} r} \right) \right), & v > P_{95}, \end{cases} \tag{9}$$

with $\mathcal{N}(v) = 0$ for $v < P_5$, where $r = P_{95} - P_5$ is the empirical 90% inter-percentile range, $\kappa = 0.92$ sets the knee at P_{95}, and $\tau_{\text{tail}} = 0.30$ is the tail scaling parameter (summarized in Table 1).

This formulation possesses four key mathematical properties:

1. **Outlier Resistance:** The central 90% of the distribution is scaled linearly against cohort percentiles rather than single-sample extremes.

Table 1 Parameters for soft-knee cohort normalization.

Parameter	Operational Value and Design Rationale
κ (Knee)	0.92 — Value at P_{95}; reserves 8% headroom for elite frontier differentiation.
τ_{tail} (Tail Scale)	0.30 — Tail reaches $1 - e^{-1} \approx 63.2\%$ of the remaining headroom at distance $\tau_{\text{tail}} r$ above P_{95}.
P_5 (Saturation Point)	5th cohort percentile; $\mathcal{N}$ saturates at 0 below it, since the lower tail is dominated by failed or near-zero evaluations.
P_{95} (Knee Point)	95th cohort percentile; the linear limb spans $[P_5, P_{95}]$.

2. **Strict Rank-Order Monotonicity on $[P_5, \infty)$:** The normalization function $N(v)$ is continuous and strictly increasing for $v \geq P_5$. Its derivative is strictly positive on the open intervals (P_5, P_{95}) and (P_{95}, ∞), but the function is not differentiable at the breakpoints P_5 and P_{95}. Consequently, for any $v_1 > v_2 \geq P_5$, we have $N(v_1) > N(v_2)$. In particular, the exponential upper-tail mapping assigns distinct normalized values to models above P_{95}, thereby avoiding tie saturation introduced by hard percentile clipping. Below P_5, normalization saturates at zero by design. Candidates in this region may receive a quality-deficit penalty during scoring, as described in Section 4.9, but are not automatically excluded unless they violate an explicit hard feasibility constraint.
3. **Cohort Invariance under Filter Shifts:** Percentiles P_5, P_{95} are computed once over the entire ingested catalog rather than post-filter candidate subsets, guaranteeing that a candidate's normalized capability does not change merely because another model was filtered.
4. **Continuity with slope changes at the saturation point and knee:** $\mathcal{N}$ is continuous but not differentiable at P_5, where the constant zero limb meets the positive-slope central limb, and at P_{95}, where its slope jumps from κ/r on the linear limb to $(1-\kappa)/(\tau_{\text{tail}} r) = 0.267/r$ on the exponential tail, an attenuation of roughly 71%. It is therefore C^0 but not C^1 at either join. The upper-knee slope change is deliberate—a softer tail would flatten the elite band the knee exists to separate—and is disclosed here because Section 1 criticises step penalties. The frontier quality factor is continuously differentiable (C^1), whereas the balanced, budget, and speed quality factors are continuous (C^0) but not differentiable at their quality threshold. The asymptotic display ceiling is C^1, and the coding-generalist quality adjustment is also C^1 with respect to quality following the smoothstep revision described in Section 4.9.

4.4 Dimensional Telemetry Formulations

For each model $m \in \mathcal{M}^*$, the system evaluates six standardized coordinates $\mathbf{x}_m = [x_{m,\text{qual}}, x_{m,\text{aff}}, x_{m,\text{spd}}, x_{m,\text{ctx}}, x_{m,\text{priv}}, x_{m,\text{avail}}]^T \in [0,1]^6$:

4.4.1 Affordability ($x_{m,\text{aff}}$)

Pricing is evaluated using a category-sensitive blended token cost. General work is weighted toward output tokens, whereas document and retrieval workloads are input-heavy:

$$P_m^{\text{blend}} = \begin{cases} 0.60P_m^{\text{in}} + 0.40P_m^{\text{out}}, & \text{general work,} \\ 0.85P_m^{\text{in}} + 0.15P_m^{\text{out}}, & \text{document work.} \end{cases} \tag{10}$$

Because token costs span several orders of magnitude (from zero for open-source models to over \$50/M tokens for reasoning APIs), affordability is mapped logarithmically:

$$x_{m,\text{aff}} = \text{clamp}\left(1.0 - \frac{\log_{10}(P_m^{\text{blend}} + 1.0)}{2.1},\ 0.45,\ 0.98\right). \tag{11}$$

For a positive blended price, the map is clamped to $[0.45, 0.98]$; the upper bound is approached at the free-price end, and the lower bound binds from roughly \$13/M tokens upward. A non-positive or unavailable blended price is treated as unmeasured rather than as a measured free service. Newly imported open-weight models with unknown self-hosting infrastructure costs receive the explicit heuristic $x_{m,\text{aff}} = 0.95$; for other records, a registry rebuild retains any previously stored affordability value, and a value still missing when scoring occurs receives the scorer's general missing-metric floor of 0.08. These are distinct fallback paths: 0.95 is an optimistic unknown-infrastructure-cost assumption, whereas 0.08 is a conservative generic missing-value floor. In the frozen 47-model cohort, all 47 have positive blended prices and use the price map; zero use the 0.95 open-weight fallback, and zero rely on a retained or missing-value fallback. Therefore, none of these unmeasured-price paths affects the budget panel.

4.4.2 Context Capacity ($x_{m,\text{ctx}}$)

Context capacity spans from 1K to 10M+ tokens. The map is a linear-fractional (rational) function of $\log_{10}(\text{ctx})$ rather than a logistic one; its

purpose is to compress the range while reserving resolution for the long-window regime:

$$x_{m,\text{ctx}} = \text{clamp}\left(1.55\,\frac{\log_{10}(\text{ctx}_m) - 2.0}{\log_{10}(\text{ctx}_m) + 1.0},\ 0.08,\ 1.0\right), \tag{12}$$

yielding $x_{\text{ctx}} \approx 0.54$ at 4K, 0.71 at 32K, 0.79 at 128K, and 0.89 at 1M tokens.

4.4.3 Inference Speed ($x_{m,\text{spd}}$)

Empirical output throughput Th_m (tokens per second) from the Artificial Analysis API is calibrated via:

$$x_{m,\text{spd}} = \text{clamp}\left(0.25 + \frac{\text{Th}_m}{300} \cdot 0.73,\ 0.25,\ 0.98\right), \tag{13}$$

yielding ≈ 0.35 at 40 t/s, ≈ 0.54 at 120 t/s, and 0.98 at ≥ 300 t/s.

4.4.4 Privacy and Availability ($x_{m,\text{priv}}, x_{m,\text{avail}}$)

Privacy is a three-level classification of the access model, $x_{m,\text{priv}} \in \{1.0, 0.75, 0.45\}$ for self-hostable open weights, zero-data-retention hosted endpoints, and standard cloud APIs, respectively. Availability is a provider-level aggregate of catalogue presence, endpoint redundancy and SLA history, with every catalogue-listed provider scored 0.90 in the revision this paper evaluates. Neither coordinate varies across the evaluation cohort (priv = 0.45, avail = 0.90 for all 47 models), so neither can move a ranking in any panel reported here, and both are recorded as classification rather than measurement for that reason; Section 6.11 reports the span that makes this auditable.

4.5 Dual-Signal Task Quality Synthesis

In specialized workloads such as coding or agentic execution, relying solely on domain syntax pass rates yields incomplete evaluations. Real-world engineering requires both domain mastery and broad algorithmic reasoning.

Let c_{norm} denote normalized domain execution capability (e.g., HumanEval, SWE-bench) and i_{norm} denote normalized general cognitive intelligence (MMLU, Chatbot Arena Elo). The combined capability index

is:

$$\tilde{c}_m = \begin{cases} 0.60\, c_{\text{norm}} + 0.40\, i_{\text{norm}}, & \text{if } c_{\text{norm}} \neq \emptyset \wedge i_{\text{norm}} \neq \emptyset, \\ c_{\text{norm}}, & \text{if } i_{\text{norm}} = \emptyset, \\ i_{\text{norm}}, & \text{if } c_{\text{norm}} = \emptyset. \end{cases} \tag{14}$$

The composite capability is then mapped into task quality via a concave curve:

$$x_{m,\text{qual}} = \text{clamp}\left(0.34 + 0.65\sqrt{\tilde{c}_m},\ 0.40,\ d\right), \quad d = \begin{cases} 0.99, & \text{coding capability}, \\ 0.98, & \text{overall quality}. \end{cases} \tag{15}$$

The square-root transformation amplifies capability gains at lower thresholds while requiring substantial advances across both reasoning and code execution to approach the upper bound.

