

Psychometric Tests for AI Agents and Their Moduli Space

Przemysław Chojecki
ulam.ai

November 24, 2025

Abstract

We develop a moduli-theoretic view of psychometric test batteries for AI agents and connect it explicitly to the AAI score developed in [1]. First, we make precise the notion of an AAI functional on a battery and set out axioms that any reasonable autonomy/general intelligence score should satisfy. Second, we show that the composite index ('AAI-Index') defined in [1] is a special case of our AAI functional. Third, we introduce the notion of a cognitive core of an agent relative to a battery and define the associated AAI_{core} score as the restriction of an AAI functional to that core. Finally, we use these notions to describe invariants of batteries under evaluation-preserving symmetries and outline how moduli of equivalent batteries are organized.

1 Introduction

In psychology psychometric evaluation aims to measure capability via structured batteries of tasks. In large-agent settings (LLMs, tool-using systems, embodied agents), a battery can be understood as structured data plus evaluation rules plus resource accounting. This paper builds on the AAI score [1] by: (i) formalizing batteries as objects and morphisms, (ii) defining AAI functionals on batteries, (iii) proving that the AAI-Index from AAI is an instance of such a functional, (iv) introducing AAI_{core} , a score attached to the agent's cognitive core, and (v) describing the moduli space of batteries and its continuous structure.

In this framework, AGI is not a single-task milestone but a property of an agent's performance over the moduli of batteries. A frequent critique of AI evaluation is the "overfitting" to specific benchmarks. *This paper argues mathematically that one need not obsess over the specifics of the tests themselves, provided one samples enough of them.* This is akin to students training for a Math Olympiad by solving problems from past competitions: they saturate the space of known problem types not to memorize answers, but to internalize the underlying logic required to solve new, similar problems.

Our determinacy result (Theorem 6.12) implies a practical certification rule: if the family of scoring functionals is "regular" (Lipschitz continuous on the moduli space), then performance on a sufficiently dense finite panel (a δ -net of canonicalized batteries) mathematically certifies the agent's performance across the entire continuum of possible tests, up to a controllable error term.

Specifically, if an agent's AAI score meets or exceeds the AGI threshold on every panel instance with margin m , then its worst-case AAI over the whole moduli is at least threshold $-2L\delta$ (where L is the Lipschitz modulus). In other words, sufficiently many strong results on a well-designed, diverse panel certify the level globally. Because the functional is resource-aware (via the projective geometry of resource rays and explicit cost terms), a certified level is about capabilities per unit cost, not just raw wins, aligning with the operational AAI score [1].

A second insight is that the cognitive core makes “true generality” testable rather than anecdotal. High AAI_{core} across heterogeneous task families, small gaps between the full AAI and AAI_{core} (indicating competence rather than scaffolding), low dispersion penalties, and stability under battery symmetries and seeded drifts together provide necessary evidence patterns for AGI. Moreover, when the induced core aligns with broad CHC-style factors and remains low-dimensional while still reconstructing threshold decisions across families, it indicates transferable structure rather than overfit tricks [2].

Many of our structural results are not specific to the particular moduli space of test batteries. In fact, the determinacy, regularity, and concentration theorems hold for any Lipschitz-regular functionals on a metric space (S, d) with a dense family of “simple” configurations and, for the axis-based scores, a finite collection of axis functionals whose expectations are Lipschitz in d .

Notation and conventions

For a measurable space X , let $\mathcal{P}(X)$ denote probability laws on X . For task families $T = \bigsqcup_k F_k$, write $u(t) \in [0, 1]$ for a PIT-uniformized per-task score (defined below), $q(t) \in [0, 1]$ for threshold-aligned success, and $r \in \mathbb{R}_+^{d_R}$ for resources. Expectations $\mathbb{E}[\cdot]$ are taken over seeds/drifts unless stated otherwise.

2 Batteries and Morphisms

Definition 2.1 (Battery). A *battery* is an octuple

$$\mathcal{B} = (T, \mathcal{F}, \mathcal{S}, Q^*, \mu, \mathcal{D}, \Pi, \mathcal{R}),$$

where:

- T is a finite set of tasks; $\mathcal{F} = \{F_k\}$ is a partition of T into families.
- $\mathcal{S} = \{S_t : \Omega_t \rightarrow [0, 1]\}_{t \in T}$ are task-specific scoring maps.
- $Q^* : T \rightarrow [0, 1]$ are task thresholds.
- μ is a sampling law on $T \times \Pi \times \mathcal{D}$ (tasks, seeds, drifts).
- \mathcal{D} (drifts) and Π (seeds) are measurable spaces.
- $\mathcal{R} \cong \mathbb{R}^{d_R}$ are resource coordinates (e.g., time, tokens, cost), recorded nonnegatively.

Definition 2.2 (Agent representation on a battery). Fix \mathcal{B} and an agent \mathcal{A} . A run on $t \in T$ with seed $s \in \Pi$ and drift $\delta \in \mathcal{D}$ and resource vector $r \in \mathcal{R}$ produces a score $q(t) = S_t(\text{Run}(\mathcal{A}; t, s, \delta, r)) \in [0, 1]$. Write

$$X_{\mathcal{B}} := [0, 1]^T \times \mathbb{R}_{\geq 0}^{d_R}.$$

Let \mathbb{P} be the joint probability over (t, s, δ) drawn from μ and any internal randomness of \mathcal{A} . The *agent representation* is the pushforward (image) measure

$$\rho_{\mathcal{B}}(\mathcal{A}) := \mathbb{P} \circ ((q(t))_{t \in T}, r)^{-1} \in \mathcal{P}(X_{\mathcal{B}}).$$

Definition 2.3 (Morphisms of batteries). A morphism $f : \mathcal{B} \rightarrow \mathcal{B}'$ is a tuple

$$f = (f_T, f_{\Pi}, f_{\mathcal{D}}, f_{\mathcal{R}}, \{\phi_t\}_{t \in T})$$

satisfying:

1. $f_T : T \rightarrow T'$ respects families ($t \in F_k \Rightarrow f_T(t) \in F'_k$).
2. $f_{\Pi} : \Pi \rightarrow \Pi'$ and $f_{\mathcal{D}} : \mathcal{D} \rightarrow \mathcal{D}'$ are measurable and measure-preserving relative to μ .

3. $f_R : \mathbb{R} \rightarrow \mathbb{R}'$ is linear and unit-consistent.
4. $\phi_t : [0, 1] \rightarrow [0, 1]$ are strictly increasing and continuous.
5. (Evaluation preservation) For all agents \mathcal{A} and $t \in T$,

$$S'_{f_T(t)}(\text{Run}(\mathcal{A}; f_T(t), f_\Pi(s), f_D(\delta), f_R(r))) = \phi_t(S_t(\text{Run}(\mathcal{A}; t, s, \delta, r))),$$

almost surely (w.r.t. $\mu(\cdot | t)$).

Proposition 2.4. *Batteries and their morphisms form a category \mathbf{Bat} under composition and identities.*

Proof. Identity: take identity maps with $\phi_t = \text{id}$. Composition is componentwise; evaluation preservation is stable under composition. Associativity and identity laws are immediate. \square

3 Symmetries and the Moduli Space

Definition 3.1 (Symmetry group). Let

$$G := \left(\prod_k \text{Sym}(F_k) \right) \times \text{Aut}_\mu(\Pi) \times \text{Aut}_\mu(D) \times \left(\prod_{t \in T} \text{Homeo}^+([0, 1]) \right) \times (\mathbb{R}_{>0})^{d_R},$$

acting by within-family task permutations, measure-preserving relabelings of seeds and drifts, strictly increasing per-task score reparameterizations, and positive unit rescalings of resources.

Definition 3.2 (Moduli space). The *moduli space of batteries* (coarse, set-level) is the set of isomorphism classes

$$\mathfrak{M} := \mathbf{Bat}/G.$$

Remark. If automorphisms matter, the appropriate object is the quotient stack (action groupoid) $[\mathbf{Bat}/G]$. Here we work with \mathfrak{M} .

4 PIT Normalization and Canonical Representation

Definition 4.1 (Randomized PIT). Given a scalar task score $s(t)$ with conditional CDF F_t , define $u(t) = F_t(s(t)^-) + \xi(F_t(s(t)) - F_t(s(t)^-))$ with $\xi \sim \text{Unif}[0, 1]$. Then $u(t) \sim \text{Unif}[0, 1]$ conditional on drift and seed.

