Evaluating the evaluation itself.

A high-throughput, unwavering measure for evaluation.

CATALOG 2026

Evaluating the evaluation itself, from outside.

The meta-evaluation idea

Ordinary AI evaluation has the AI directly score application materials. eval000's meta-evaluation engine instead evaluates and reconstructs the evaluation itself, from outside. The review's purpose is first defined as External Principle N, and SEGT (Standard Evaluation Generation Theory) convergence is applied iteratively to the scores given by multiple reviewers and AI persona evaluators — mathematically converging on a single Evaluation Standard (v*).

BEFORE CONVERGENCE (PER-REVIEWER DRIFT)

Human reviewer	68.2
AI persona · Pragmatist	84.0
AI persona · Innovator	89.0

→

AFTER SEGT CONVERGENCE

80.3

Standard Evaluation v*

*SEGT convergence applies whether the evaluators are all human, all AI personas, or a mix of both. eval000 Essential implements this as a first-pass evaluation using multiple AI persona reviewers.

A different kind of improvement

Conventional improvements — "gather better reviewers," "refine the criteria" — remain within an existing frame. eval000 is fundamentally different: it redefines the act of evaluation itself. Human judgment never intervenes in the calculation of the Standard Evaluation.

Three things eval000 delivers

VALUE 01

An evaluation that doesn't depend on panel composition

Multiple evaluator tendencies converge to a fixed point bound to External Principle N.

VALUE 02

Measuring "how strong the convergence is"

We're defining metrics — an Evaluation Measure and Stabilization Ratio.

VALUE 03

The evaluation "policy" is written down first

Review policy is defined up front as External Principle N, making the premise explicit.

What eval000 delivers

→ **For review organizers**

Writes down judgment criteria that used to live in individuals' heads as External Principle N, keeping the whole review process consistent.

→ **For applicants**

Records the reasoning behind the convergence process and auto-generates feedback reports, building trust in the outcome.

→ **For leadership & decision-makers**

Compresses first-pass screening that used to take dozens of hours, raising both speed and transparency across the review.

Evaluation drift isn't a matter of any reviewer's ability.

The "noise" inherent in human judgment

Daniel Kahneman et al., in "Noise: A Flaw in Human Judgment" (2021), found an average 55% variance in insurance claims adjusters' assessed amounts for the same case — and significant variation even for the same adjuster depending on time of day, order, and mood.

> POINT
>
> This isn't a matter of any individual reviewer's ability — it's a structural difficulty built into the review process itself.

Generative AI has "personality" too

Individual AI models — ChatGPT, Gemini, Claude — each carry their own review-philosophy bias. A 2025 PoC demonstrated gaps of up to 12 points on the same document across models. Simply handing evaluation to generative AI doesn't solve the problem.

Rubric design shapes the outcome too

Scoring the same "Ecchu-Yao" plan under two different rubrics produced very different results: under the AG rubric, Ecchu-Yao scored 70.6 vs. EcoBottle's 77.9 (a 7.3-point gap); under the BG rubric, Ecchu-Yao scored 80.3 vs. EcoBottle's 80.1 (just 0.2 points apart). The design of the evaluation axes themselves substantially shapes the conclusion.

These issues can be organized into six structural challenges, shown on the next page.

A capacity problem, as volume grows

At the same time, many organizations see entry volume grow year over year, straining a limited pool of reviewers. After adopting eval000, since AI scores and supporting comments are pre-generated, reviewers' work narrows to "verify and judge." As a pre-PoC hypothesis, we target a 90% reduction in review workload and 12× processing capacity (actual measured values are obtained through the PoC).

Six structural challenges, and SEGT's approach.

Issue	A difficulty many organizers feel	SEGT's approach	Status
1. Absence of N	What counts as "good" stays tacit, so the bar shifts as reviewers change	Writes the review's purpose down as External Principle N	Addressed by design
2. Rubric drift	Axes exist, but level definitions and weights stay vague	Makes axes, weights, and criteria explicit as an Evaluation Rubric	Addressed (depends on design)
3. Noise & bias	Order, fatigue, and domain bias seep in unchecked	Multiple AI persona judges remove noise, spread out bias	Addressed (depends on design)
4. Committee conformity	Members converge toward senior or louder voices	Computes v* as a fixed point independent of headcount	Addressed by design
5. Accountability	Explaining rejection to every applicant is labor-intensive	Records convergence reasoning, auto-generates feedback	Addressed by design
6. Reproducibility	Results shift with reviewers or timing	Same evaluator group + same input theoretically reproduces v*	Being validated

POINT: BEING VALIDATED

Banach's fixed-point theorem guarantees reproducibility for the same evaluator group, the same input, the same N, and the same rubric. Convergence across different evaluator groups is under active study.