4.6 Multi-Objective Pareto Frontier Stratification

Prior to scalar utility aggregation, candidates are evaluated for multi-objective Pareto optimality over their core efficiency triplet $\mathbf{y}_m = [x_{m,\text{qual}}, x_{m,\text{aff}}, x_{m,\text{spd}}]^T$:

Definition 1 (Pareto Dominance)**.** *Model m_a dominates model m_b ($m_a \succ_{\mathcal{P}} m_b$) if and only if:*

$$\forall j \in \{qual, aff, spd\},\ y_{m_a,j} \geq y_{m_b,j} \quad \wedge \quad \exists k \in \{qual, aff, spd\},\ y_{m_a,k} > y_{m_b,k}. \tag{16}$$

The candidate set is partitioned into two disjoint subsets:

$$\mathcal{P}_{\text{frontier}} = \{m \in \mathcal{M}^* \mid \nexists m' \in \mathcal{M}^* : m' \succ_{\mathcal{P}} m\}, \tag{17}$$

$$\mathcal{P}_{\text{eligible}} = \mathcal{M}^* \setminus \mathcal{P}_{\text{frontier}}. \tag{18}$$

The partition is computed with the standard maxima-of-a-set procedure [18]. Models in $\mathcal{P}_{\text{frontier}}$ are non-dominated alternatives that cannot be improved in quality, affordability, or speed without sacrificing another dimension. This Pareto tier is surfaced as explanatory metadata, so a user can audit which candidates are non-dominated in the efficiency triplet. It is *not* a guarantee that frontier membership determines rank, and the difference is not cosmetic wherever a coordinate outside the triplet carries weight.

4.6.1 The Pareto tier and the score rank in different spaces: a scope condition on the dominance proposition

Definition 1 is stated over the triplet $\mathbf{y} = [x_{\text{qual}}, x_{\text{aff}}, x_{\text{spd}}]$, whereas the score $S(m)$ of Section 5 aggregates all six coordinates. Proposition 2 bridges the two only under the hypothesis that every coordinate with positive weight lies in the triplet. The shipped frontier and budget profiles both violate that hypothesis by assigning context weight 0.20 and 0.65, respectively; the balanced profile also assigns context weight 0.65. Outside the proposition's stated restriction, a model dominated on the triplet can still be preferred because of a positively weighted coordinate outside it.

A concrete case, using the square-root power mean of Eq. 21 with normalised weights 0.60 on quality, 0.20 on affordability and 0.20 on context. Let m_a have $(x_{\text{qual}}, x_{\text{aff}}, x_{\text{spd}}, x_{\text{ctx}}) = (0.90, 0.50, 0.50, 0.30)$ and m_b have $(0.85, 0.45, 0.45, 0.90)$, with all other coordinates equal. Since each of the three triplet coordinates is strictly larger for m_a and no coordinate is smaller, $m_a \succ_{\mathcal{P}} m_b$ by Definition 1. The utilities are $U(m_a) = (0.60\sqrt{0.90} + 0.20\sqrt{0.50} + 0.20\sqrt{0.30})^2 = 0.673$ and $U(m_b) = (0.60\sqrt{0.85} + 0.20\sqrt{0.45} + 0.20\sqrt{0.90})^2 = 0.769$, so m_b outranks m_a despite being triplet-dominated. This does not contradict Proposition 1, which requires coordinate-wise dominance over the coordinates that carry weight and therefore does not apply when one model is better on context and worse on quality. It does mean that the two properties are separate ones: the concatenation of the Pareto filter with the score is not itself a dominance-consistent order, and citing frontier membership as a justification for a rank would be wrong. Frontier status is consequently reported throughout as metadata.

4.7 Cobb–Douglas/CES Utility Aggregation

Given normalized weights $\bar{w}_k = w_k / \sum_j w_j$ and attribute coordinates $x_{m,k} \in [0, 1]$, the multi-attribute utility is evaluated using the CES family:

$$U_\rho(\mathbf{x}_m) = \left(\sum_{k=1}^{K} \bar{w}_k x_{m,k}^{\rho} \right)^{1/\rho}, \quad 0 < \rho \leq 1, \tag{19}$$

$$U_0(\mathbf{x}_m) = \prod_{k=1}^{K} \max(x_{m,k}, \varepsilon)^{\bar{w}_k}, \quad \varepsilon = 0.05. \tag{20}$$

Our default configuration employs $\rho = 0.5$, which simplifies to the square-root power mean:

$$U_{0.5}(\mathbf{x}_m) = \left(\sum_{k=1}^{K} \bar{w}_k \sqrt{x_{m,k}} \right)^2 . \tag{21}$$

By the generalized power-mean inequality [11], for $0 < \rho \leq 1$, the regularized Cobb–Douglas utility, the CES utility, and the weighted arithmetic mean satisfy the following ordering whenever $x_{m,k} \geq \varepsilon$ for every attribute with positive weight $\bar{w}_k > 0$:

$$U_0(\mathbf{x}_m) \leq U_\rho(\mathbf{x}_m) \leq U_1(\mathbf{x}_m) = \sum_{k=1}^{K} \bar{w}_k x_{m,k}. \tag{22}$$

Under this condition, the regularization floor in U_0 is inactive, so the standard power-mean inequality applies. For $0 < \rho < 1$, equality throughout holds if and only if all positively weighted attribute values are equal. The operational attribute mappings in Section 4.3 satisfy this condition because their minimum mapped value is 0.08, exceeding $\varepsilon = 0.05$. For counterfactual values below ε, the inequality between the floored Cobb–Douglas endpoint and the unfloored CES utility is not guaranteed.

For the default choice $\rho = 0.5$, the elasticity of substitution is $\sigma = 1/(1-\rho) = 2.0$. The resulting CES aggregator permits more substitution than the Cobb–Douglas limit ($\sigma = 1$), but penalizes attribute imbalances relative to the weighted arithmetic mean. Unlike the unfloored Cobb–Douglas product, it remains well-defined when an attribute is zero.

4.8 Multiplicative Coupling and Confidence Discounting

Base utility is scaled into an initial raw score via multiplicative Cobb–Douglas coupling:

$$S_m^{\text{base}} = 100 \cdot U_m \cdot F_m^{\alpha} \cdot \gamma(C_m) \cdot T_m, \tag{23}$$

where:

- **Category Fit (F_m):** Measures alignment with the query category $\mathcal{C}^*$. A primary-use-case match gives $F_m = 1.00$ and a secondary match gives 0.88; failing both, the category pairs give 1.00 for an exact category match, 0.72 for a general model applied to document work, 0.58 for a document or voice model applied to general work, and 0.16 for any

other mismatch. With elasticity parameter $\alpha = 1.2$, an incompatible model is attenuated by $0.16^{1.2} \approx 0.11$.

- **Source Trust (T_m):** A multiplicative provenance factor, $T_m = \text{clamp}(a_m \cdot r_m \cdot l_m, 0.45, 1.0)$, where a_m is source authority (1.00 first-party, 0.99 benchmark, 0.98 aggregator or curated, 0.95 seed, 0.85 heuristic), r_m is record type (1.00 API model, 0.98 open-weight model, 0.95 repository or model-family record) and l_m is link health (1.00 verified, 0.98 catalogue, 0.97 unverified, 0.50 broken). An unclassified record defaults to 0.98 on each factor, and a hidden or non-deployable record type is scored $T_m = 0$ and removed from contention. The clamp floors the factor at 0.45, so provenance can move a score by at most a factor of 2.2.
- **Confidence Discount ($\gamma(C_m)$):** Metadata verification maturity $C_m \in [0, 1]$ is discounted quadratically:

$$\gamma(C_m) = 1.0 - \lambda(1.0 - C_m)^2, \quad \lambda = 0.20. \tag{24}$$

  This ensures $\gamma'(1.0) = 0$ (no variance penalty for fully verified models), while discounting unverified models by up to 20% ($\gamma(0) = 0.80$).