Proposition 4.2 (Uniformity and invariance). *PIT-normalized $u(t)$ is invariant to strictly increasing reparameterizations of $s(t)$ and enables copula-based comparisons across tasks.*

Proof. Uniformity follows from the probability integral transform with randomization at discontinuities. We define the transformed CDF

$$F_t^\phi(y) := \mathbb{P}(\phi(s(t)) \leq y | \text{drift, seed}).$$

If ϕ is strictly increasing, then

$$F_t^\phi(y) = \mathbb{P}(\phi(s(t)) \leq y | \dots) = \mathbb{P}(s(t) \leq \phi^{-1}(y) | \dots) = F_t(\phi^{-1}(y)).$$

In particular, for $y = \phi(s)$ we have

$$F_t^\phi(\phi(s)) = F_t(s),$$

so the distribution of $u(t)$ is unchanged; the copula of $(u(t))$ captures dependence invariantly. See [4] for a proof. \square

Definition 4.3 (Canonical representation). Let X^{can} contain $(u(t))_{t \in T}$, threshold-aligned $q(t) = \mathbb{1}\{u(t) \geq \tau(t)\}$ with $\tau(t) = F_t(Q^*(t))$, and the resource ray $[r]$. The pushforward law on X^{can} is the canonical representation of an agent on \mathcal{B} .

5 Topological Structure of the Moduli

Proposition 5.1 (Continuous parameters). *Fix the discrete task skeleton (families, anchors) and threshold structure. The moduli space decomposes as:*

$$\mathfrak{M} \simeq (\text{Discrete Data}) \times (\text{Thresholds } \tau \in [0, 1]^T) \times (\text{Copulas } C_u \in C) \times (\text{Rays } [r] \in \mathbb{P}(\mathbb{R}_+^{d_R})),$$

where C denotes the set of all copulas on $[0, 1]^T$, endowed with the W_1 -topology.

Proof. Discrete invariants (family sizes, anchor structure, threshold ordering) partition the space into strata. Within a fixed stratum, every battery is specified by the coordinates $(\tau, C_u, [r])$: the threshold vector τ , the copula C_u of the canonical PIT scores, and the projective resource ray $[r]$. The symmetry group G acts trivially on τ (once the discrete threshold structure is fixed), while C_u varies continuously in the Wasserstein topology and $[r]$ in the projective topology. This yields the stated decomposition. \square

Remark 5.2 (Wasserstein geodesics). Viewing each copula C_u as a probability measure on $[0, 1]^T$ and equipping the copula space with the Wasserstein–1 metric W_1 , any two copulas C_0, C_1 can be connected by a W_1 -geodesic $(C_t)_{t \in [0, 1]}$, also called a displacement interpolation. The path $t \mapsto C_t$ describes a controlled drift of dependence across tasks: C_t deforms C_0 into C_1 by moving probability mass along optimal transport plans. Evaluating our AAI functionals along such geodesics provides a natural way to stress test robustness against gradual changes in task dependence. See [5].

Discrete and continuous moduli. Fixing a battery \mathcal{B} determines a collection of discrete invariants: the partition of tasks into families $\{F_k\}$, the choice of anchors within each family, and the qualitative threshold structure (e.g. which thresholds coincide and the induced partial order on families). We refer to this finite combinatorial datum as the *battery skeleton* and denote it by S . Let \mathcal{S} be the set of all such skeletons, modulo the action of the symmetry group G . For a fixed skeleton $S \in \mathcal{S}$, Proposition 5.1 shows that the corresponding configurations form a continuous parameter space

$$\mathfrak{M}_S \simeq \{\tau \in [0, 1]^T\} \times \{C_u \in C\} \times \mathbb{P}(\mathbb{R}_+^{d_R}),$$

where τ collects the task thresholds, C_u is the copula of the canonical PIT scores, and $[r] \in \mathbb{P}(\mathbb{R}_+^{d_R})$ is the resource ray. In particular, once S is fixed, the remaining degrees of freedom are purely continuous and live in a product of metric and measure-theoretic spaces (thresholds in $[0, 1]^T$, copulas in the Wasserstein space (C, W_1) , and rays in the positive projective space).

Stratification by combinatorial type. It is therefore natural to view the full moduli space as a stratified space indexed by battery skeletons:

$$\mathfrak{M} \simeq \bigsqcup_{S \in \mathcal{S}} \mathfrak{M}_S \approx \bigsqcup_{S \in \mathcal{S}} \underbrace{[0, 1]^T}_{\tau \text{ (thresholds)}} \times \underbrace{C}_{C_u \text{ (copulas)}} \times \underbrace{\mathbb{P}(\mathbb{R}_+^{d_R})}_{[r]},$$

where each stratum \mathfrak{M}_S is the continuous parameter space associated with a fixed combinatorial type S .

In this perspective, the “discrete moduli” is the set \mathcal{S} of skeletons, while the “continuous moduli” over each S is given by the coordinates $(\tau, C_u, [r])$ described above. Passing from one skeleton to another corresponds to degenerations of the discrete data: thresholds colliding (e.g. $\tau_i = \tau_j$), families merging or splitting, or anchors appearing and disappearing. This is directly analogous to the familiar picture in algebraic geometry in which a moduli space is stratified by combinatorial types (dual graphs, incidence data), with each stratum carrying a continuous family of parameters and the boundaries between strata encoding degenerations of that combinatorial structure.

6 AAI as a Functional on Representations

Definition 6.1 (AAI functional). For each battery \mathcal{B} , an AAI functional is a measurable map $\Phi_{\mathcal{B}} : \mathcal{P}(X_{\mathcal{B}}) \rightarrow \mathbb{R}$ assigning $\text{AAI}_{\mathcal{B}}(\mathcal{A}) = \Phi_{\mathcal{B}}(\rho_{\mathcal{B}}(\mathcal{A}))$ and satisfying axioms:

- (A1) **Naturality.** If f is a symmetry with pushforward $f_* : \mathcal{P}(X_{\mathcal{B}}) \rightarrow \mathcal{P}(X_{\mathcal{B}'})$, then $\Phi_{\mathcal{B}'}(f_*\nu) = \Phi_{\mathcal{B}}(\nu)$.
- (A2) **Restricted Monotonicity.** If ν' dominates ν in increasing concave order (second-order stochastic dominance) with respect to the success indicators, and

$$\text{Var}_k(\mathbb{E}_{\nu'}[\bar{q}(F_k)]) \leq \text{Var}_k(\mathbb{E}_{\nu}[\bar{q}(F_k)])$$

(i.e., dispersion does not increase), and the expected resource cost is non-increasing, then $\Phi_{\mathcal{B}}(\nu') \geq \Phi_{\mathcal{B}}(\nu)$.

- (A3) **Threshold calibration.** Increasing $\mathbb{P}_{\nu}\{q(t) \geq Q^*(t)\}$ increases $\Phi_{\mathcal{B}}(\nu)$, with highest sensitivity near thresholds.
- (A4) **Generality.** Family means enter symmetrically; dispersion penalties discourage unfair concentration.

Definition 6.2 (Tractable instance). Let $\bar{q}(F_k) = |F_k|^{-1} \sum_{t \in F_k} q(t)$. Define

$$\Phi_{\mathcal{B}}(\nu) = \sum_k w_k \mathbb{E}_{\nu} \left[\frac{1}{|F_k|} \sum_{t \in F_k} \psi_t(q(t), Q^*(t)) \right] - \lambda \text{Var}_k(\mathbb{E}_{\nu}[\bar{q}(F_k)]) - \gamma \mathbb{E}_{\nu}[\text{Cost}(r)]. \quad (1)$$

Remark 6.3 (Absolute vs. Projective Resources). For this specific tractable instance, we treat absolute resource usage as part of the evaluation data. Accordingly, for the analysis of this functional, we restrict the symmetry group G to exclude resource rescalings, thereby defining the moduli over absolute resource vectors rather than projective rays. Alternatively, one could define the cost term as $\mathbb{E}_{\nu}[\text{Cost}([r])]$, where $\text{Cost}([r])$ depends only on the projective resource ray $[r] \in \mathbb{P}(\mathbb{R}_+^{d_R})$ (i.e., is invariant under positive rescalings $r \mapsto cr$). This would make $\Phi_{\mathcal{B}}$ strictly well-defined on the projective moduli space, though at the expense of ignoring absolute resource constraints.