Five layers — N, EC, R, v*, EM — that reconstruct evaluation.

eval000 organizes five layers — External Principle N, Evaluation Criterion, Evaluation Rubric, Evaluation Standard (v*), and Evaluation Measure — under a single theory: SEGT (Standard Evaluation Generation Theory).

Term	Definition
N	A governing principle, set by the Introducing Company, that supplies the purpose and direction of the review from outside the system. Constrains the entire evaluation process.
EC	A criterion, set in light of External Principle N, that defines the evaluation's viewpoints. Forms the basis for designing the Evaluation Rubric.
R	A specification, designed in accordance with N, that defines evaluation items, weights, and judgment criteria.
v*	An evaluation result, free of bias, noise, and error, derived by the meta-evaluation engine from N and R.
EM	A mechanism for continuing to use a generated Standard Evaluation (v*) as a reusable evaluation measure.

A mathematical guarantee of convergence

$$v^* = F_N(v^*, R)$$

Convergence guaranteed by Banach's fixed-point theorem

eval000's theoretical foundation rests on Banach's fixed-point theorem (a mathematical guarantee of convergence) together with Gödel's incompleteness theorem — the observation that a complete set of evaluation criteria cannot be proven from within the system itself. That's precisely why an externally-supplied principle, N, is needed.

> POINT
>
> This mathematical guarantee holds for the same evaluator group, the same input conditions, the same N, and the same rubric. Convergence across different evaluator groups is under active study.

A 20.8-point gap converged to a 0.2-point difference in v*.

At the eval000 hands-on seminar held in August 2026, two regional-revitalization business plans were scored by three AI persona judges under the same External Principle N and the same Evaluation Rubric (the "BG" rubric).

Ecchu-Yao (20.8-point gap between judges)

BEFORE CONVERGENCE (WEIGHTED TOTALS)

Pragmatist	68.2
Balancer	84.0
Innovator	89.0

Gap: 20.8 pts

→

AFTER SEGT CONVERGENCE

80.3

Standard Evaluation v*

EcoBottle (11.0-point gap between judges)

BEFORE CONVERGENCE (WEIGHTED TOTALS)

Pragmatist	73.0
Balancer	84.0
Innovator	84.0

Gap: 11.0 pts

→

AFTER SEGT CONVERGENCE

80.1

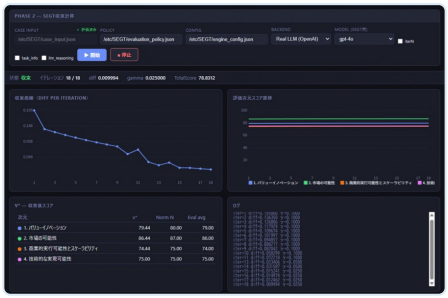
Standard Evaluation v*

POINT1

Despite very different pre-convergence gaps (20.8 vs. 11.0 points), both plans settled onto standard evaluations just 0.2 points apart — suggesting a stable benchmark regardless of how scattered the individual judges were.

The example above was run on our actual operating console, "SEGT Monitor." PHASE 1 runs the first-pass evaluation by AI persona judges; PHASE 2 runs the SEGT convergence calculation.

PHASE 1 — Document review (3-LLM evaluation): first-pass scoring by the Pragmatist, Innovator, and Balancer personas



PHASE 2 — SEGT convergence: converging to v* over 18 iterations (diff trend and per-dimension score trend)

Four tiers, matched to your scale and purpose.

Essential

Overview	Evaluation outsourcing (pay-per-project)
Scale	Per-project (tens to hundreds/year)
Term	One-off (no annual contract)
Time to launch	1–2 weeks from quote

Platform

Overview	Full-feature SaaS (annual contract)
Scale	50+ reviews/year, multiple programs
Term	Annual contract
Time to launch	From a 3-month PoC

License

Overview	Site license (on-premise available)
Scale	Government, large municipalities
Term	Annual (multi-year negotiable)
Time to launch	Individual schedule

OSS

Overview	Selected tech released as open source
Scale	Research & educational institutions
Term	Joint research agreement, etc.
Time to launch	By consultation

Start with Essential

eval000 Essential delivers the core technology of the meta-evaluation engine in a simpler, faster, more accessible form. No SaaS contract required — get started right from a quote. If continuous use is expected, we can also propose moving to the full Platform tier.

A rule of thumb for choosing

Essential is a good fit for a one-off contest or an annual review. Platform suits organizations running multiple programs year-round. License fits government bodies with data-sovereignty and security requirements. OSS suits research institutions looking for academic collaboration. If you're not sure which fits, just get in touch — we're happy to help you figure it out.