4.9 Continuous Quality Deficit Penalties and Frontier Rewards

To reduce compensation for a quality shortfall without introducing step-function cliff discontinuities, the frontier archetype uses a continuous rational deficit penalty:

$$\psi_{\text{frontier}}(q) = \frac{1}{1.0 + \lambda_f (q_f - q)_+^p}, \tag{25}$$

where $(q_f, \lambda_f, p) = (0.92, 3.0, 1.2)$. Balanced, budget, and speed use the shared logarithmic deficit factor $1/[1 + 1.8 \ln(1 + 3.5(0.70 - q)_+)]$, which is gentler for mild shortfalls.

For frontier queries, models meeting or exceeding the threshold ($q \geq 0.92$) receive a convex super-frontier ceiling reward:

$$\beta(q) = 1.0 + 0.25 \left(\frac{q - 0.92}{0.08} \right)^{1.5}. \tag{26}$$

Define the effective quality factor for all four archetypes $\theta \in \{\text{frontier}, \text{balanced}, \text{budget}, \text{speed}\}$. For frontier, $g_{\text{frontier}}(q) = \psi_{\text{frontier}}(q)$

when $q < 0.92$ and $g_{\text{frontier}}(q) = \beta(q)$ when $q \geq 0.92$. For balanced, budget, and speed, $g_\theta(q) = 1/[1+1.8\ln(1+3.5(0.70-q)_+)]$ for every q, which equals 1 above the shared threshold. For frontier, $p = 1.2 > 1$ makes the left derivative zero at the threshold, and the right derivative of β is also zero, so g_{frontier} is C^1. For the other three archetypes, the logarithmic limb meets the constant 1 continuously with finite one-sided slope, so g_θ is C^0 and strictly increasing in quality below 0.70. The implemented factor is therefore defined for all four values of θ.

For coding queries in the general category, three specialisation terms stack additively: a coding-specialist weight (+4%), chain-of-thought reasoning (+3%) and agentic tool use (+2%), so $B_m \leq 1.09$. A generalist that is neither a code specialist nor a reasoning model receives a smooth quality factor. Define $t(q) = \text{clip}((q-0.95)/0.05, 0, 1)$ and $h(t) = 3t^2 - 2t^3$; then $B_m = 0.92 + 0.08h(t(q))$. This factor is 0.92 for $q \leq 0.95$, reaches 1 at $q = 1$, and has zero derivative at both endpoints. The adjusted score is $S_m^{\text{adj}} = S_m^{\text{base}} \cdot g_\theta(x_{m,\text{qual}}) \cdot B_m$.

Continuity of the coding quality factor. The cubic smoothstep removes the former 8% jump at $q = 0.95$: the generalist coding factor is continuously differentiable with respect to quality at both $q = 0.95$ and $q = 1$. This statement concerns the quality-dependent factor; specialist and reasoning bonuses remain discrete metadata terms. All empirical tables were regenerated with this revised factor.

4.10 Continuously Differentiable (C^1) Asymptotic Display Ceiling

When high-quality models accumulate specialist bonuses and frontier rewards, composite scores can exceed 100. Applying a hard clip $\min(99, S)$ causes catastrophic score saturation, collapsing multiple elite models onto identical integer scores. We implement a smooth asymptotic exponential soft-ceiling:

$$\mathcal{C}(S) = \begin{cases} S, & S \leq s_0, \\ s_0 + H\left(1 - \exp\left(-\dfrac{S - s_0}{M}\right)\right), & S > s_0, \end{cases} \tag{27}$$

with knee threshold $s_0 = 86.0$, asymptotic headroom $H = 99.4 - 86.0 = 13.4$, and smoothing scale $M = 13.4$. Setting $M = H$ guarantees that the

derivative is exactly 1.0 on both sides of the knee:

$$\lim_{S\to 86^-} \mathcal{C}'(S) = 1.0, \qquad \lim_{S\to 86^+} \mathcal{C}'(S) = \frac{H}{M} e^0 = 1.0. \tag{28}$$

Final recommendation scores lie strictly within $[0, 99.4)$, preserving the natural linear calibration below 86.0 while guaranteeing strict rank monotonicity across the elite frontier.

5 Theoretical Properties

Throughout this section, the full recommendation score is:

$$S(m) = \mathcal{C}\Big(100 \cdot U_\rho(\mathbf{x}_m) \cdot F_m^\alpha \cdot \gamma(C_m) \cdot T_m \cdot g_\theta(x_{m,\text{qual}}) \cdot B_m\Big). \tag{29}$$

Proposition 1 (Monotonicity, Ceteris Paribus)**.** *Let models m_a, m_b share identical category fit F, confidence C, trust T, specialist bonus B, and archetype θ. If $\mathbf{x}_{m_a} \geq \mathbf{x}_{m_b}$ coordinate-wise with $x_{m_a,k} > x_{m_b,k}$ for some attribute with priority weight $w_k > 0$, then $S(m_a) > S(m_b)$.*

Proof. For $\rho \in (0, 1]$, the mapping $t \mapsto t^\rho$ is strictly increasing on $\mathbb{R}_+$. Let k be the coordinate named in the hypothesis, so that $w_k > 0$ and hence $\bar{w}_k = w_k / \sum_j w_j > 0$, with $x_{m_a,k} > x_{m_b,k} \geq 0$. Then $\bar{w}_k x_{m_a,k}^\rho > \bar{w}_k x_{m_b,k}^\rho$, while for every $j \neq k$ the term is weakly larger, $\bar{w}_j x_{m_a,j}^\rho \geq \bar{w}_j x_{m_b,j}^\rho$, because $x_{m_a,j} \geq x_{m_b,j}$ and weights are non-negative. Summing, $\sum_j \bar{w}_j x_{m_a,j}^\rho > \sum_j \bar{w}_j x_{m_b,j}^\rho$, and raising to the strictly increasing power $1/\rho$ preserves the strict inequality. The argument uses only the existence of *one* positively weighted coordinate that is strictly larger; coordinates with $w_j = 0$ contribute nothing to either side and the parsed vectors, which set some weights to zero, therefore satisfy the hypothesis. For $\rho = 0$ the same step applies to the floored product provided $x_{m_b,k} \geq \varepsilon$; every clamp in Section 4.4 returns at least 0.08, so this holds for every coordinate the pipeline maps. It does not hold for a coordinate set to zero by hand, which the pipeline never emits: the stress test of Section 6.3 does exactly that, and is labelled there as a counterfactual bound rather than as a state the method produces. The common factor $100F^\alpha\gamma(C)TB > 0$ preserves ordering. Furthermore, since $x_{m_a,\text{qual}} \geq x_{m_b,\text{qual}}$ and $g_\theta(q) > 0$ is nondecreasing, Ug_θ is strictly larger for m_a. Because the display ceiling $\mathcal{C}(S)$ has a strictly positive first derivative $\mathcal{C}'(S) > 0$ everywhere, $S(m_a) > S(m_b)$. □

Proposition 2 (Dominance Consistency)**.** *If all attributes with positive weights $w_k > 0$ belong to the efficiency triplet $\mathbf{y} = [x_{qual}, x_{aff}, x_{spd}]^T$ and $m_a \succ_{\mathcal{P}} m_b$ with equal F, C, T, B, then $S(m_a) \geq S(m_b)$. The inequality is strict when at least one strictly improved triplet coordinate has positive weight. Equality is possible when all positively weighted coordinates tie and the only strict gains are on zero-weight coordinates that do not change the quality factor (for example, an affordability-only gain when quality is unchanged).*