Proposition 6.4. Let $\Phi_{\mathcal{B}}$ be given by the tractable instance (1) in Definition 6.2. Assume:

1. For each task t , the map $q \mapsto \psi_t(q, Q^*(t))$ is measurable, nondecreasing, and concave in q , with maximal local sensitivity in a neighbourhood of the threshold $Q^*(t)$.
2. The cost functional $\text{Cost}(r)$ is measurable and nonincreasing along any improvement of capability in the sense of Axiom (A2) (so that moving from ν to ν' with “non-increasing cost” implies $\mathbb{E}_{\nu'}[\text{Cost}(r)] \leq \mathbb{E}_{\nu}[\text{Cost}(r)]$).
3. Symmetries f of \mathcal{B} act by permuting tasks within families and applying resource rescalings that preserve the sets $\{F_k\}$, the weights w_k , the thresholds $Q^*(t)$, and the cost structure.

Then $\Phi_{\mathcal{B}}$ is an AAI functional in the sense of Definition 6.1 and satisfies axioms (A1)-(A4).

Proof. By construction, $\Phi_{\mathcal{B}}$ is a measurable map $\mathcal{P}(X_{\mathcal{B}}) \rightarrow \mathbb{R}$, since it is obtained by applying measurable functions $(\psi_t, \bar{q}, \text{Cost})$ and finite sums, expectations, and a finite variance operator to the underlying law ν . It remains to verify axioms (A1)-(A4).

(A1) Naturality. Let f be a symmetry of \mathcal{B} with pushforward $f_* : \mathcal{P}(X_{\mathcal{B}}) \rightarrow \mathcal{P}(X_{\mathcal{B}'})$. By assumption (iii) f acts by permuting tasks within families and applying resource rescalings that preserve the family partition $\{F_k\}$, the weights w_k , the thresholds $Q^*(t)$, and the cost structure.

Hence, under f , the collection of random variables $\{q(t), Q^*(t), r\}_{t \in T}$ is mapped to a relabelled copy with the same joint law. In particular, for every k ,

$$\frac{1}{|F_k|} \sum_{t \in F_k} \psi_t(q(t), Q^*(t)), \quad \mathbb{E}_\nu[\bar{q}(F_k)], \quad \text{Cost}(r)$$

are invariant in distribution under f , and so are any finite linear combinations and the variance across families. Therefore $\Phi_{\mathcal{B}'}(f_*\nu) = \Phi_{\mathcal{B}}(\nu)$, which is exactly naturality.

(A2) Monotonicity. Suppose that ν' dominates ν in increasing concave order with respect to the success indicators, and that the associated resource usage has non-increasing expected cost $\mathbb{E}_{\nu'}[\text{Cost}(r)] \leq \mathbb{E}_\nu[\text{Cost}(r)]$. By definition of increasing concave order, for every bounded, increasing, concave function ψ we have

$$\mathbb{E}_{\nu'}[\psi(q(t))] \geq \mathbb{E}_\nu[\psi(q(t))] \quad \text{for all tasks } t.$$

In particular, taking $\psi(\cdot) = \psi_t(\cdot, Q^*(t))$ and using assumption (i) that $q \mapsto \psi_t(q, Q^*(t))$ is increasing and concave, we obtain

$$\mathbb{E}_{\nu'}\left[\frac{1}{|F_k|} \sum_{t \in F_k} \psi_t(q(t), Q^*(t))\right] \geq \mathbb{E}_\nu\left[\frac{1}{|F_k|} \sum_{t \in F_k} \psi_t(q(t), Q^*(t))\right]$$

for every family F_k . Hence the first term in $\Phi_{\mathcal{B}}$ is weakly larger under ν' than under ν .

As improvements in the $q(t)$ under ν' are assumed not to worsen the dispersion of family means, we have

$$\text{Var}_k(\mathbb{E}_{\nu'}[\bar{q}(F_k)]) \leq \text{Var}_k(\mathbb{E}_\nu[\bar{q}(F_k)]),$$

so the dispersion penalty $-\lambda \text{Var}_k(\cdot)$ is also weakly larger under ν' than under ν . Finally, the cost condition $\mathbb{E}_{\nu'}[\text{Cost}(r)] \leq \mathbb{E}_\nu[\text{Cost}(r)]$ implies

$$-\gamma \mathbb{E}_{\nu'}[\text{Cost}(r)] \geq -\gamma \mathbb{E}_\nu[\text{Cost}(r)].$$

Combining these three inequalities yields $\Phi_{\mathcal{B}}(\nu') \geq \Phi_{\mathcal{B}}(\nu)$.

(A3) Threshold calibration. Fix a task t and consider the effect of increasing $\mathbb{P}_\nu\{q(t) \geq Q^*(t)\}$ while holding the other tasks fixed. By assumption (i), the function $q \mapsto \psi_t(q, Q^*(t))$ is nondecreasing and has its largest local slope in a neighbourhood of $Q^*(t)$. Thus increasing the success probability near $Q^*(t)$ strictly increases the contribution of task t to the first term in (1), and the marginal effect is maximised when the current success probability is close to $Q^*(t)$. Since the dispersion and cost terms depend on family means and on resources, not on local perturbations of a single task at fixed cost, their contribution is negligible for such infinitesimal changes. Hence $\Phi_{\mathcal{B}}(\nu)$ is strictly increasing in $\mathbb{P}_\nu\{q(t) \geq Q^*(t)\}$, with highest sensitivity near the threshold.

(A4) Generality. The first term in (1) is a weighted sum of familywise averages with symmetric treatment of tasks within each family; any permutation of tasks that preserves the family partition leaves it unchanged. The second term penalises dispersion of the family means $\{\mathbb{E}_\nu[\bar{q}(F_k)]\}_k$ via the variance across k , thereby discouraging unfair concentration of performance on a small subset of families. The cost term treats resources through $\text{Cost}(r)$ without privileging any particular task family. Together, these design choices implement axiom (A4): family means enter symmetrically and the dispersion penalty discourages unfair concentration.

Combining the four parts shows that the tractable functional $\Phi_{\mathcal{B}}$ in (1) is an AAI functional satisfying axioms (A1)-(A4) under assumptions (i)-(iii). \square

Remark 6.5 (Dual and risk-sensitive variants). The tractable AAI functional in (1) is concave in the success indicators and admits a standard Fenchel-type dual representation in terms of linear scores and the concave conjugates ψ_t^* (see, e.g., [3]). One can also obtain risk-sensitive variants by applying entropic or other convex transforms to the scalar base score. We do not pursue these functional-analytic aspects here, focusing instead on the geometric and probabilistic structure of the moduli space.

6.1 Determinacy from dense agreement on batteries and laws

We now define the metric structure on the moduli space and prove that a regular functional is determined by its values on a countable dense subset. This provides the rigorous justification for using finite test panels to certify general intelligence.

Definition 6.6 (Canonical metric on the moduli). Write the canonical representative of a battery as $(C_u, \tau, [r])$, where C_u is the copula of PIT scores, τ the threshold vector, and $[r]$ the resource ray. Fix weights $\alpha, \beta, \gamma > 0$ and define

$$d_{\mathfrak{M}}((C_u, \tau, [r]), (C'_u, \tau', [r'])) := \alpha W_1(C_u, C'_u) + \beta \|\tau - \tau'\|_1 + \gamma d_{\text{ray}}([r], [r']).$$

Here W_1 is the 1-Wasserstein distance on $[0, 1]^T$, $\|\cdot\|_1$ is extended absolutely to countable T , and d_{ray} is any standard projective metric on rays in $\mathbb{R}_+^{d_R}$.

Definition 6.7 (Canonical pushforward of a law). For a battery \mathcal{B} with canonical PIT map u and resource readout r , let

$$S_{\mathcal{B}} : X_{\mathcal{B}} \longrightarrow [0, 1]^T \times \mathbb{R}_+^{d_R}, \quad x \mapsto (u(t)(x))_{t \in T} \oplus r(x).$$

For $\nu \in \mathcal{P}(X_{\mathcal{B}})$, write $\mu_{\mathcal{B}, \nu} := (S_{\mathcal{B}})_* \nu$.