06 — APPLICABLE DOMAINS

A shared standard, across five domains.

01 **Contests, Judging & Awards** — **¥1M–5M /year**

Business contests, startup-support organizations, university/corporate contests, employee recognition. Public-sector grant review and policy evaluation also fall here.

02 **HR Evaluation & Recruitment** — **¥1M–5M /year**

360° reviews, first-pass hiring screens, promotion reviews. A domain where consistency and accountability both matter.

03 **Comparison & Procurement Platforms** `New Domain` — **By consultation**

For BtoB information-search and comparison platforms (product comparison sites, industry portals, etc.): standardizing listing review, lead-quality scoring, and objective comparison scores. The customer here is the platform operator itself, not an applicant — a different audience and use case from the other domains. Proposed since 2026.

04 **New Service & Business Development** — **By consultation**

Integration into existing AI services, core design for new business ventures. Can underpin a product's own evaluation feature.

05 **Research, Education & Nonprofit** — **¥1M–5M /year**

Research grant review, joint research with academic institutions. Collaboration via OSS is also envisioned.

GOVERNMENT & PUBLIC SECTOR

We propose a site license (¥5M–30M/year) meeting data-sovereignty and security requirements, with on-premise implementation available.

For organizations handling 50+ reviews a year

eval000 is built for organizations handling 50 or more review documents per year across these domains. Rather than replacing your existing process, you can start simply by re-evaluating past documents with AI and comparing against the human evaluation.

Combining multiple domains is also possible. For example, you could run HR evaluation and an internal new-business proposal program on the same eval000 environment, sharing External Principle N design know-how and evaluation infrastructure across the organization.

Transparent pricing, and a low-risk path to adoption.

eval000 Essential pricing (pay-per-project)

Domain	Setup fee	Per-item (up to 100 items)
Business contests	¥0 (pre-configured template)	¥4,000–¥5,000/item
Other domains	¥80,000/project	¥4,000–¥5,000/item

Platform: the 3-month PoC flow

MONTH 1

Setup

Design External Principle N and the rubric. Design the AI persona reviewer composition.

MONTH 2

Validation

Parallel operation with your existing process. Measure AI–human consistency and Kappa.

MONTH 3

Decision

Go/No-Go decision. Target: ≥60% workload reduction, ≥4.0/5 satisfaction, ≥80% consistency.

The human role: HITL Lv0–4

Level	Description
Lv0	Human-only review (no AI involvement)
Lv1	AI provides reference information; humans make all judgments
Lv2	AI proposes scores; humans verify every entry
Lv3	AI performs the first pass; humans check a sample
Lv4	**Humans define N and give final sign-off on the SEGT-converged result (eval000 recommended)**

POINT

After you contact us, an NDA is signed electronically within one business day and remains in effect for 3 years after service use ends. Payment is due within 30 days of invoicing.

Over 30 years in business, backed by academic rigor.

Three generations of products

GEN.1	→	GEN.2	→	GEN.3
mo4ma		**award-of**		**eval000**
Since 1995		Official Japan Partner, Award Force		Since 2026

Founded in 1995, TEN PROXY Co., Ltd. has over 30 years of experience centered on IT, marketing, business development, and go-to-market support. eval000 (GEN.3) is its third-generation product, following mo4ma (GEN.1) and award-of (GEN.2). mo4ma provided IT and marketing support; award-of operates the contest-management platform "award-of.net" and serves as the official Japan partner of Award Force, an Australian platform used in 50+ countries.

Leadership

PARTY A / PATENT HOLDER & TECHNICAL SUPERVISION

Ryuichi Satoyoshi

Completed a master's degree at the Graduate School of Economics, Yokohama City University. Currently affiliated with Tokyo Fukushi University. As patent applicant for the meta-evaluation engine, oversees technical supervision and quality assurance.

PARTY B / BUSINESS OPERATOR

Seiho Budo

Representative Director of TEN PROXY Co., Ltd. A fellow alumnus of the same seminar under Prof. Yoshiyuki Senga as Satoyoshi. Over 30 years of experience in IT, business development, and go-to-market support since founding the company in 1995.

POINT

Filed as Japanese Patent Application No. 2026-096693, "Standard Evaluation Calculation Method, Apparatus, and Program Based on Evaluation Reconstruction Using Generative AI," covering the SEGT convergence algorithm. Presented at IOTEIR 2026 (Bridgewater State University, July 28–29, 2026).

GET IN TOUCH

Bring us the review process you are least happy with

It's fine even if your volume or evaluation criteria aren't finalized yet. Consultations and quotes are free. NDAs can be signed electronically, often the same day.

https://eval000.ai ／ info@eval000.ai

TEN PROXY, INC. eval000 team