Proof. By Definition 1, $m_a \succ_{\mathcal{P}} m_b$ implies $\mathbf{y}_{m_a} \geq \mathbf{y}_{m_b}$ coordinate-wise. All positively weighted coordinates lie in this triplet, so the score cannot decrease, by the weak form of coordinate-wise monotonicity. If at least one strict gain is on a positively weighted coordinate, Proposition 1 gives a strict score increase. If all positively weighted coordinates tie, equality follows when the quality factor also ties; a strict gain on a zero-weight quality coordinate can still increase the score through g_θ, so it is not an equality case. □

Proposition 3 (Quality-Conditional Score Ceiling)**.** *Suppose $F_m, \gamma(C_m), T_m, B_m \leq 1.0$ and task quality $x_{m,qual} = q$. Then, regardless of the values of the other coordinates, the recommendation score is bounded by:*

$$S(m) \leq 100 g_\theta(q). \tag{30}$$

Proof. Because $x_{m,k} \leq 1.0$ for all k, $U_\rho(\mathbf{x}_m) \leq 1.0$. Multiplicative factors satisfy $F^\alpha \gamma T B \leq 1.0$. Furthermore, $\mathcal{C}(s) \leq s$ since $H(1 - e^{-z/H}) \leq z$. Thus, $S(m) \leq S^{\text{base}} g_\theta(q) \leq 100 g_\theta(q)$, completing the proof. □

For a frontier query with $q_f = 0.92, \lambda_f = 3.0, p = 1.2$, a model with quality $q = 0.89$ is capped at 95.7, and a model with $q = 0.71$ is capped at 68.4, even if its other coordinates are perfect (1.0). These examples instantiate the frontier penalty; g_θ gives the corresponding bound for balanced, budget, and speed. The proposition assumes $B_m \leq 1$ and therefore does not cover the coding-specialist bonus case (B_m can reach 1.09); the display ceiling still bounds the final score below 99.4. Proposition 1 separately establishes ordering under a common B.

Proposition 4 (Non-Compensation of the Aggregator)**.** *Let attribute k have normalized weight $\bar{w}_k$ and $x_k = 0$. For $\rho \in (0, 1)$, $U_\rho \leq (1 - \bar{w}_k)^{1/\rho}$. For $\rho = 0$ with floor ε, $U_0 \leq \varepsilon^{\bar{w}_k}$; without the floor, $U_0 = 0$.*

Proof. For $\rho \in (0,1)$, $U_\rho = (\sum_{j\neq k} \bar{w}_j x_j^\rho)^{1/\rho} \leq (\sum_{j\neq k} \bar{w}_j)^{1/\rho} = (1-\bar{w}_k)^{1/\rho}$ since $x_j \leq 1.0$. For $\rho = 0$, the k-th factor is $\varepsilon^{\bar{w}_k}$ and remaining factors are ≤ 1.0. □

Proposition 5 (Properties of the Asymptotic Display Ceiling)**.** *The soft-ceiling $\mathcal{C}(S)$ is C^1 continuous on $\mathbb{R}$, strictly increasing, satisfies $\mathcal{C}(S) \leq S$, and satisfies $\mathcal{C}(S) < 99.4$ for all finite S.*

Proof. For $S \leq s_0$, $\mathcal{C}'(S) = 1.0$. For $S > s_0$, $\mathcal{C}'(S) = \frac{H}{M} e^{-(S-s_0)/M} = e^{-(S-s_0)/H} \in (0,1)$ since $M = H$. As $S \downarrow s_0$, $\mathcal{C}'(S) \to 1.0$. Hence the first derivative is continuous everywhere. The upper bound follows from $\lim_{S\to\infty} \mathcal{C}(S) = s_0 + H = 86.0 + 13.4 = 99.4$. □

6 Empirical Evaluation and Validation

The empirical results measure outcome alignment with Agent Arena's cost-blind headline net-improvement proxy, not user satisfaction. The frontier-intent panel is the primary result because this proxy most closely matches a quality-first objective. Budget and balanced Arena-regret panels are boundary results: the proxy does not measure the cost-sensitive trade-offs those intents express. We therefore also report a secondary evaluation using one explicit cost-sensitive utility and the same hard cost constraints for every method.

6.1 Evaluation Setup and Live Telemetry Sources

To ensure empirical validity, candidate alternatives are derived from our live registry store, which is populated with catalog metadata and token pricing from the OpenRouter API [20] and with repository telemetry (architecture, licence, download and like counts) from the Hugging Face API [13]. The repository telemetry informs model identity and the source-trust factor of Section 4.8 only; no evaluation outcome reported in this section is computed from it. Capability and throughput metrics are calibrated against benchmark observations from the Artificial Analysis API [2], including Artificial Analysis Intelligence Indices, Coding Benchmark pass rates, and verified output throughput (tokens per second).

6.2 Illustrative frontier-coding recommendation

In the illustrative nine-model coding case, the parsed frontier profile assigns zero weight to affordability and ranks Claude Opus 5.5 first (score

98.32); two sub-frontier controls are demoted by the continuous quality-deficit penalty. This example illustrates pipeline behavior only and is not evidence of user-preference satisfaction; the comparative evidence is the frozen-cohort evaluation below. The archive contains the analysis scripts but omits the frozen inputs and generated complete rankings described in the Data Availability statement.

6.3 Exponent, missing-telemetry, and display stress tests

Across the nine-model exponent ablation, the ordering is unchanged for $\rho \in \{0, 0.25, 0.5, 1\}$ ($r_s = 1.00$). In a deliberately adversarial missing-context counterfactual, the $\rho = 0.5$ score is 75.5, compared with 69.2 for regularized Cobb–Douglas; this is a stress test, not an operating missing-data state. The soft display ceiling preserves distinct scores for three high-scoring models (98.32, 97.49, and 90.74). Inverting the ceiling in Eq. 27 gives pre-ceiling values of approximately 119.7, 112.1, and 91.9, respectively; a hard cap at 99 would collapse only the first two, while the third would remain at 91.9. These tests characterize score behavior and do not establish comparative recommendation quality; the analysis scripts are in the research archive, while generated outputs and frozen inputs are omitted as described in the Data Availability statement.

6.4 Outcome definition and provenance

The reference is the Agent Arena Overall leaderboard's headline net-improvement column. The payload also reports per-signal contrasts; an earlier extractor read the first signal rather than the headline. We corrected that provenance error, retained the original snapshot locally, and recomputed all reported panels from the headline column [3]. A separate harness defect ranked the proposed method on its *rounded* display score while baselines used unrounded scores; all current analyses rank on the exact shipped score (`scoreExact`). Both corrections affect reported results and are disclosed; the analysis scripts and parser outputs are in the research archive, while the frozen payloads are withheld as described in the Data Availability statement.

The leaderboard's headline outcome y_m is the equal-weighted mean of the per-model, per-signal treatment effects reported in the frozen payload; it is expressed in net-improvement percentage points. The 47-model cohort is evaluated at six per-task cost caps, $B_b \in \{\$0.10, \$0.25, \$0.50, \$1, \$2, \$5\}$. At each cap, the common eligible set $\mathcal{M}_b$ contains cohort models whose observed median Agent Arena task cost

is no greater than B_b. For method r, $\hat{m}_{r,b}$ is its highest-ranked eligible model, and $m_b^* = \arg\max_{m \in \mathcal{M}_b} y_m$ is the outcome oracle. Per-tier regret is the oracle outcome minus the selected outcome; the reported mean gives each of the six tiers equal weight:

$$R_{r,b} = y_{m_b^*} - y_{\hat{m}_{r,b}} = \max_{m \in \mathcal{M}_b} y_m - y_{\hat{m}_{r,b}}, \qquad \overline{R}_r = \frac{1}{6} \sum_{b=1}^{6} R_{r,b}. \tag{31}$$

Thus, $R_{r,b} = y_{m_b^*} - y_{\hat{m}_{r,b}}$ measures how far the recommendation falls below the best eligible outcome in that tier. This is outcome alignment with the cost-blind proxy, not user satisfaction; budget caps define eligibility, but the regret itself does not reward lower cost within a tier.