Definition 6.8 (Pair metric). Let $\mathcal{P}^\sharp := \{(\mathcal{B}, \nu) : \nu \in \mathcal{P}(X_{\mathcal{B}})\}$. Define

$$d_\sharp((\mathcal{B}, \nu), (\mathcal{B}', \nu')) := \alpha W_1(\mu_{\mathcal{B}, \nu}, \mu_{\mathcal{B}', \nu'}) + \beta \|\tau - \tau'\|_1 + \gamma d_{\text{ray}}([r], [r']).$$

Definition 6.9 (Simple pairs). Let $\mathfrak{M}_{\text{simp}} \subset \mathfrak{M}$ be the batteries with (i) finite T ; (ii) $\tau(t) \in \mathbb{Q} \cap [0, 1]$ for all t ; (iii) C_u finitely supported on a rational grid in $[0, 1]^T$; (iv) $[r]$ represented by a rational vector in $\mathbb{Q}_+^{d_R}$. Let $\mathcal{P}_{\text{simp}}^\sharp \subset \mathcal{P}^\sharp$ be the pairs (\mathcal{B}, ν) with $\mathcal{B} \in \mathfrak{M}_{\text{simp}}$ and $\mu_{\mathcal{B}, \nu}$ finitely supported on a rational grid in $[0, 1]^T \times \mathbb{Q}_+^{d_R}$.

Lemma 6.10 (Countable density). $\mathfrak{M}_{\text{simp}}$ is countable and dense in $(\mathfrak{M}, d_{\mathfrak{M}})$, and $\mathcal{P}_{\text{simp}}^\sharp$ is dense in $(\mathcal{P}^\sharp, d_\sharp)$.

Proof. For any measure μ on the compact metric space $[0, 1]^T$, there exists a sequence of measures μ_n with finite support on rational grids such that $W_1(\mu_n, \mu) \rightarrow 0$ (standard quantization, e.g. [5] Lemma 6.18). Similarly, thresholds τ and rays $[r]$ can be approximated by rational vectors. Since d_\sharp is a weighted sum of these metrics, the product of dense subsets is dense. \square

Definition 6.11 (Regular families of AAI functionals). A family $\{\Phi_{\mathcal{B}}\}_{\mathcal{B}}$ of AAI functionals (Definition 6.1) is regular if there exists $L < \infty$ such that for all pairs

$$|\Phi_{\mathcal{B}}(\nu) - \Phi_{\mathcal{B}'}(\nu')| \leq L d_\sharp((\mathcal{B}, \nu), (\mathcal{B}', \nu')).$$

Theorem 6.12 (Determinacy from dense agreement). *Let $\{\Phi_{\mathcal{B}}\}$ and $\{\Psi_{\mathcal{B}}\}$ be regular families. If they agree on $\mathcal{P}_{\text{simp}}^{\sharp}$, they agree everywhere.*

Proof. Fix $(\mathcal{B}, \nu) \in \mathcal{P}^{\sharp}$ and choose a sequence $(\mathcal{B}_n, \nu_n) \in \mathcal{P}_{\text{simp}}^{\sharp}$ with $d_{\sharp}((\mathcal{B}_n, \nu_n), (\mathcal{B}, \nu)) \rightarrow 0$ (using Lemma 6.10). Regularity implies Lipschitz continuity:

$$|\Phi_{\mathcal{B}}(\nu) - \Phi_{\mathcal{B}_n}(\nu_n)| \leq L_{\Phi} d_{\sharp}((\mathcal{B}, \nu), (\mathcal{B}_n, \nu_n)),$$

and similarly for Ψ . Since $\Phi_{\mathcal{B}_n}(\nu_n) = \Psi_{\mathcal{B}_n}(\nu_n)$ on simple pairs, we have

$$|\Phi_{\mathcal{B}}(\nu) - \Psi_{\mathcal{B}}(\nu)| \leq |\Phi_{\mathcal{B}}(\nu) - \Phi_{\mathcal{B}_n}(\nu_n)| + |\Psi_{\mathcal{B}_n}(\nu_n) - \Psi_{\mathcal{B}}(\nu)| \leq (L_{\Phi} + L_{\Psi}) d_{\sharp}((\mathcal{B}, \nu), (\mathcal{B}_n, \nu_n)) \rightarrow 0.$$

Thus $\Phi_{\mathcal{B}}(\nu) = \Psi_{\mathcal{B}}(\nu)$. \square

Definition 6.13 (Order Regularity). A family is order regular if $\Phi_{\mathcal{B}_n}(\nu_n) \rightarrow \Phi_{\mathcal{B}}(\nu)$ whenever (\mathcal{B}_n, ν_n) is a monotone refining sequence (monotone convergence of canonical scores in convex order, thresholds, and rays).

Theorem 6.14 (Determinacy from monotone-class agreement). *Let $\{\Phi_{\mathcal{B}}\}$ and $\{\Psi_{\mathcal{B}}\}$ be natural, bounded, order-regular families. If $\Phi_{\mathcal{B}}(\nu) = \Psi_{\mathcal{B}}(\nu)$ for all simple pairs (\mathcal{B}, ν) (i.e. those where canonical scores are simple with respect to a finite partition), then $\Phi_{\mathcal{B}}(\nu) = \Psi_{\mathcal{B}}(\nu)$ for all $(\mathcal{B}, \nu) \in \mathcal{P}^{\sharp}$.*

Proof. Fix an arbitrary pair (\mathcal{B}, ν) . We construct a monotone approximating sequence using dyadic filtrations. Let U be the canonical score variable on $[0, 1]^T$. Define $U_n = \mathbb{E}[U | \mathcal{G}_n]$, where \mathcal{G}_n is the dyadic partition of the hypercube at scale 2^{-n} . We keep thresholds and resource rays fixed, so that (\mathcal{B}_n, ν_n) shares the same discrete structure, thresholds and rays as (\mathcal{B}, ν) , and only the canonical score law is replaced by that of U_n .

The variables U_n are simple (constant on finite partition cells). By Jensen's inequality for conditional expectations, $U_n \preceq_{cx} U_{n+1}$. By the Martingale Convergence Theorem, $U_n \rightarrow U$ in L^1 (hence W_1). Thus, any pair (\mathcal{B}, ν) is the limit of a monotone sequence of simple pairs (\mathcal{B}_n, ν_n) . Since the functionals agree on the sequence and are order regular, uniqueness of limits implies $\Phi_{\mathcal{B}}(\nu) = \Psi_{\mathcal{B}}(\nu)$. \square

Remark 6.15 (Operational corollary). Because $\mathcal{P}_{\text{simp}}^{\sharp}$ is countable and dense, a regular family of AAI functionals is uniquely determined by its values on a fixed countable catalog of simple batteries. In practice, one can calibrate Φ by tabulating this panel once.

7 The AAI Score as an AAI Functional

Let \mathcal{X} be axes with weights $w_x > 0$ and $W = \sum_x w_x$. The AAI-Index from AAI is the weighted geometric mean $\mathcal{C} = (\prod_{x \in \mathcal{X}} x^{w_x})^{1/W}$. Let $\pi_x : \mathcal{P}(X_{\mathcal{B}}) \rightarrow [0, 1]$ be the functional corresponding to the operational definition of axis x as defined in [1]. Specifically:

- For **Autonomy** ($x = A$), $\pi_A(\nu)$ computes the mean horizon-capped action count $\mathbb{E}_{\nu}[\min(a(t)/H, 1)]$ composed with calibration ϕ_A .
- For **Generality** ($x = G$), $\pi_G(\nu)$ calculates the fraction of task families F_i where the mean quality $\bar{q}(F_i)$ meets the family-specific threshold τ_i .
- For **Planning** ($x = P$), $\pi_P(\nu)$ averages the effective plan depths $d(t)$ normalized by the target depth anchor D for successful traces.
- For **Memory** ($x = M$), $\pi_M(\nu)$ aggregates the retention half-life derived from performance decay over time lags Δ and the immediate retrieval recall $\text{Rec}@K$.

- For **Tool Economy** ($x = T$), $\pi_T(\nu)$ computes the geometric mean of tool category coverage, success under drift perturbations δ , and logarithmic discovery size.
- For **Self-Revision** ($x = R$), $\pi_R(\nu)$ aggregates the autonomy-weighted difference-in-differences $\rho_r \Delta C_r$ (capability gain over control) for self-initiated patches.
- For **Sociality** ($x = S$), $\pi_S(\nu)$ measures the performance lift of multi-agent configurations over the single-agent baseline, penalized by deadlock and chatter indicators.
- For **Embodiment** ($x = E$), $\pi_E(\nu)$ forms the geometric mean of real-world actuation reliability, safety incident rates (weighted by severity), and sim-to-real transfer agreement.
- For **World-Model** ($x = W$), $\pi_W(\nu)$ computes the probabilistic calibration via the Brier score of agent predictions relative to a reference baseline.
- For **Economics** ($x = \$$), $\pi_{\$}(\nu)$ calculates the ratio of quality-adjusted throughput (TPH_{Q^*}) to monetary cost (CPH).

