



測定の方法・条件・限界

2026 年 9 月 株式会社 AGEST

1. 本書について

本書は、AGEST ZERO TEST について当社が公表している数値の「どのように測ったか」「どの条件で測ったか」「何が測れていないか」を示す文書です。

公表値をそのまま信じていただくためではなく、同じ手順で追試できる状態にすることを目的としています。

本書の扱い

- ・ 記載の数値は、すべて当社の検証環境での測定です。お客様のシステムでの実績ではありません
- ・ 当社が公表している数値のうち、測定条件が記録として残っているものだけを掲載しています。掲載していない数値とその理由は §7 に記載します
- ・ 測定に用いた検証システムは当社が自作したもので、意図的に不具合を埋め込んであります

2. 測定環境

項目	内容
ハードウェア	NVIDIA DGX Spark
AI モデル	Qwen3 30B（社内 GPU 上のローカル実行。外部 API を使用しない構成）
ブラウザ	Playwright 同梱の Chromium
対象システム	当社が自作した検証用サイト 24 業種（demo-sites）／販売管理ベンチ（卸販売・請求管理システム）
配信方法	作業ツリーを一時 HTTP サーバで配信（外部ホスティングを使わない）

配信方法を作業ツリー直配信にしているのは、外部ホスティングの CDN キャッシュにより「直したはずの版ではなく古い版を測っていても数字だけは出る」状態を避けるためです。

3. 本書に掲載する 3 つの数値

指標	値	測定日	対象	測定の種類
テスト設計量	1.26 倍	2026-08-25	検証用サイト 23 業種	大規模 LLM（Claude Opus 4.8）との比較
不具合の検出率	98%（到達可能 45 件中 44 件）	2026-08-19	検証用サイト 24 業種	埋め込み欠陥の検出
誤検知	0 件	2026-08-19	検証用サイト 24 業種	正常版どうしの比較

★ 3 つは測定日・対象・条件がそれぞれ異なります。1 つの表にまとめて「総合性能」として読まないでください。

4. 指標① テスト設計量 — 1.26 倍

4.1 目的

同じ画面を見せたときに、本ツールと大規模 LLM のどちらが多く異なるテストケースを作れるかを測ります。量の測定であり、網羅率ではありません。

4.2 方法

比較を成立させるために、次の 4 点を両者へ同一に適用しています。

#	適用したこと	なぜ必要か
1	同一観測	画面の探索（到達）を本ツールに固定し、同じ観測データを両者へ渡してから各自にケースを生成させる
2	同一分類器	決定論ルール分類器（LLM を使わない・再現可能）で観点を付与する
3	機械変動の除外	「入力＝正常値→送信」の反復、ナビゲーションの空振り、Enter 送信などを除外する
4	近重複の畳み込み	数値・括弧内・助詞・記号を落とした正規化キーで、観点内の重複を 1 つに畳む

★ 3 と 4 を入れないと、同じ内容のケースが数として積み上がり、量の比較が成立しません。

4.3 結果（観点別・畳み込み後）

観点	本ツール	Claude Opus 4.8	倍率
同値分割／データ網羅	87	34	2.56
状態遷移／データ整合	57	26	2.19
境界値	100	55	1.82
必須チェック（空送信）	77	55	1.40
業務ルール	51	40	1.27
正常系（成立）	72	60	1.20
認証（誤認証・権限保護）	2	4	0.50
形式チェック	4	28	0.14
料金／計算の検証	0	55	—
合計（9 観点・重複除外）	450	357	1.26

4.4 この測定の限界

★ 観点別の倍率を単独で引用しないでください。言い回しの差を両方向に含みます。

- ・ 正常系は本ツールが過大に出やすい：機械変動が分類器のフォールバックで正常系に寄るため。水増しを除いた後でも、この傾向は残ります
- ・ 料金／計算・形式チェックは本ツールが過小に出る：本ツールの計算系は「金額を変えたら合計が変わる」という挙動ベースの確認であり、「計算式を検証する」という言い回しの分類には乗りにくいからです
- ・ 測れているのは量であり、質ではありません。1 ケースあたりどれだけ欠陥を見つけられるかは、この測定では分かりません
- ・ n=1 の測定です。複数回の反復による分散は取っていません

したがって当社が言えるのは、次の範囲です。

「同じ画面（観測）から、重複を除いても約 1.26 倍の異なるテストケースを生成できる。強みは入力の組合せ網羅（同値分割・境界値・空送信・状態遷移）であり、意味の理解が要る観点は大規模 LLM のほうが上で、相補的である」