6.5 Coverage of the catalogue

The outcome panels above are bounded by the leaderboard they grade: Agent Arena lists only models that appear in agentic traffic, so the cohort is the full intersection available at the time of retrieval rather than a sample of the catalogue [3]. Every recommendation in Sections 6.6–6.12 is therefore a comparison of aggregation rules on the measured stratum, and none of them can say anything about a model the leaderboard does not list. The coverage statistic reported here is the complement of that statement: it is computed on the whole catalogue, it needs no outcome reference, and it records a property of the registry rather than of any aggregation rule.

A tier is unserved when its top eligible record is structural. Partition the catalogue into budget tiers by the maximum per-task cost an admissibility filter admits. Within each tier, rank the eligible records by the method's own score and take the top one. The tier is *unserved* if that top record carries a structurally estimated capability score while a measured record was eligible to be chosen—that is, if the recommendation is forced to a model whose quality index rests on a specification rather than on a benchmark, not because no measured model qualifies, but because the measured models are ranked below the estimate. The statistic counts tiers, not models, because a single stale estimate can displace a measured model in every tier in which it is eligible.

Repair moves the unserved tiers to zero. Before repair, duplicate records for the same underlying model—the same checkpoint surfaced under more than one provider or effort variant—inflate the eligible set and let a structural estimate displace measured records it should not have

competed against; the same duplication also discards confidence when two records disagree and neither provenance is preferred. After de-duplication against the registry and the assignment of a provenance-derived confidence to each remaining record, the number of unserved tiers falls from 4/24 to 0/24: on the repaired catalogue no budget tier has a structural estimate at its top while a measured record was eligible. The same repair moves unmeasured records inside the method's top ten from 8/40 to 1/40. Both figures are properties of the registry as repaired, and neither is a claim that the structural estimate is accurate where it is used; they say only that the regression to the estimate is no longer forced by a duplicate or an unresolved provenance conflict.

Why this section needs no outcome. An outcome-based panel can grade an aggregation rule only on the models the outcome covers, so the outcome panels of this paper are silent on the 366 of 671 catalogue records (54.5%) whose score rests on a structural estimate; that stratum is in the catalogue but outside every leaderboard used here. The coverage statistic is the only evidence available for it, and because it is computed without an outcome it can be reported even where the outcome cannot. Its favourable direction is a consequence of the registry repair rather than of any method, and it is reported as such: it restores the measured stratum to the tiers from which duplicates had pushed it, and it does not demonstrate that a remaining structural estimate would outperform a measured model if one were eligible.

6.6 Does Preference Integration Destroy Capability Alignment? Validation against Agent Arena

We first hold the intent fixed at the frontier, quality-maximising profile of Section 6.2 and grade every method on the headline outcome. Table 2 reports the result.

Three findings follow. In Table 2, Δ_{P-M} is proposed mean regret minus method mean regret; the displayed intervals are paired 95% bootstrap percentile intervals over model resamples. Point estimates equal to a confidence-interval endpoint can occur in this discrete resampling distribution.

- **Under a capability-first intent the composite leads the original unpenalized integrative rules on point-estimate regret, but ties the matched WSM variant.** It attains mean regret 0.09 and worst-tier regret 0.27, versus 0.49 and 2.38 for the strongest original baseline values; captures 0.989 of the tier oracle versus 0.943;

Table 2 Validation on the 2026-10-08 headline net-improvement column (n=47 distinct registry models): mean and worst-tier recommendation regret over per-task budgets \$0.10–\$5.00, paired 95% bootstrap percentile interval for Δ_{P-M}, defined as proposed mean regret minus method mean regret (2,000 model resamples), Spearman r_s with a 95% bootstrap interval, Kendall τ_b, and the fraction of the outcome's top three models contained in the method's top three. **Bold** marks the best *integrative* rule in the mean-regret column; the quality-only row is a cost-blind reference ordering. An interval containing zero means the gap is unresolved by this cohort. The composite is scored on its full six-coordinate profile with category fit, confidence discount and trust, whereas every baseline is scored on the three-coordinate triplet; Section 6.7 measures how much of the difference that asymmetry accounts for.

Decision Method	**Mean Regret**	**Worst**	Δ_{P-M} **(95% CI)**	r_s **(95% CI)**	τ_b	**Top-3**
Proposed (ρ=0.5)	**0.09**	**0.27**	—	0.816 [0.649, 0.911]	0.650	2/3
WSM (Linear)	0.49	2.38	[−0.40, 1.96]	**0.837** [0.676, 0.922]	0.674	1/3
WSM + Hard Floor	0.49	2.38	[−2.12, 1.64]	0.785 [0.612, 0.895]	0.643	1/3
TOPSIS	0.49	2.38	[−0.40, 1.96]	0.836 [0.680, 0.922]	0.673	1/3
Cobb–Douglas (ρ=0)	0.49	2.38	[−0.40, 1.96]	0.830 [0.669, 0.916]	0.661	1/3
CES only (no fit/penalty)	0.49	2.38	[−0.40, 1.96]	0.835 [0.675, 0.921]	0.671	1/3
Quality only (ref.)	0.09	0.27	[−0.22, 2.03]	0.838 [0.680, 0.926]	0.680	1/3

and attains the oracle at four of six tiers versus three for each original baseline. WSM with both the deficit factor and frontier reward also has mean regret 0.09, although its top-three overlap is 1/3. A separate descriptive value-for-money check gives the composite and quality-only ordering the same 24.2 net-improvement points per dollar, with neither recommendation dominated in any tier. Every paired regret interval contains zero, so these are unresolved point-estimate differences. Section 6.7 shows that removing the deficit factor alone leaves mean regret unchanged; removing the frontier reward raises it to 1.28.

- **Rank fidelity is undifferentiated, and the reference ordering is nominally the best predictor.** Spearman correlations lie in $[0.785, 0.838]$ with heavily overlapping intervals; the linear sum is nominally highest at 0.837, the proposed score is 0.816, and every interval contains the others. The quality-only ordering reaches 0.838 with mean regret 0.09. This is the expected behaviour of a cost-blind outcome and not evidence about the aggregator: net improvement prices neither cost nor latency, so the regret-minimising ordering of an admissible set is the quality ordering up to proxy noise, and any rule that spends weight on affordability or speed forfeits fidelity here *by construction*. The quality-only row is therefore reported as a reference ordering rather than entered in the comparison.
- **The composite's shortlist fidelity is its clearest descriptive advantage in this panel.** Its top three contains two of the outcome's three highest-ranked models (2/3), whereas each original baseline places one there (1/3). Spearman is not a lead: WSM is 0.837 and the proposed score is 0.816. The linear-aggregator ablation has the same mean regret (0.09) and the same shortlist overlap (2/3); therefore, this cohort does not show a regret or top-three overlap improvement attributable to the CES aggregator. Removing the deficit factor leaves regret unchanged, while removing the frontier reward raises it; the paired interval for the latter difference includes zero. WSM with the deficit factor and TOPSIS with the deficit factor both remain at 0.49 mean regret. These are cohort-specific comparisons, not evidence of a general causal effect. The shortlist statistic is coarse and needs external replication.

6.7 Component Ablation and Baseline Parity

Two questions bound the capability-first panel, and this subsection answers both on the same frozen cohort.

Are the baselines handicapped? The composite is scored on six coordinates with category fit, a confidence discount and a source-trust factor; the initial baselines are scored on the three-coordinate efficiency triplet. Table 3 therefore re-runs each baseline on the same six coordinates with the same parsed weights. The six-coordinate and triplet variants attain identical mean regret, worst-tier regret and value captured (0.49, 2.38 and 0.943); the frontier profile assigns 0.20 to context, 0.05 to privacy, and 0.15 to availability outside the triplet, or $0.40/(3.60 + 0.05 + 0.20 + 0.05 + 0.15) = 9.9\%$ of its nominal weight. These coordinates move no tier's pick in this panel. This parity check addresses coordinate coverage, but does not show that CES is responsible for the proposed score's lead.