Define

$$\Phi_{\mathcal{B}}^{\text{geom}}(\nu) = \exp\left(\frac{1}{W} \sum_{x \in \mathcal{X}} w_x \mathbb{E}_{\nu}[\log \pi_x]\right). \quad (2)$$

Proposition 7.1. $\Phi_{\mathcal{B}}^{\text{geom}}$ satisfies the axioms and for $\nu = \delta_x$ equals the AAI-Index with x 's axis values.

Proof. Each $\log \pi_x$ increases with axis performance; the average preserves symmetry across axes; exponentiation preserves order. For deterministic x , (2) reduces to the geometric mean. \square

Remark 7.2 (Embedding into the tractable instance). Take $\psi_t(q, Q^*) = \log \pi_x$ for tasks in axis x and set $\lambda = \gamma = 0$; grouping tasks by axes gives (2).

7.1 Regularity of the AAI score

We now show that our axiomatic AAI functional is stable with respect to the moduli space metric d_{\sharp} . We first prove a general Lipschitz-regularity result for geometric aggregators built from axis functionals, and then check that the concrete AAI score of [1] satisfies its assumptions.

Theorem 7.3 (General regularity of geometric AAI scores). *Assume each axis map π_x is L_x -Lipschitz (as a function of the canonical variables) and bounded away from zero, $\pi_x \geq \varepsilon > 0$. Define the geometric aggregator*

$$\Phi_{\mathcal{B}}^{\text{geom}}(\nu) = \exp\left(\frac{1}{W} \sum_{x \in \mathcal{X}} w_x \mathbb{E}_{\mu_{\mathcal{B}, \nu}}[\log \pi_x]\right), \quad W := \sum_{x \in \mathcal{X}} w_x.$$

Then the family $\{\Phi_{\mathcal{B}}^{\text{geom}}\}$ is Lipschitz-regular with respect to d_{\sharp} .

Proof. Since $\pi_x \in [\varepsilon, 1]$, the map $u \mapsto \log u$ has derivative bounded by $1/\varepsilon$ on $[\varepsilon, 1]$. By the chain rule, the composition $\log \pi_x$ is (L_x/ε) -Lipschitz in the canonical variables. By Kantorovich–Rubinstein duality, we obtain

$$|\mathbb{E}_{\mu_{\mathcal{B}, \nu}}[\log \pi_x] - \mathbb{E}_{\mu_{\mathcal{B}', \nu'}}[\log \pi_x]| \leq \frac{L_x}{\varepsilon} W_1(\mu_{\mathcal{B}, \nu}, \mu_{\mathcal{B}', \nu'}).$$

Multiplying by w_x and averaging over x gives

$$|\Phi_{\mathcal{B}}^{\text{geom}}(\nu) - \Phi_{\mathcal{B}'}^{\text{geom}}(\nu')| \leq \frac{1}{W} \sum_{x \in \mathcal{X}} w_x \frac{L_x}{\varepsilon} W_1(\mu_{\mathcal{B}, \nu}, \mu_{\mathcal{B}', \nu'}).$$

By definition of d_{\sharp} , there exists a constant $\alpha > 0$ such that

$$W_1(\mu_{\mathcal{B}, \nu}, \mu_{\mathcal{B}', \nu'}) \leq \alpha^{-1} d_{\sharp}((\mathcal{B}, \nu), (\mathcal{B}', \nu')).$$

Combining the two displays yields

$$|\Phi_{\mathcal{B}}^{\text{geom}}(\nu) - \Phi_{\mathcal{B}'}^{\text{geom}}(\nu')| \leq \frac{1}{\alpha W_{\varepsilon}} \left(\sum_{x \in \mathcal{X}} w_x L_x \right) d_{\sharp}((\mathcal{B}, \nu), (\mathcal{B}', \nu')),$$

so $\{\Phi_{\mathcal{B}}^{\text{geom}}\}$ is Lipschitz-regular with respect to d_{\sharp} . \square

Corollary 7.4 (Regularity of the AAI score of [1]). *Let $\Phi_{\mathcal{B}}^{\text{AAI}}(\nu)$ denote the AAI score of [1], given by the geometric aggregator in Eq. (2). Under the design assumptions of [1] (in particular, that each axis map π_x is a Lipschitz function of the canonical evaluation variables and is clipped below at $\varepsilon > 0$), the family $\{\Phi_{\mathcal{B}}^{\text{AAI}}\}$ is Lipschitz-regular with respect to d_{\sharp} .*

Proof. By construction in [1], each axis map π_x is implemented as a finite composition of Lipschitz operations (affine transforms, thresholding, min/max, and smooth aggregations) applied to the canonical scores and resource variables. Hence each π_x is L_x -Lipschitz for some finite L_x . The clipping step $\pi_x^{(\varepsilon)} = \max\{\varepsilon, \pi_x\}$ ensures $\pi_x^{(\varepsilon)} \geq \varepsilon > 0$ for all inputs. Thus the hypotheses of Theorem 7.3 are satisfied with π_x replaced by $\pi_x^{(\varepsilon)}$, and the geometric AAI score of [1] is Lipschitz-regular with respect to d_{\sharp} . \square

8 Cognitive Cores and the Score AAI_{core}

In [1] a concrete AAI_{core} score is defined following [2] by equal-weight aggregation over a fixed collection of CHC-style domains $\mathcal{CC} = \{\text{Gc}, \text{Grw}, \dots, \text{Gwm}, \text{Gls}, \text{Glr}\}$, referred to there as the *cognitive core*. In the present framework this core score corresponds to a functional $\Psi_{\mathcal{B}}$ on the core factor $X_{\mathcal{B}}^{\text{core}}$, and the maps p_{cc} formalize the projection from full, interface-rich batteries to their underlying cognitive core. The following extension theory describes all ways of lifting such a given core score to evaluations on full batteries.

Given a measurable map $p : X \rightarrow Y$ and $\nu \in \mathcal{P}(X)$, we write $p_{\#}\nu \in \mathcal{P}(Y)$ for the pushforward measure:

$$p_{\#}\nu(A) := \nu(p^{-1}(A)), \quad A \in \mathcal{B}(Y).$$

Definition 8.1 (Cognitive core). A cognitive core for \mathcal{B} is a factor $p_{\text{cc}} : X_{\mathcal{B}} \rightarrow X_{\mathcal{B}}^{\text{core}}$ with sigma-algebra \mathcal{C} such that, for every task t , $\mathbb{1}\{q(t) \geq Q^*(t)\}$ is \mathcal{C} -measurable and \mathcal{C} is minimal with this property (up to null sets). Write $\nu^{\text{core}} := p_{\text{cc}}_{\#}\nu$.

Definition 8.2 (Core score). Let $p_{\text{cc}} : X_{\mathcal{B}} \rightarrow X_{\mathcal{B}}^{\text{core}}$ be the core projection and $\nu^{\text{core}} := p_{\text{cc}}_{\#}\nu$ the induced core law. Fix a measurable lifting map $\mathcal{L} : \mathcal{P}(X_{\mathcal{B}}^{\text{core}}) \rightarrow \mathcal{P}(X_{\mathcal{B}})$ such that $p_{\text{cc}}_{\#} \circ \mathcal{L} = \text{id}$. Given an AAI functional $\Phi_{\mathcal{B}} : \mathcal{P}(X_{\mathcal{B}}) \rightarrow \mathbb{R}$, define the *core AAI score* by

$$\text{AAI}_{\text{core}, \mathcal{B}}(\nu) := \Phi_{\mathcal{B}}(\mathcal{L}(\nu^{\text{core}})).$$

This definition can equivalently be viewed as evaluating the induced core functional $\Psi_{\mathcal{B}}(\nu^{\text{core}})$ obtained by restriction of $\Phi_{\mathcal{B}}$ to the image of \mathcal{L} .

Remark 8.3. In applications, we typically choose \mathcal{L} as the maximum-entropy lift consistent with the core marginals.

