評価をメタ評価する。

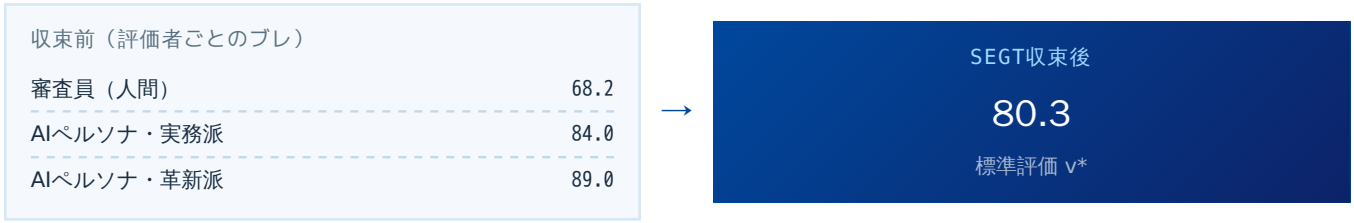
高生産性で、揺るぎない評価ができる、ものさし。

CATALOG 2026

評価そのものを、外側から評価する。

メタ評価という発想

通常のAI評価は、AIが直接応募書類を採点します。eval000のメタ評価エンジンは、評価そのものを外側から評価・再構成するアプローチです。まず審査目的を外生の原理N（External Principle N）として定義し、複数の審査員・AIペルソナ審査員が出した評価値に、標準評価生成理論（SEGT : Standard Evaluation Generation Theory）に基づく収束計算を反復適用します。これにより、数学的に唯一の標準評価（Evaluation Standard／v*）へ収束させます。



※SEGT収束は、人間の審査員のみ、AIペルソナ審査員のみ、またはその組み合わせのいずれにも適用できます。eval000 Essentialでは、複数のAIペルソナ審査員による一次評価としてこの仕組みを実装しています。

従来の改善とは異なるアプローチ

「より良い審査員を集める」「基準を精緻化する」という従来の改善は、既存フレームの中での工夫にとどまります。eval000は根本から異なり、評価という行為そのものの定義を刷新します。人間の判断は標準評価の算出そのものには介入しません。

eval000が提供する3つの価値

VALUE 01

審査員構成に左右されないブレない評価

複数の評価傾向を、外生の原理Nに拘束された固定点へ収束させます。

VALUE 02

「収束の強さ」を数値で測る取り組み

Evaluation Measure・Stabilization Ratioという指標の定義を進めています。

VALUE 03

評価の「方針」を最初に明文化する

外生の原理Nとして審査方針を定義し、審査の前提を明確にします。

eval000が届けるもの

→ **審査運営者へ**

属人化した判断基準を外生の原理Nとして明文化し、審査プロセス全体の一貫性を担保します。

→ **応募者へ**

収束過程の根拠を記録し、フィードバックレポートを自動生成することで、結果への納得感を高めます。

→ **経営層・意思決定者へ**

数十時間かかっていた一次選考の工数を大幅に圧縮し、審査全体のスピードと透明性を高めます。

評価のブレは、審査員個人の資質の問題ではない。

人間の評価に内在する「ノイズ」

ダニエル・カーネマンら著『Noise: A Flaw in Human Judgment』（2021年）では、保険会社の損害査定員を対象にした調査で、同一ケースに対する査定額が担当者によって平均55%のばらつきがあることが示されています。同一の担当者でも、時間帯・順序・気分によって有意な変動が生じます。

> POINT
> これは審査員個人の能力の問題ではなく、審査という仕組みそのものに内在する構造的な難しさです。

生成AIにも「個性」がある

ChatGPT・Gemini・Claudeなど、個々のAIモデルには「審査思想の偏り（AIの個性）」があります。同一書類に対して、モデル間で最大12点の差が生じることが2025年のPoCで実証されています。生成AIに評価を任せれば解決する、という単純な話ではありません。

ルーブリックの設計も、評価の方向性を左右する

同じ「越中八尾」という事業計画を、2種類のルーブリックで評価したところ、AGルーブリックでは越中八尾70.6点 vs EcoBottle 77.9点（差7.3点）でしたが、BGルーブリックでは越中八尾80.3点 vs EcoBottle 80.1点（差0.2点）という結果になりました。評価軸の設計そのものが、結論に大きく影響します。