Which component changes the result? Table 3 separates the below-threshold deficit factor from the frontier reward. Removing the deficit factor while retaining the reward leaves mean regret at 0.09 and changes no tier pick; removing the reward while retaining the deficit factor changes two of six picks and raises mean regret to 1.28 (paired difference versus the full score 1.19, 95% interval $[0, 2.49]$). Removing both also gives 1.28. Thus the old ablation attributed the change to the wrong component: on this cohort the point-estimate difference tracks the frontier reward, not the deficit factor, and its paired interval includes zero. WSM plus the deficit factor and TOPSIS plus the deficit factor each retain all six baseline picks and remain at 0.49. Adding both the deficit factor and frontier reward to WSM changes one pick and gives 0.09, but its top-three overlap remains 1/3; the corresponding TOPSIS variant remains at 0.49. Under budget and balanced intents, adding the production deficit factor to WSM or TOPSIS changes no tier picks or mean regrets (WSM 6.66 and 0.79; TOPSIS 5.10 and 1.52, respectively). Replacing the power mean by a linear aggregator leaves mean regret unchanged at 0.09 and retains top-three overlap at 2/3, matching the full pipeline. The confidence discount, category fit and source trust each move mean regret by less than 0.005 on this cohort. Under budget intent, removing the specialist bonus improves regret from 9.07 to 6.66, another cohort-specific result rather than a general mechanism.

6.8 Intent-Conditioned Regret across Archetypes

The purpose of the method is intent-conditioned recommendation, and the preceding test holds the intent fixed. We therefore repeat the analysis under three requests—*"best frontier coding models"*, *"budget coding models"* and *"balanced coding models"*—all parsed by the hosted intent parser

Table 3 Corrected component ablation and matched-penalty baseline comparison on the frozen cohort (n=47, 2026-10-08 headline column), under the capability-first intent of Table 2. WSM and TOPSIS use the same six normalized benefit coordinates and pinned weights; the deficit factor is the production archetype-specific multiplicative quality factor below threshold. The combined frontier variant also applies the above-threshold reward. Δ is each row's mean regret minus the full-pipeline mean; its 95% percentile interval uses 2,000 paired model bootstrap resamples with the tier oracle recomputed. Mean and worst are regret in net-improvement points; lower is better.

Variant	**Mean**	**Worst**	**Value capt.**	**Δ (95% CI)**	r_s	**Top-3**
Full pipeline (shipped)	**0.09**	**0.27**	**0.989**	—	0.816	**2/3**
minus deficit (reward retained)	0.09	0.27	0.989	0.00 [0.00, 0.00]	0.795	2/3
minus frontier reward	1.28	4.77	0.851	1.19 [0.00, 2.49]	0.763	2/3
minus deficit and reward	1.28	4.77	0.851	1.19 [0.00, 2.49]	0.731	2/3
linear aggregator	0.09	0.27	0.989	0.00 [0.00, 0.00]	0.817	2/3
WSM (six coordinates)	0.49	2.38	0.943	0.40 [−1.96, 0.40]	0.838	1/3
WSM + deficit	0.49	2.38	0.943	0.40 [−1.96, 0.40]	0.835	1/3
WSM + deficit + reward	**0.09**	**0.27**	**0.989**	0.00 [−2.03, 0.09]	0.836	1/3
TOPSIS (six coordinates)	0.49	2.38	0.943	0.40 [−1.96, 0.40]	0.836	1/3
TOPSIS + deficit	0.49	2.38	0.943	0.40 [−1.96, 0.40]	0.836	1/3
TOPSIS + deficit + reward	0.49	2.38	0.943	0.40 [−1.96, 0.48]	0.837	1/3

of Section 3. The parser returns the target category and an ordered priority list; the archetype classifier maps the request onto one of four weight profiles, and the priority list determines the vector when no archetype is named. The requests resolve on the shipped implementation to frontier (quality 3.60, affordability 0, speed 0.05, context 0.20, privacy 0.05, availability 0.15), budget (affordability 2.50, quality 0.85, speed 0.60, context 0.65, privacy 0.05, availability 0.05), and balanced (quality 1.25, affordability 0.05, speed 0.05, context 0.65, privacy 0.05, availability 0.05) profiles. The proposed score uses the full parsed vector; triplet baselines use the quality, affordability, and speed projection, so these original panels compare bundled scoring and coordinate choices. The vectors and parser outputs are pinned in the research artifact. Section 6.10 summarizes parser reliability; the cohort and budget tiers are as in Section 6.6.

Table 4 Mean regret, worst-tier regret and value captured under the *budget* intent (prompt "budget coding models"; parsed weights: affordability 2.50, quality 0.85, speed 0.60, context 0.65, privacy 0.05, availability 0.05) on the 2026-10-08 headline column. Regret is in net-improvement points; lower regret, higher value captured and a shorter worst tier are better. Each Δ interval is proposed regret minus method regret, using paired bootstrap resampling of models (95% percentile interval). **Bold** marks the best integrative rule.

Decision Method	Mean Regret	Worst	Value captured	Δ_{P-M} (95% CI)	r_s	τ_b
Proposed (ρ=0.5)	9.07	15.30	−0.057	—	−0.602	−0.430
WSM (Linear)	**6.66**	12.40	0.225	[−14.88, 5.33]	−0.582	−0.404
WSM + Hard Floor	9.64	15.30	−0.123	[−4.21, 6.31]	0.386	0.226
TOPSIS	**5.10**	**10.54**	0.406	[−11.45, 5.26]	−0.622	−0.447
Cobb–Douglas (ρ=0)	**5.10**	**10.54**	0.406	[−11.45, 5.26]	−0.623	−0.441
CES only	**6.66**	12.40	0.225	[−11.45, 6.18]	−0.584	−0.410
Quality only (ref.)	0.09	0.27	0.989	[3.22, 14.31]	0.838	0.680

Table 5 Mean regret, worst-tier regret and value captured under the *balanced* intent (prompt "balanced coding models"; parsed weights: quality 1.25, affordability 0.05, speed 0.05, context 0.65, privacy 0.05, availability 0.05) on the 2026-10-08 headline column. Regret is in net-improvement points; lower regret, higher value captured and a shorter worst tier are better. Each Δ interval is proposed regret minus method regret, using paired bootstrap resampling of models (95% percentile interval). **Bold** marks the best integrative rule on mean regret.

Decision Method	Mean Regret	Worst	Value captured	Δ_{P-M} (95% CI)	r_s	τ_b
Proposed (ρ=0.5)	3.06	7.68	0.644	—	0.604	0.430
WSM (Linear)	**0.79**	2.38	0.908	[−0.40, 7.70]	0.760	0.573
WSM + Hard Floor	**0.79**	2.38	0.908	[−1.01, 7.70]	0.701	0.534
TOPSIS	1.52	4.41	0.822	[−2.31, 8.04]	0.765	0.586
Cobb–Douglas (ρ=0)	6.47	16.43	0.246	[−5.75, 4.61]	0.693	0.523
CES only	1.73	5.90	0.799	[−0.40, 6.76]	0.736	0.558
Quality only (ref.)	0.09	0.27	0.989	[0.17, 8.64]	0.838	0.680

These budget and balanced tables are boundary results for alignment with the cost-blind Arena outcome. On that outcome, the proposed method's regret is 9.07 under budget (versus 5.10 for TOPSIS and Cobb–Douglas) and 3.06 under balanced (versus 0.79 for WSM and floored WSM). The comparisons preserve the shipped parsed profiles and are not evidence about which method better satisfies cost-sensitive users. In the next analysis, all methods are instead graded with the same explicit cost-sensitive utility.

6.9 Shared cost-sensitive utility evaluation

As a secondary outcome independent of the Arena net-improvement column, we evaluated every ranking on the same four normalized attributes: predicted coding quality q_m, log-inverse observed Agent Arena median session cost c_m, speed s_m, and context x_m. The common linear utility is

$$U_{m,p} = \sum_{k \in \{q,c,s,x\}} \tilde{w}_{p,k} z_{m,k}, \qquad \sum_{k} \tilde{w}_{p,k} = 1, \tag{32}$$

where $z_{m,k}$ are min–max normalized over the frozen 47-model cohort (cost uses $1 - \log(C_m/C_{\min})/\log(C_{\max}/C_{\min})$). The budget utility uses normalized weights $(0.1848, 0.5435, 0.1304, 0.1413)$ for quality, cost utility, speed, and context, taken from the parsed budget profile. The parser supplied no explicit balanced weights for these four evaluation attributes, so balanced uses the equal-weight assumption $(0.25, 0.25, 0.25, 0.25)$. Neither vector is validated against observed user choices. Every method first receives the same hard eligibility filter, observed median task cost $\leq B_b$, for each of the six caps B_b defined above. Utility regret is $\max_{m \in \mathcal{M}_b} U_{m,p} - U_{\hat{m}_{r,b},p}$, averaged equally over the six tiers. We also report arithmetic mean quality, mean observed session cost, and the number of selected tier recommendations dominated on quality and cost (lower cost and at least equal quality, with one strict).