5. 指標② 不具合の検出率 — 到達可能 45 件中 44 件（98%）

5.1 目的

当社が自作した 24 業種の検証システムに意図的な不具合を埋め込み、改修前（正常版）と改修後（不具合版）を比較させて、いくつ見つけれられるかを測ります。

5.2 条件

項目	内容
測定日	2026-08-18 ～ 2026-08-19（24 業種を完走）
対象	当社検証用サイト 24 業種（会計・受発注・CRM・EC・教育・金融・保険・在庫・物流・医療・給与・生産・案件管理・公共・購買・不動産・予約・飲食・販売・チケット・旅行・管理画面・CMS・コミュニティ）
埋め込み欠陥	46 件
探索時間の上限	1 業種あたり 90 秒
AI の利用	有効（フル）
業務チェック	有効
版表示の扱い	「正常版」「不具合版」等の版を示す文字は比較対象外に指定

★ 版表示を比較対象外に指定しないと、全画面が差分として出て本物の欠陥が埋もれます。この指定を外した測定は、数値として意味を持ちません。

5.3 結果

数え方	結果	備考
埋め込み欠陥 全件に対して	44 / 46（96%）	分母に到達不能な 1 件を含む
到達可能な欠陥に対して	44 / 45（98%）	★公表値はこちら
誤検知（正常版どうし）	0 / 24 業種	§ 6 に詳述

5.4 見逃した 2 件の所在

★ 見逃しの所在を示せない検出率は、数値として意味を持ちません。

セル	欠陥の種類	状態
給与 / negative	負値・不正値を許容	★黒箱では原理的に到達不能。画面操作から当該状態へ至る経路が存在しないため、公表値では分母から除外している
飲食 / validation	必須・形式検証の欠落	未調査。前夜の測定が当該業種へ到達する前に停止しており、前後比較ができていない

5.5 欠陥種別ごとの内訳（別測定）

★ 以下は § 5.2 とは別の測定です。分母も条件も違うため、上の表と混ぜないでください。

項目	内容
測定日	2026-08-13 19:45 ~ 2026-08-14 16:24
対象・規模	24 業種 / 欠陥セル 136 + 誤検知セル 24 = 160 セル
探索時間の上限	1 セルあたり 180 秒
その他	非機能比較 有効 / 巡回上限 40 ページ / セル所要合計 約 12.5 時間

欠陥の種類	内容	期待区分	検出	要確認	見逃し	検出率
calc	金額・数量の計算誤り	detect	18	0	0	100%
price	価格・単価の表示改ざん	detect	16	0	0	100%
order	一覧の並び順の変更	detect	16	0	0	100%
label	ボタン・見出しの文言変更	detect	15	0	0	100%
missingfield	入力欄が消える	detect	15	0	0	100%
各種計算誤り	税・試算・在庫・控除・合計	detect	各 1	0	0	100%
brokenlink	ナビゲーションのリンク切れ	detect	0	0	1	0%
validation	必須・形式検証の欠落	hard	11	0	8	58%
dupe	重複データが登録できる	hard	1	0	0	100%
sameacct	同一勘定への振替を許容	hard	1	0	0	100%
negative	負値・不正値を許容	hard	0	0	2	0%
console	読み込み時の JS エラー	oob	0	16	0	0%
slow	応答の遅延（非機能）	oob	0	2	9	0%

期待区分	検出率	意味
detect（画面から見える変化）	99%（85 / 86）	本ツールが本来対象とする範囲。唯一の見逃しは 販売 / brokenlink の 1 セル
hard（別のオラクルが必要）	57%（13 / 23）	差分の比較だけでは原理的に届きにくい種類。 validation・dupe・negative・sameacct
oob（設計上の対象外）	0%（0 / 27）	JS エラー・応答遅延など。本ツールの対象範囲外として設計している

5.6 業種別の検出率を掲載しない理由

業種ごとに埋め込んだ欠陥の数と種類が異なるため、業種別の率は業種間で比較できません。50%~100%の幅が出ますが、これは業種による強弱ではなく、その業種にどの種類の欠陥を埋め込んだかで決まります。

当社の一次記録にも「業種別の値を単独で引用しない」と明記しています。

5.7 この測定の限界

- ・ 本ツールは差分オラクルであり、欠陥に名前を付けません。同一画面の複数の欠陥は1つの差分として提示されます。したがって「どの欠陥を見つけたか」を1件ずつ名指しできない場合があります
- ・ hard 区分は原理的に届きにくい種類です。「入力チェックが無い」ことは、改修前も改修後も同じようになってしまうため、前後の比較では差が出ません
- ・ oob 区分は対象範囲外として設計しています。応答遅延や JS エラーは別の手法（負荷試験・監視）で見べき対象です
- ・ 改修前から誤っていた挙動は見つかりません。本ツールは改修前を正とするため、改修前の時点で仕様に反していた動作は差分として出ません