これらの問題は、次のページに示す6つの構造的な課題に整理できます。

件数増加というキャパシティの問題

同時に、多くの現場では応募・エントリー件数が年々増加し、限られた人数の審査員では対応しきれないというキャパシティの問題も生じています。eval000導入後は、AIスコアと根拠コメントが事前生成されているため、審査員の作業は「確認・判断」に絞られます。PoC前の仮説値として、審査工数90%削減・処理能力12倍という目標を掲げています（正式な実測値はPoCを通じて取得します）。

6つの構造的な課題と、SEGTのアプローチ。

課題	多くの主催者が感じている難しさ	SEGTのアプローチ	状況
①Nの不在	「良い評価とは何か」が暗黙知化し、審査員が変わるたびに基準が揺れる	外生の原理Nとして審査目的を明文化する	仕組みで対応
②ルーブリック形骸化	評価軸はあっても水準の定義や重み付けが曖昧	Evaluation Rubricとして軸・配点・判定基準を明示	対応（設計品質に依存）
③ノイズ・バイアス	審査順番・疲労・専門分野の偏りが対策なく入り込む	複数のAIペルソナ審査員でノイズを排し偏りを分散	対応（設計品質に依存）
④合議での同調	経験豊富な審査員や声の大きい意見に同調しやすい	v*を人数比に依存しない固定点として算出	仕組みで対応
⑤説明責任	落選理由を応募者一人ひとりに伝えるのは工数的に難しい	収束過程の根拠を記録し、FBレポートを自動生成	仕組みで対応
⑥再現性	審査員や実施タイミングが変わると結果が変わる	同一評価者集団・同一入力ならv*が理論的に再現	検証中

POINT：検証中

バナッハの固定点定理が保証するのは、同一の評価者集団・同一の入力・同一のN・同一のRのもとでの再現性です。異なる評価者集団間の収束性は、実証研究中です。

N・EC・R・v*・EMという5階層で、評価を再構成する。

eval000は、外生の原理N・Evaluation Criterion・Evaluation Rubric・Evaluation Standard（v*）・Evaluation Measureという5階層の構造を、SEGT（Standard Evaluation Generation Theory）という一つの理論名のもとに整理しています。

N
外生の原理
（最上位）

→

EC
Evaluation
Criterion

→

R
Evaluation
Rubric

→

v*
Evaluation
Standard
（標準評価）

→

EM
Evaluation
Measure

用語	定義
N	導入企業が定める、審査・評価の目的及び方向性を体系の外部から与える上位原理。評価プロセス全体を拘束する。
EC	外生の原理Nを踏まえて設定される、評価の観点・視点を定めた基準。Evaluation Rubric設計の土台となる。
R	外生の原理Nを受けて設計される、評価項目・配点・判定基準を定めた仕様。
v*	メタ評価エンジンが、N及びRに基づき導出する、バイアス・ノイズ・誤差を除去した評価結果。
EM	生成された標準評価（v*）を、再利用可能な評価尺度として継続的に用いる仕組み。

数学的な収束保証

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

バナッハの固定点定理による収束保証

eval000の理論的根拠は、バナッハの固定点定理（収束の数学的保証）に加え、ゲーデルの不完全性定理（完全な評価基準は体系の内部だけでは証明できないという指摘）にも基づいています。完全な評価基準を体系の内部だけで構成することはできない——だからこそ、外生の原理Nという体系外部からの原理が必要になります。

> POINT
> この数学的保証が成立するのは、同一の評価者集団・同一の入力条件・同一のN・同一のRのもとです。異なる評価者集団間の収束性については、実証研究中です。

20.8点のブレが、0.2点差のv*に収束した実例。

2026年8月開催のeval000体験セミナー・ライブデモでは、地域活性化に関する2つの事業計画を、同じ外生の原理N・同じEvaluation Rubric（BGルーブリック）のもとで、3種類のAIペルソナ審査員に評価してもらいました。