The "Proposed (harmonized)" row retains the full scoring pipeline and intent-specific four-axis weights, replacing catalogue-price affordability with normalized observed session-cost utility. In this offline harness, both prompts use the balanced archetype; thus the proposed budget row also receives the shared balanced/budget/speed quality-deficit penalty, rather than a budget-specific archetype penalty. The dominated-tiers count compares only predicted quality and observed session cost, so it ignores speed

Table 6 Common cost-sensitive utility results on the frozen 47-model cohort. Every method uses the same hard observed-cost cap within each tier and is graded by the same intent-specific utility. Utility regret is averaged over six tiers; quality is the mean normalized quality; cost is mean observed median session cost in USD; dominated reports quality–cost Pareto-dominated selected tiers. The proposed row is the harmonized re-scoring of the proposed method using the shared observed-cost utility, rather than its shipped score. WSM's zero regret for both profiles follows by construction because it maximizes the same linear utility over each eligible set.

Intent	Method	Utility regret	Mean quality	Mean cost (USD)	Dominated tiers
Budget	Proposed (harmonized)	0.0231	0.8481	0.0520	6/6
	WSM	0.0000	0.9247	0.0738	0/6
	TOPSIS	0.0105	0.8983	0.0750	3/6
	Cobb–Douglas	0.0139	0.9205	0.1163	5/6
	CES ($\rho = 0.5$)	0.0000	0.9247	0.0738	0/6
Balanced	Proposed (harmonized)	0.0181	0.8969	0.2138	6/6
	WSM	0.0000	0.9349	0.3517	5/6
	TOPSIS	0.0074	0.9097	0.2175	5/6
	Cobb–Douglas	0.0000	0.9349	0.3517	5/6
	CES ($\rho = 0.5$)	0.0000	0.9349	0.3517	5/6

and context. WSM is also dominated at 5/6 balanced tiers, so 6/6 versus 5/6 is not a meaningful gap.

The proposed harmonized ranking has nonzero shared-utility regret for both profiles and its selected recommendations are quality–cost dominated at all six tiers. Under balanced weights it has lower mean cost than the alternatives but also lower mean quality; the scalar utility expresses the chosen trade-off, while the dominance count exposes quality–cost inefficiency without assigning a weight. WSM's zero regret for each profile is an optimizer check rather than independent comparative evidence. These descriptive estimates share the cohort and observed cost source, have no user-preference ground truth or uncertainty interval, and should be treated as a secondary evaluation rather than independent external validation.

6.10 Parser reliability and weight diagnostics

Across 17 prompts parsed four times, 16 returned identical categories, priority lists, and derived weights on all runs; one varied in a single run, changing its context/privacy priorities. Category was stable for all 17. Paraphrase tests exposed archetype-classification instability for budget and latency requests, so repeatability of the evaluated prompts does not establish invariance to wording or fidelity to user intent. The archive includes the pinned parser outputs used by the reported panels; full per-repeat and paraphrase diagnostics are not included.

6.11 Nominal weights and realized influence

Nominal weights do not directly equal realized influence after normalization, nonlinear transforms, and multiplicative factors. For this 47-model cohort, privacy and availability are constant and cannot affect any reported ranking. The archive includes a script to compute per-coordinate influence from the frozen inputs; because those inputs are withheld, the resulting decomposition is not included here.

6.12 Cost and value-for-money results

The cost-sensitive Arena-regret panels are boundary results rather than evidence of intent satisfaction. Under the separate common-utility evaluation (Table 6), the harmonized proposed ranking has nonzero utility regret and is quality–cost dominated at all six selected tiers for both profiles. Under balanced weights it recommends lower-cost models on average, with lower mean quality than WSM, Cobb–Douglas, and CES.

The utility analysis adds cost, quality, and Pareto evidence to the cost-blind outcome alignment, but remains descriptive because its weights are researcher-specified and its cost and quality inputs come from the same frozen cohort.

7 Discussion and Limitations

The formal results establish properties of the scoring functions under their stated assumptions; they do not establish that the chosen profiles reflect user preferences. The primary empirical claim is narrow: on this cohort, the shipped method has the lowest capability-first point-estimate regret among integrative baselines, with paired intervals containing zero. Budget and balanced Arena-regret results are boundary comparisons against a proxy that omits cost-sensitive trade-offs. The common-utility analysis addresses that gap by applying the same explicit utility and hard cost caps to every method, but the proposed harmonized ranking has nonzero utility regret and is quality–cost dominated at all six tiers for each profile. The utility weights are assumptions rather than observed preferences, so neither analysis establishes intent satisfaction.

Several implementation and study limits matter. The parser is hosted and its outputs are recorded rather than regenerated. The evaluation uses one dated 47-model cohort and six nested budget tiers, so the effective number of independent decisions is small and uncertainty intervals are wide; we reuse one inspected cohort for all analyses, so this is not a temporal holdout. Privacy and availability are constant across the 47 models, so their ranking effects are untested. The Arena cohort does not cover the full catalogue; coverage describes registry behavior but cannot validate structurally estimated quality. Prices may be missing or stale, and the current budget filter uses input price rather than blended per-task cost. The matched ablation applies the same six-coordinate weights and production deficit factor to WSM and TOPSIS, and tests the frontier reward separately; it does not transfer category fit, confidence discount, source trust, or the coding-specialist adjustment to those baselines. Simple maximum-quality-under-a-common-budget and lexicographic rules are not evaluated. No independent expert or direct pairwise preference study is available. Telemetry maps and penalty constants are not comprehensively calibrated. The parser outputs and analysis scripts are in the accompanying archive; frozen third-party inputs are omitted as described in the Data Availability statement.

7.1 Threats to validity

The main threats are outcome-column dependence, snapshot dependence, small and nested tier counts, parser instability, incomplete catalogue coverage, untested privacy and availability variation, and the absence of user or expert preference ground truth. There is no temporal hold-out or direct pairwise-choice cohort. The scripts document the extractor and rounded-ranking corrections; frozen payloads are not redistributed. Capability-first differences remain unresolved point estimates; no claim of human-preference validation is made. Further fully factor-matched comparisons, incorporating category fit, confidence discounting, source trust, and coding-specialist adjustments, together with simple cost-feasible decision rules, are needed before attributing improvements to the aggregation design.

7.2 Scope of the claims

Mean regret against Agent Arena's outcome is an outcome-alignment measure; because that outcome is cost-blind, it cannot establish intent fidelity for budget or balanced requests. We therefore add a common-utility evaluation with the same hard cost constraints for every method and report utility regret, cost, quality, and Pareto dominance. This is stronger evidence about cost-sensitive trade-offs than the Arena panels alone, but its researcher-specified weights and shared cohort limit the claim. Direct pairwise choices from budget and balanced users remain the appropriate next validation.

8 Conclusion

This paper presents a foundation-model recommendation pipeline that combines hard feasibility constraints, Pareto stratification, CES aggregation, a continuous quality-deficit penalty, and a bounded display mapping. We prove monotonicity, a quality-conditional score ceiling, and display-ceiling properties under stated assumptions, and show by counterexample that Pareto-tier membership does not determine score rank when coordinates outside the efficiency triplet carry weight.