Proposition 8.4 (Identifiability of cores). *The cognitive core is unique up to isomorphism. Specifically, if two core maps $p_{CC} : X_{\mathcal{B}} \rightarrow X_{\mathcal{B}}^{\text{core}}$ and $p'_{CC} : X_{\mathcal{B}} \rightarrow X'_{\mathcal{B}}^{\text{core}}$ generate the same threshold-indicator sigma-algebra almost surely, and both core sigma-algebras are minimal with this property, then there exists a measurable isomorphism $h : X_{\mathcal{B}}^{\text{core}} \rightarrow X'_{\mathcal{B}}^{\text{core}}$ such that $p'_{CC} = h \circ p_{CC}$ almost surely.*

Proof. By definition of the core (Definition 8.1), the sigma-algebra $\mathcal{C} := \sigma(p_{CC})$ is the minimal sigma-algebra making all threshold indicators $\mathbf{1}\{q(t) \geq Q^*(t)\}$ measurable, and similarly for $\mathcal{C}' := \sigma(p'_{CC})$. If the two cores generate the same threshold-indicator sigma-algebra almost surely and are both minimal, then \mathcal{C} and \mathcal{C}' coincide up to null sets. The corresponding factor spaces $(X_{\mathcal{B}}^{\text{core}}, \mathcal{C})$ and $(X'_{\mathcal{B}}^{\text{core}}, \mathcal{C}')$ are therefore isomorphic as measurable spaces: one may define h on $X_{\mathcal{B}}^{\text{core}}$ by sending each p_{CC} -fiber to the corresponding p'_{CC} -fiber, which is well defined up to null sets, and standard measure-theoretic arguments show that h is measurable with measurable inverse. By construction $p'_{CC} = h \circ p_{CC}$ almost surely. \square

8.1 Given Core Scores and Continuations

We formalize the relationship between the "true" core score and the observable score using extension theory. This isolates "interface artifacts" from "core capability."

Definition 8.5 (Given core score and continuations). Fix a battery \mathcal{B} with factor $p_{CC} : X_{\mathcal{B}} \rightarrow X_{\mathcal{B}}^{\text{core}}$ and core sigma-algebra \mathcal{C} . Let the given core score be the functional

$$\Psi_{\mathcal{B}} : \mathcal{P}(X_{\mathcal{B}}^{\text{core}}) \rightarrow \mathbb{R}, \quad \Psi_{\mathcal{B}}(\nu^{\text{core}}) = \text{AAI}_{\text{core}, \mathcal{B}}(\nu),$$

where $\nu^{\text{core}} := p_{CC} \# \nu$. A continuation of $\Psi_{\mathcal{B}}$ is a functional $\Phi_{\mathcal{B}} : \mathcal{P}(X_{\mathcal{B}}) \rightarrow \mathbb{R}$ such that

$$\Phi_{\mathcal{B}}(\nu) = \Psi_{\mathcal{B}}(\nu^{\text{core}}) \quad \text{whenever } \nu \text{ coincides with the canonical lift, i.e., } \nu = \mathcal{L}(\nu^{\text{core}}).$$

Definition 8.6 (Core sufficiency). A continuation $\Phi_{\mathcal{B}}$ is *core-sufficient* if it depends only on core information, i.e.

$$\Phi_{\mathcal{B}}(\nu) = \Phi_{\mathcal{B}}(\nu') \quad \text{whenever } \nu^{\text{core}} = (\nu')^{\text{core}}.$$

Lemma 8.7 (Minimal continuation). *The assignment*

$$\Phi_{\mathcal{B}}^{\min}(\nu) := \Psi_{\mathcal{B}}(\nu^{\text{core}})$$

is a continuation. It is the unique core-sufficient continuation.

Proof. Well-defined by composition with p_{CC} ; if ν is a canonical lift, then $\nu = \mathcal{L}(\nu^{\text{core}})$, hence $\Phi_{\mathcal{B}}^{\min}(\nu) = \Psi_{\mathcal{B}}(\nu^{\text{core}})$ by definition. If $\Phi_{\mathcal{B}}$ is core-sufficient and a continuation, then for any ν , using the sufficiency property, we have $\Phi_{\mathcal{B}}(\nu) = \Phi_{\mathcal{B}}(\mathcal{L}(\nu^{\text{core}})) = \Psi_{\mathcal{B}}(\nu^{\text{core}}) = \Phi_{\mathcal{B}}^{\min}(\nu)$. \square

Proposition 8.8 (Dominance of Extensions). *Let $\nu \in \mathcal{P}(X_{\mathcal{B}})$. If ν dominates the canonical lift $\mathcal{L}(\nu^{\text{core}})$ in the sense of Axiom (A2) (i.e., ν represents a capability improvement over the baseline lift), then:*

$$\text{AAI}_{\text{core}, \mathcal{B}}(\nu) = \Phi_{\mathcal{B}}^{\min}(\nu) \leq \Phi_{\mathcal{B}}(\nu).$$

Equality holds if and only if $\Phi_{\mathcal{B}}(\nu) = \Phi_{\mathcal{B}}(\mathcal{L}(\nu^{\text{core}}))$, i.e., the specific fiber structure of ν yields no score advantage over the lift.

Proof. By Lemma 8.7, $\Phi_{\mathcal{B}}^{\min}(\nu)$ is the value of the core score functional $\Psi_{\mathcal{B}}$ on ν^{core} . By Definition 8.2, this is defined as $\Phi_{\mathcal{B}}(\mathcal{L}(\nu^{\text{core}}))$. Since ν dominates $\mathcal{L}(\nu^{\text{core}})$ by hypothesis, the Restricted Monotonicity axiom (A2) implies $\Phi_{\mathcal{B}}(\nu) \geq \Phi_{\mathcal{B}}(\mathcal{L}(\nu^{\text{core}}))$. \square

Definition 8.9 (Non-core invariants). A measurable random variable $Z : X_{\mathcal{B}} \rightarrow \mathbb{R}$ is a *non-core invariant* if

$$\mathbb{E}[Z | \mathcal{C}] = 0 \quad \text{and} \quad Z \text{ is invariant under the evaluation-preserving symmetry group } G.$$

Write $\mathcal{I}_{nc}(\mathcal{B})$ for the linear space of such invariants with $\mathbb{E}|Z| < \infty$.

Theorem 8.10 (Additive decomposition of continuations). *Any continuation $\Phi_{\mathcal{B}}$ that is natural with respect to G can be written as*

$$\Phi_{\mathcal{B}}(\nu) = \Phi_{\mathcal{B}}^{\min}(\nu) + F_{\mathcal{B}}(\nu),$$

where $F_{\mathcal{B}}$ is G -invariant and vanishes on canonical lifts: $F_{\mathcal{B}}(\nu) = 0$ whenever $\nu = \mathcal{L}(\nu^{\text{core}})$. Conversely, any such $F_{\mathcal{B}}$ yields a valid continuation when added to $\Phi_{\mathcal{B}}^{\min}$.

Proof. Define $F_{\mathcal{B}}(\nu) := \Phi_{\mathcal{B}}(\nu) - \Phi_{\mathcal{B}}^{\min}(\nu)$. Then $F_{\mathcal{B}} \equiv 0$ on lifted laws by the continuation property. Naturality is inherited. The converse is immediate. \square

Proposition 8.11 (Parametric families and cardinality). *Fix $Z_1, \dots, Z_m \in \mathcal{I}_{nc}(\mathcal{B})$. For any $\theta \in \mathbb{R}^m$ set*

$$\Phi_{\mathcal{B}}^{\theta}(\nu) := \Phi_{\mathcal{B}}^{\min}(\nu) + \sum_{j=1}^m \theta_j \mathbb{E}_{\nu}[Z_j].$$

Then $\Phi_{\mathcal{B}}^{\theta}$ is a continuation that is natural with respect to G . If there exists a nonzero $Z \in \mathcal{I}_{nc}(\mathcal{B})$, then the set of continuations has the cardinality of the continuum.