越中八尾（審査員間のバラつき 20.8点）

収束前（3ペルソナの加重合計）

実務派	68.2
バランス派	84.0
革新派	89.0
差 20.8点	

→

SEGT収束後

80.3

標準評価 v*

EcoBottle（審査員間のバラつき 11.0点）

収束前（3ペルソナの加重合計）

実務派	73.0
バランス派	84.0
革新派	84.0
差 11.0点	

→

SEGT収束後

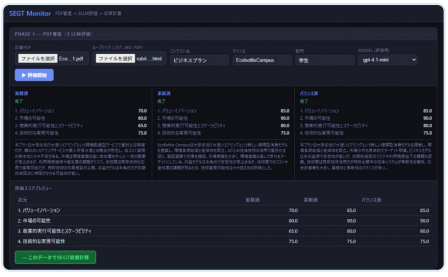
80.1

標準評価 v*

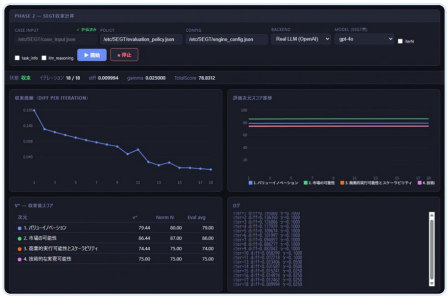
POINT1

審査員間のバラつきの大きさ（20.8点 vs 11.0点）が異なる2つの計画が、収束後にはわずか0.2点差の標準評価に収まりました。個々の審査員のブレの大きさに依存せず、安定した基準点を導けることがうかがえます。

上記の実例は、実際の管理画面「SEGT Monitor」上で実行されたものです。PHASE 1でAIペルソナ審査員による一次評価を行い、PHASE 2でSEGT収束計算を実行します。



PHASE 1 – PDF審査（3 LLM評価）：実務派・革新派・バランス派による一次評価



PHASE 2 – SEGT収束計算：18回の反復でv*に収束（diff値の推移・評価次元スコア推移）

導入規模・目的に応じた4つの提供形態。

Essential

概要	評価アウトソーシング（都度課金）
対象規模	案件単位（年数十〜数百件）
契約期間	単発（年間契約不要）
導入までの期間	お見積りから最短1〜2週間

Platform

概要	フル機能SaaS（年間契約）
対象規模	年間50件以上・複数プログラム
契約期間	年間契約
導入までの期間	3ヶ月のPoCから

License

概要	サイトライセンス（オンプレミス可）
対象規模	行政・大規模自治体
契約期間	年間契約（複数年も相談可）
導入までの期間	個別スケジュールによる

OSS

概要	技術の一部をOSS公開
対象規模	研究・教育機関
契約期間	共同研究契約等
導入までの期間	個別相談

まずは**Essential**から

eval000メタ評価エンジンの核心技術を、よりシンプルに・より速く・ご利用いただきやすい形でお届けするのがeval000 Essentialです。SaaS契約不要、お見積りからすぐに始められます。継続的な利用が前提となる場合は、フル機能のPlatformへの移行もご提案します。

選び方の目安

単発のコンテストや年1回の審査であればEssentialが適しています。複数のプログラムを年間通じて運用する場合はPlatform、データ主権・セキュリティ要件のある行政機関はLicense、学術連携を模索する研究機関にはOSSが目安です。どの形態が合うか分からない場合も、まずはお気軽にご相談ください。

5つの分野に、標準評価という共通言語を。

01 コンテスト・審査・表彰

100〜500 万円/年

ビジネスコンテスト、スタートアップ支援機関、大学・企業主催コンテスト、従業員表彰制度など。公的機関・行政の補助金審査・政策評価もこの分野に含まれます。

02 人事評価・採用選考

100〜500 万円/年

360度評価、採用一次選考、昇格審査など、社内の評価プロセス全般。評価の一貫性と説明責任の両立が求められる領域です。

03 比較評価・ベンダー選定・プラットフォーム 新領域

個別相談

BtoB情報検索・比較プラットフォーム（製品比較サイト、業界ポータル等）向けに、掲載審査の標準化・リード品質スコアリング・比較表示用スコアを提供する分野です。応募者ではなくプラットフォーム運営者自身が対象で、他分野とは対象・活用形態が異なります。2026年に提案を開始した新領域です。