On a frozen 47-model Agent Arena cohort, the pipeline attains mean regret 0.09 under capability-first intent, versus 0.49 for the original unpenalized WSM and TOPSIS baselines; paired intervals include zero. WSM with both matched quality factors ties the pipeline at 0.09. Corrected ablation shows that removing the deficit factor leaves mean regret at 0.09

and changes no tier pick, whereas removing the frontier reward changes two of six picks and raises mean regret to 1.28; the paired interval for this difference includes zero. Adding the deficit factor alone to WSM or TOPSIS leaves mean regret at 0.49 and changes no baseline pick. The WSM variant with both factors has top-three overlap of 1/3; the corresponding TOPSIS variant remains at 0.49. Under budget and balanced intents, adding the deficit factor to either baseline changes no tier picks or mean regrets. Replacing the power mean with a linear aggregator leaves mean regret unchanged and retains top-three overlap at 2/3, matching the full pipeline in this cohort. These results identify a cohort-specific point-estimate pattern, not a general causal advantage. Under budget and balanced intents the pipeline trails several baselines (9.07 versus 5.10 under budget, and 3.06 versus 0.79 under balanced intent). Because the outcome prices neither cost nor latency, these comparisons identify a weakness against the available proxy and do not establish what a cost-sensitive user would prefer.

In the secondary shared-utility evaluation, the harmonized proposed scorer has mean utility regret of 0.0231 for budget and 0.0181 for balanced intent; its recommendations are quality–cost dominated in all six tiers for both profiles. These values use explicit evaluation weights, not observed user preferences.

The study is limited to one dated cohort and six budget tiers; telemetry maps and penalty constants are not comprehensively calibrated; privacy and availability are constant in the cohort; and parser outputs are recorded rather than regenerated. We therefore claim outcome alignment under capability-first intent, not user-intent satisfaction. Future work includes temporal validation, direct pairwise preference studies involving budget- and balanced-intent users, validation of shared-utility weights against observed user choices, and fully factor-matched baseline comparisons.

Declarations. **Funding.** "The authors received no specific funding for this work."

Competing interests. The authors declare that they have no competing financial or non-financial interests that could have influenced the research presented in this manuscript.

Ethics approval and consent to participate. Not Applicable.

Consent for publication. Not applicable. This study does not include any personally identifiable information, individual person or any objectionable images, or other material requiring consent for publication.

Clinical trial registration. Not applicable. This study is not a clinical trial; therefore, no clinical trial registration number is applicable

Data availability. The accompanying `Model-Compass-Research` source archive contains the analysis code and the recorded intent-parser outputs, but it does not contain the frozen model registry or the Agent Arena snapshots used for the reported tables. The two local Arena snapshots were produced by parsing the leaderboard's rendered-page payload; the current Arena Terms of Use restrict automated extraction and redistribution of service output. We therefore do not redistribute those snapshots or claim that the archive alone reproduces the numerical results. Arena's publisher separately provides a historical leaderboard dataset through Hugging Face under CC BY 4.0 (https://huggingface.co/datasets/lmarena-ai/leaderboard-dataset); we have not established that its released records are equivalent to the specific frozen headline values used here, so it is not presented as a drop-in reproduction input. The registry also contains observations from OpenRouter, Hugging Face, and Artificial Analysis and is not bundled. The exact external repository/DOI is not deposited for this version. The hosted parser is not rerun: the frozen parser outputs are included in the archive.Zenodo

Code availability. The analysis code is supplied in the accompanying `Model-Compass-Research` source archive. It preserves the scoring implementation used for the reported analyses; the live application has since evolved, and its current registry and scores are not the basis of these paper results. No public repository URL or DOI has been deposited for this version. The archive includes the frozen analysis scripts and parser outputs but excludes the third-party snapshots and registry described above; running the full analyses therefore requires those exact inputs to be obtained through an authorized route. The archive's `LICENSE` file states the current reuse terms for the authors' code; third-party material remains subject to its own terms.Zenodo

Author contributions. Surbhi Agrawal: Conceptualization, methodology development, theoretical formulation, formal analysis, research supervision, validation, interpretation of results, writing—original draft, and writing—review and editing.

Divya Chouksey.: Software development and implementation, data collection and curation, experimental execution, computational analysis, visualization, reproducibility documentation, and contributions to manuscript preparation.

Both authors reviewed and approved the final version of the manuscript.

AI use in manuscript preparation. AI assisted with manuscript formatting and language editing to improve readability, presentation, and clarity.

References

[1] Arrow, K.J., Chenery, H.B., Minhas, B.S., Solow, R.M.: Capital-labor substitution and economic efficiency. *The Review of Economics and Statistics* **43**(3), 225–250 (1961). doi:10.2307/1927286

[2] Artificial Analysis: LLM quality, speed, and pricing benchmarks API (2024). https://artificialanalysis.ai/api-reference. Accessed 30 September 2026

[3] Arena: Agent Arena leaderboard: net improvement and median cost per task. https://arena.ai/leaderboard/agent and https://arena.ai/leaderboard/agent/pareto. Payload last updated 8 October 2026, retrieved 11 October 2026

[4] LMArena: Arena leaderboard dataset (`lmarena-ai/leaderboard-dataset`). Hugging Face (2026). https://huggingface.co/datasets/lmarena-ai/leaderboard-dataset

[5] Bommasani, R., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021). arXiv:2108.07258

[6] Brans, J.P., Vincke, P.: A preference ranking organisation method: the PROMETHEE method for multiple criteria decision-making. *Management Science* **31**(6), 647–656 (1985). doi:10.1287/mnsc.31.6.647

[7] Chen, M., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021). arXiv:2107.03374

[8] Chen, L., Zaharia, M., Zou, J.: FrugalGPT: How to use large language models while reducing cost and improving performance. arXiv preprint arXiv:2305.05176 (2023). arXiv:2305.05176

[9] Cobbe, K., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021). arXiv:2110.14168

[10] Cobb, C.W., Douglas, P.H.: A theory of production. *The American Economic Review* **18**(1), 139–165 (1928). doi:10.2307/1811556

[11] Hardy, G.H., Littlewood, J.E., Pólya, G.: *Inequalities*. Cambridge University Press, Cambridge (1934)

[12] Hendrycks, D., et al.: Measuring massive multitask language understanding. In: *International Conference on Learning Representations (ICLR)* (2021). openreview:d7KBjmI3GmQ

[13] Hugging Face: Open LLM Leaderboard and Hub API. Hugging Face Spaces (2024). https://huggingface.co/docs/hub/api

[14] Hwang, C.L., Yoon, K.: *Multiple Attribute Decision Making: Methods and Applications*. Springer, Berlin (1981). doi:10.1007/978-3-642-48318-9

[15] Jimenez, C.E., et al.: SWE-bench: Can language models resolve real-world GitHub issues? In: *International Conference on Learning Representations (ICLR)* (2024). openreview:shMODneKD4

[16] Kaplan, J., et al.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020). arXiv:2001.08361

[17] Keeney, R.L., Raiffa, H.: *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, Cambridge (1993). doi:10.1017/CBO9780511983832

[18] Kung, H.T., Luccio, F., Preparata, F.P.: On finding the maxima of a set of vectors. *Journal of the ACM* **22**(4), 469–476 (1975). doi:10.1145/321892.321899

[19] Ong, I., et al.: RouteLLM: Learning to route LLMs with preference data. arXiv preprint arXiv:2406.18665 (2024). arXiv:2406.18665

[20] OpenRouter: Unified API and model benchmarks endpoint (2024). https://openrouter.ai/docs/api-reference/overview. Accessed 30 September 2026

[21] Opricovic, S., Tzeng, G.-H.: Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research* **156**(2), 445–455 (2004). doi:10.1016/S0377-2217(03)00020-1

[22] Roy, B.: Classement et choix en présence de points de vue multiples (la méthode ELECTRE). *Revue Française d'Informatique et de Recherche Opérationnelle* **2**(8), 57–75 (1968). doi:10.1051/ro/196802V100571

[23] Saaty, T.L.: *The Analytic Hierarchy Process*. McGraw-Hill, New York (1980)

[24] Tversky, A.: Elimination by aspects: A theory of choice. *Psychological Review* **79**(4), 281–299 (1972). doi:10.1037/h0032955

[25] Zheng, L., et al.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2023). NeurIPS:2023