Proof. Since $\mathbb{E}[Z_j | \mathcal{C}] = 0$, for any lift $\nu = \mathcal{L}(\mu)$ we have $\mathbb{E}_{\nu}[Z_j] = 0$ (assuming \mathcal{L} respects the conditional expectation structure, e.g., max-entropy). Thus $\Phi_{\mathcal{B}}^{\theta}$ agrees with $\Phi_{\mathcal{B}}^{\min}$ on the core. G -invariance follows from the invariance of Z_j . Varying θ gives uncountably many distinct functionals when some $Z \neq 0$. \square

Corollary 8.12 (Uniqueness under core sufficiency). *If continuations are restricted to be core-sufficient, then $\Phi_{\mathcal{B}} = \Phi_{\mathcal{B}}^{\min}$ is the unique continuation.*

Lemma 8.13 (Calibration of finite families). *Let $Z_1, \dots, Z_m \in \mathcal{I}_{nc}(\mathcal{B})$ and choose reference laws $\nu^{(1)}, \dots, \nu^{(M)}$. Suppose the $M \times m$ matrix with entries $[\mathbb{E}_{\nu^{(i)}}(Z_j)]$ has full column rank. Then the parameters θ in $\Phi_{\mathcal{B}}^{\theta}$ can be identified uniquely from M linear calibration conditions on $\Phi_{\mathcal{B}}^{\theta}(\nu^{(i)})$.*

Proof. From $\Phi_{\mathcal{B}}^{\theta}(\nu^{(i)}) - \Phi_{\mathcal{B}}^{\min}(\nu^{(i)}) = \sum_j \theta_j \mathbb{E}_{\nu^{(i)}}[Z_j]$ we obtain a full-rank linear system for θ . \square

Proposition 8.14 (Lower and upper envelopes). *Let \mathcal{F} be any class of G -invariant $F_{\mathcal{B}}$ that vanish on canonical lifts and satisfy given monotonicity or resource constraints. Define*

$$\Phi_{\mathcal{B}}^{\inf}(\nu) := \Phi_{\mathcal{B}}^{\min}(\nu) + \inf_{F \in \mathcal{F}} F_{\mathcal{B}}(\nu), \quad \Phi_{\mathcal{B}}^{\sup}(\nu) := \Phi_{\mathcal{B}}^{\min}(\nu) + \sup_{F \in \mathcal{F}} F_{\mathcal{B}}(\nu).$$

Then for any admissible continuation $\Phi_{\mathcal{B}}$ with $F_{\mathcal{B}} \in \mathcal{F}$,

$$\Phi_{\mathcal{B}}^{\inf}(\nu) \leq \Phi_{\mathcal{B}}(\nu) \leq \Phi_{\mathcal{B}}^{\sup}(\nu) \quad \text{for all } \nu.$$

Remark 8.15 (Categorical view: right Kan extension). Let $U : \mathbf{Bat} \rightarrow \mathbf{Core}$ send each battery to its core. The minimal continuation $\Phi^{\min} = \Psi \circ U$ is the right Kan extension of Ψ along U . Non-core terms $F_{\mathcal{B}}$ are precisely G -invariant functionals on fibers that vanish on canonical lifts.

9 Statistical Guarantees and Stability

To certify that the AAI score is robust and practically estimable from finite data, we provide concentration bounds and stability guarantees under battery perturbations.

Theorem 9.1 (Finite-Sample Concentration). *Let $\widehat{\Phi}_{\mathcal{B}}$ be the plug-in estimator of the tractable instance (1) computed over n independent seeds per task. Assume the component functions ψ_t are L -Lipschitz and bounded on $[0, 1]$, and that resources and weights are uniformly bounded and normalized so that the effective per-task weights $w_k/|F_k|$ are $O(1/|T|)$. Then for any failure probability $\delta \in (0, 1)$,*

$$\mathbb{P}\left(|\widehat{\Phi}_{\mathcal{B}}(\rho) - \Phi_{\mathcal{B}}(\rho)| > L\sqrt{\frac{C \log(2/\delta)}{n|T|}} + \lambda \frac{K}{\sqrt{n}}\right) \leq \delta, \quad (3)$$

for constants $C, K > 0$ depending only on the weights, the number of families, and the bounds on scores and resources.

Proof. We write $\widehat{\Phi}_{\mathcal{B}}$ as the sum of a “linear” part and a variance penalty:

$$\widehat{\Phi}_{\mathcal{B}} = \widehat{\Phi}_{\text{lin}} - \lambda \widehat{\text{Var}}_k(\widehat{m}_k),$$

where

$$\widehat{\Phi}_{\text{lin}} := \sum_k w_k \frac{1}{|F_k|} \sum_{t \in F_k} \widehat{\mathbb{E}}[\psi_t(q(t), Q^*(t))] - \gamma \widehat{\mathbb{E}}[\text{Cost}(r)], \quad \widehat{m}_k := \widehat{\mathbb{E}}[\bar{q}(F_k)].$$

Here $\widehat{\mathbb{E}}[\cdot]$ denotes the corresponding empirical average.

Step 1: Linear term via McDiarmid. Index the $N = n|T|$ independent seed draws as $(s_{t,i})_{t \in T, 1 \leq i \leq n}$ and regard $\widehat{\Phi}_{\text{lin}}$ as a function f of these N variables. Fix a pair (t_0, i_0) and change only the seed s_{t_0, i_0} . This affects only the empirical average

$$\widehat{\mathbb{E}}[\psi_{t_0}(q(t_0), Q^*(t_0))] = \frac{1}{n} \sum_{i=1}^n \psi_{t_0}(q(t_0; s_{t_0, i}), Q^*(t_0)),$$

changing at most one summand. Since $q(t_0; s) \in [0, 1]$ and $\psi_{t_0}(\cdot, Q^*(t_0))$ is L -Lipschitz in its first argument, the difference in that summand is at most L , so the empirical average changes by at most L/n .

In the global functional $\widehat{\Phi}_{\text{lin}}$, this term is weighted by $w_k/|F_k|$. By the normalization assumption on the weights, there exists a constant $C_1 > 0$ such that the sensitivity to a single seed is

$$|\widehat{\Phi}_{\text{lin}}(s) - \widehat{\Phi}_{\text{lin}}(s')| \leq \frac{C_1}{n|T|} = \frac{C_1}{N}.$$

Applying McDiarmid’s inequality with sensitivity $\Delta = C_1/N$: The sum of squared differences is $\sum_{i=1}^N \Delta^2 = N(C_1/N)^2 = C_1^2/N$. Thus, for any $\varepsilon > 0$:

$$\mathbb{P}(|\widehat{\Phi}_{\text{lin}} - \mathbb{E}\widehat{\Phi}_{\text{lin}}| > \varepsilon) \leq 2 \exp\left(-\frac{2\varepsilon^2}{C_1^2/N}\right) = 2 \exp\left(-\frac{2N\varepsilon^2}{C_1^2}\right).$$

Setting the right-hand side equal to $\delta/2$ and solving for ε gives:

$$|\widehat{\Phi}_{\text{lin}} - \mathbb{E}\widehat{\Phi}_{\text{lin}}| \leq \frac{C_1}{\sqrt{2N}} \sqrt{\log \frac{4}{\delta}} = L \sqrt{\frac{C \log(2/\delta)}{n|T|}}$$

for a suitable constant C , matching the theorem statement. Since $\widehat{\Phi}_{\text{lin}}$ is a plug-in estimator of sums of expectations, we have $\mathbb{E}\widehat{\Phi}_{\text{lin}} = \Phi_{\text{lin}}$, so the same bound holds for $|\widehat{\Phi}_{\text{lin}} - \Phi_{\text{lin}}|$ with probability at least $1 - \delta/2$.

Step 2: Variance penalty. For each family F_k , the empirical mean \widehat{m}_k is an average of $n|F_k|$ bounded variables and therefore satisfies, by Hoeffding's inequality,

$$\mathbb{P}(|\widehat{m}_k - m_k| > \varepsilon) \leq 2 \exp(-2n|F_k|\varepsilon^2) \leq 2 \exp(-2n\varepsilon^2), \quad m_k := \mathbb{E}_\nu[\bar{q}(F_k)].$$

Applying a union bound over the K' families and setting the right-hand side to $\delta/2$ yields, for

$$\varepsilon = \sqrt{\frac{1}{2n} \log \frac{4K'}{\delta}},$$

that

$$\mathbb{P}\left(\max_k |\widehat{m}_k - m_k| > \varepsilon\right) \leq \frac{\delta}{2}.$$

The variance functional

$$v(m_1, \dots, m_{K'}) := \text{Var}_k(m_k)$$

is Lipschitz on $[0, 1]^{K'}$ (with K' the number of families): changing a single coordinate m_k by δ changes $v(m)$ by at most $C_2|\delta|$ for some constant $C_2 > 0$. Consequently, on the event from Hoeffding's bound,

$$|\widehat{\text{Var}}_k(\widehat{m}_k) - \text{Var}_k(m_k)| \leq C_2 \max_k |\widehat{m}_k - m_k| \leq C_2 \sqrt{\frac{1}{2n} \log \frac{4K'}{\delta}}.$$

Multiplying by λ and absorbing constants gives a bound of the form $\lambda K/\sqrt{n}$ for some $K > 0$, with probability at least $1 - \delta/2$.