04 新サービス・新事業開発

個別相談

既存AIサービスへの組み込み、新規事業の骨格設計。プロダクトの評価機能そのものを支える基盤として活用いただけます。

05 研究・教育・非営利

100〜500 万円/年

研究費・グラント審査、学術機関との共同研究など。OSS提供による連携も想定しています。

公的機関・行政向け

データ主権・セキュリティ要件に対応するサイトライセンス（500〜3,000万円/年）としてご提案します。オンプレミス実装にも対応可能です。

年間50件以上の審査書類を扱う組織へ

これらの分野で年間50件以上の審査・評価業務を扱う組織が、eval000の主な対象です。既存の審査プロセスを置き換えるのではなく、まずは過去の書類をAIで再評価し、人手評価と比較するところから始められます。

複数の分野を組み合わせた導入も可能です。例えば人事評価と社内の新規事業提案制度を同一のeval000環境で運用し、外生の原理Nの設計ノウハウや評価基盤を組織横断的に共有することもできます。

明確な料金体系と、リスクの小さい導入フロー。

eval000 Essential の料金（都度課金）

分野	セットアップ費	件数課金（〜100件）
ビジネスコンテスト分野	¥0（設定済みテンプレート）	¥4,000〜¥5,000／件
その他分野	¥80,000／案件	¥4,000〜¥5,000／件

Platform：3ヶ月PoCフロー

MONTH 1

セットアップ

外生の原理N・ルーブリック設計。AIペルソナ審査員の構成設計。

MONTH 2

検証

既存プロセスとの並行運用。AI・人間評価一致率、Kappa係数の計測。

MONTH 3

判断

Go/No-Go判定。工数削減60%+・満足度4.0/5+・一致率80%+が目標基準。

人間の役割：HITL Lv0〜4

レベル	説明
Lv0	人間のみで審査（AI活用なし）
Lv1	AIが参考情報を提示するが、判断はすべて人間が行う
Lv2	AIが補助的にスコアを提示、人間が全件確認
Lv3	AIが一次評価を実施、人間は抽出的にサンプル確認
Lv4	人間が**N**を定義し、**SEGT**収束結果を最終確認（**eval000**推奨）

POINT

お問い合わせ後、1営業日以内に電子署名にてNDAを締結します。機密保持義務はサービス利用終了後3年間継続。お支払期限は請求書発行後30日以内です。

30年超の事業実績と、学術的な理論的根拠。

3世代にわたるプロダクトライン

GEN.1	→	GEN.2	→	GEN.3
mo4ma		**award-of**		**eval000**
1995〜		Award Force 日本正規パートナー		2026〜

株式会社テンプロクシーは、1995年の創業以来30年超の実績を持ち、IT・マーケティング・事業開発・Go-to-Market支援を主軸に、mo4ma（GEN.1）・award-of（GEN.2）に続く3世代目のプロダクトとしてeval000（GEN.3）を展開しています。mo4ma（GEN.1）はIT・マーケティング支援事業、award-of（GEN.2）はコンテスト管理プラットフォーム「award-of.net」の運営及び、オーストラリア発・50カ国以上で実績のあるAward Forceの日本正規パートナー事業です。

運営体制

甲／特許権者・技術監修

里吉 竜一

横浜市立大学大学院経済学研究科修士課程修了。現在、東京福祉大学に所属。メタ評価エンジンの特許出願人として技術監修・評価設計の品質保証を担当。

乙／ビジネス主体・事業運営

武道 誠芳

株式会社テンプロクシー代表取締役。里吉氏と同じ千賀重義教授ゼミの同窓。1995年の同社設立以来、IT・事業開発・Go-to-Market支援を主軸に30年超の実績。

POINT

特願2026-096693「生成AIによる評価再構成に基づく標準評価算出方法、装置、プログラム」は、SEGT収束アルゴリズム（評価再構成と固定点収束）を対象として出願中です。IOTEIR 2026国際学会（Bridgewater State University、2026年7月28日〜29日開催）にて、学術的な発表を行っています。

GET IN TOUCH

最も改善したいレビュープロセスを お聞かせください。

件数・評価基準が固まっていない段階でも構いません。お見積り・ヒアリングは無償です。NDAは最短即日、電子署名で対応します。

https://eval000.ai ／ info@eval000.ai
株式会社テンプロクシー eval000チーム