Step 3: Combination. Combining the two parts, we have

$$|\widehat{\Phi}_{\mathcal{B}} - \Phi_{\mathcal{B}}| \leq |\widehat{\Phi}_{\text{lin}} - \Phi_{\text{lin}}| + \lambda |\widehat{\text{Var}}_k(\widehat{m}_k) - \text{Var}_k(m_k)|.$$

By Step 1, the first term is bounded by $L\sqrt{C \log(2/\delta)/(n|T|)}$ with probability at least $1 - \delta/2$. By Step 2, the second term is bounded by $\lambda K/\sqrt{n}$ with probability at least $1 - \delta/2$. A union bound then implies that both inequalities hold simultaneously with probability at least $1 - \delta$, which is exactly the claimed concentration bound (3). \square

Proposition 9.2 (Stability under drift). *Assume $\{\Phi_{\mathcal{B}}\}$ is a regular family with Lipschitz modulus L_Φ with respect to d_\sharp . If the score copula changes by ε in 1-Wasserstein distance and thresholds Q^* shift by at most δ in the ℓ_∞ norm, then*

$$|\Delta\Phi_{\mathcal{B}}| \leq L_\Phi(\alpha \varepsilon + \beta |T| \delta) + \gamma B |\Delta\text{scale}|,$$

where α, β, γ are the weights from the definition of d_\sharp , B is the resource bound, and $|\Delta\text{scale}|$ denotes the induced change in the resource ray.

Proof. For any two pairs (\mathcal{B}, ν) and (\mathcal{B}', ν') , regularity (Definition 6.11) gives

$$|\Phi_{\mathcal{B}}(\nu) - \Phi_{\mathcal{B}'}(\nu')| \leq L_\Phi d_\sharp((\mathcal{B}, \nu), (\mathcal{B}', \nu')).$$

By Definition 6.8,

$$d_\sharp((\mathcal{B}, \nu), (\mathcal{B}', \nu')) = \alpha W_1(C_u, C'_u) + \beta \|\tau - \tau'\|_1 + \gamma d_{\text{ray}}([r], [r']).$$

If the copulas differ by at most ε in W_1 , $W_1(C_u, C'_u) \leq \varepsilon$. If the thresholds shift by at most δ in ℓ_∞ , then $\|\tau - \tau'\|_1 \leq |T|\delta$. Finally, $d_{\text{ray}}([r], [r'])$ is controlled by the change in resource scale; under the resource bound B this yields $d_{\text{ray}}([r], [r']) \leq B |\Delta\text{scale}|$. Substituting these bounds into the expression for d_\sharp and applying the Lipschitz estimate gives the stated inequality. \square

10 Operational Implementation

This section translates the theoretical framework into concrete evaluation protocols, computational procedures, and extension mechanisms.

10.1 Estimation Procedures

Procedure 1: The AAI-Index (Φ^{geom}). To estimate the geometric composite score:

1. For each axis $x \in \mathcal{X}$, compute the empirical expectation $\hat{\mathbb{E}}[\log \pi_x]$ over sampled seeds and drifts.
2. Aggregate via the exponential map:

$$\hat{\Phi}_{\mathcal{B}}^{\text{geom}} = \exp\left(\frac{1}{W} \sum_{x \in \mathcal{X}} w_x \hat{\mathbb{E}}[\log \pi_x]\right).$$

3. **Uncertainty:** Report a bootstrap confidence interval by resampling tasks within families and seeds within tasks.

Procedure 2: The Tractable Functional ($\Phi_{\mathcal{B}}$). To estimate the generalized functional (Eq. 1):

1. **Score:** For each task t , estimate success $q(t)$ and compute the non-linear utility $\psi_t(q(t), Q^*(t))$.
2. **Fairness:** Compute family means $\hat{q}(F_k)$ and subtract the dispersion penalty $\lambda \cdot \bar{\text{Var}}_k(\hat{q}(F_k))$.
3. **Cost:** Subtract the resource penalty $\gamma \cdot \hat{\mathbb{E}}[\text{Cost}(r)]$ using a fixed cost model.

Procedure 3: Copula Estimation. To analyze task correlations (required for defining the moduli metric):

1. Transform raw scores to $u(t)$ via Randomized PIT (Def. 4.1).
2. Estimate the dependency structure using rank-based empirical copulas or vine-copula estimators.
3. Validate using Goodness-of-Fit tests on pairwise margins.

10.2 Drift and Recalibration Protocol

To maintain the "don't obsess over tests" guarantee (Theorem 6.12), the evaluation must cover the moduli space locally:

- **Robustness Region:** Model allowable test variance as a ball of radius ρ in the copula space (Wasserstein metric) and a box of width δ around thresholds. Report the *worst-case* Φ over this region.
- **Anchor Maintenance:** Trigger a battery recalibration when the rank-concordance of anchor agents falls below a pre-registered threshold. Freeze new anchors only during specific leaderboard windows.

10.3 Complexity

For a battery with $|T| = m$ tasks and n seeds per task:

- Naive evaluation of Φ is $O(mn)$.
- Bootstrap Confidence Intervals scale linearly with the number of resamples B : $O(Bmn)$.
- Copula estimation scales as $O(mn \log n)$ for empirical ranks, or higher for parametric fitting.

Reproducibility requires publishing all seeds, anchor identities, and the exact cost model R .

11 Extensions

Multi-Agent Games. The framework extends to N agents evaluated on shared resources. If tasks interfere (coordination or congestion), outcomes depend on joint actions. Provided each agent’s evaluation functional Φ_B^i is continuous in the joint mixed strategy and quasi-concave in the agent’s own mixed strategy, and strategy sets are compact and convex, Glicksberg’s theorem guarantees the existence of a mixed-strategy Nash equilibrium in the induced evaluation game.

Coalitional Compositionality. Coalition scores can be derived via Minkowski sums of resource sets and convolutions of success laws. The dispersion penalties in our functional imply superadditivity for complementary coalitions (specialists combining to reduce variance) and subadditivity for redundant ones.

Dynamics of self-improvement κ . The AAI score $\Phi_B(\mathcal{A})$ measures the static capability of an agent instance on a fixed battery. Following [1], we can model self-improvement as a path of laws $(\nu_r)_{r \geq 0}$ indexed by cumulative resource r (e.g. training compute, code-rewrite budget), with $\nu_0 = \rho_B(\mathcal{A})$ and $\nu_{r+\Delta r} = T_{\Delta r}\nu_r$ for a family of transition operators $T_{\Delta r} : \mathcal{P}(X_B) \rightarrow \mathcal{P}(X_B)$. The absolute self-improvement rate at \mathcal{A} is then

$$\kappa_{\text{abs}}(\mathcal{A}) := \limsup_{\Delta r \downarrow 0} \frac{\Phi_B(T_{\Delta r}\nu_0) - \Phi_B(\nu_0)}{\Delta r},$$

and, whenever $\Phi_B(\nu_0) > 0$, the relative rate

$$\kappa_{\text{rel}}(\mathcal{A}) := \frac{\kappa_{\text{abs}}(\mathcal{A})}{\Phi_B(\nu_0)}$$

measures proportional gain per unit resource. Positive κ_{abs} or κ_{rel} indicates that the agent can, on average, increase its AAI score by investing additional resource. In the moduli picture (Section 5), the path $r \mapsto (\mathcal{B}, \nu_r)$ traces a trajectory in \mathcal{P}^\sharp . Under our regularity assumptions, Φ_B is Lipschitz with respect to d_\sharp , so κ_{abs} controls the tangent velocity of this trajectory. When self-improvement primarily increases success probabilities at threshold (i.e. raises $\mathbb{P}_{\nu_r}(q(t) \geq Q^*(t))$), the dominant contribution to this velocity comes from the Wasserstein motion of the copula component in the direction that increases threshold-aligned mass.

References

- [1] P. Chojecki. *An Operational Kardashev-Style Scale for Autonomous AI - Towards AGI and Superintelligence*. arXiv preprint arXiv:2511.13411.
- [2] D. Hendrycks *et al.* *A Definition of AGI*. arXiv preprint arXiv:2510.18212.
- [3] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Springer, 1998.
- [4] L. Rüschendorf, *On the Distributional Transform, Sklar’s Theorem, and the Empirical Copula Process*, J. Stat. Plan. Inference, 2009.
- [5] C. Villani, *Optimal Transport: Old and New*, Springer, 2009.