6. 指標③ 誤検知 — 0 件

6.1 目的

改修していない版どうしを比較させ、「変わっていないのに変わったと言わないか」を測ります。検出率と対で見る必要がある指標です。

6.2 結果

測定	結果	条件
24 業種のマトリクス（正常版 vs 正常版）	誤検知なし 24 / 24 業種	§ 5.2 と同条件
検証欠落の再測（正常版 vs 正常版）	誤検知なし 11 / 11 業種	探索上限 150 秒 / AI 誘導 / 非機能比較 有効
販売管理ベンチ（正常版 vs 正常版）	明確な差分 0 件 / 仕様不合格 0 件	32 ケース全て「同じ」判定 / 環境由来のノイズ 0

6.3 ★この測定の限界（重要）

上記の 0 件は、次の条件下での値です。

- ・ 静的なサイト（実行のたびに表示が変わらない）
- ・ 版を示す文字を比較対象外に指定した状態

★ 動的な要素を含む実システムで同じ結果になる保証はありません。

実システムでは、日時・採番（伝票番号・注文番号）・在庫数・広告・セッション由来の値などが実行のたびに変わります。これらは差分として出ます。

本ツールには抑制する手段（比較対象外テキストの指定、感度の設定、動的値の正規化、画像差分のしきい値、API の除外、不安定なケースの除外）を備えていますが、実システムでどう設定すればどこまで抑えられるかの実測は、現時点で行っていません。

したがって当社は「誤検知ゼロの製品」という表現を用いません。動的な要素を含むシステムでは、比較から外す設定を行った上で、効き方を対象システムごとに確認する運用を前提としています。

7. 本書に掲載していない数値と、その理由

当社が他の資料で公表している数値のうち、次の2つは本書に掲載していません。測定条件を記録から復元できないためです。

数値	掲載しない理由
1 システム分のテストを 38.9 分で完了	当時の測定手順の記録が見つからず、同じ条件を再現できません（2026-09-01 に確認）。値そのものは実測ですが、本書の目的（追試できる状態にする）を満たせません
画面到達率 82～100% ／ 操作到達率 84%	当時の探索時間の上限などの条件が記録に残っていません。また操作到達率の分母は「発見できた操作対象」であり、発見できなかったものは分子にも分母にも現れません

★ これらを他の資料から削除したわけではありません。本書は「方法・条件・限界を示す文書」であるため、条件を書けない数値を含めない、という判断です。

今後の測定では、測定開始時のソースコードの版と条件一式を記録し、公表値に併記する運用へ切り替えています。

8. 本ツールの原理から来る限界

以下は測定の限界ではなく、本ツールの仕組みから来る性質です。

限界	内容
改修前の誤りは見つからない	改修前のふるまいを正とするため、改修前から仕様に反していた動作は差分に出ません
新規に作った機能の期待値は決められない	比較する相手が存在しないため、新機能の正しさは人が決める必要があります
画面に出ない変化は範囲外	画面・文言・通信・データベースの観測範囲を超える内部状態の変化は扱いません
タイミング・並行性の不具合	決定論的な操作の再現では再現しにくい領域です（負荷試験・並行試験は別手法）
状態の組合せは観測できた範囲が上限	全状態は踏めません。境界値・同値分割・業務観点の付加で削減しています
画面が DOM 要素として表現されていない場合	Canvas・WebGL・独自描画の UI は、クリック対象として観測できません

9. 追試の手順

本書の数値は、次の材料があれば追試できます。当社にご請求ください。

- ・ 検証用サイト（24 業種）のソースと、埋め込んだ欠陥の一覧
- ・ 販売管理ベンチのソースと、埋め込んだ欠陥のフラグ定義
- ・ 測定に用いた設定（探索時間の上限・巡回上限・AI の有効無効・比較対象外の指定）
- ・ 測定時のソースコードの版

なお当社の検証用サイトは、意図的な不具合を含んだ状態で維持しています。測定基盤であるため、これらの不具合は修正しません。

10. 用語

用語	意味
デグレード	改修前は動いていた機能が、改修によって壊れる現象
リグレッションテスト	デグレードを見つけるために、変更後に既存部分を再テストする行為（JSTQB の用語法）
差分オラクル	仕様書ではなく、改修前のふるまいそのものを「正しい姿」として比較する方式
到達不能	画面の操作からは当該状態へ至る経路が存在しないこと。黒箱の手法では原理的に届かない
detect / hard / oob	欠陥の期待区分。detect = 画面から見える変化、hard = 別のオラクルが必要、oob = 設計上の対象外

株式会社 AGEST <https://agest.co.jp/zerotest>

※ 本書の数値はすべて当社検証環境での測定であり、お客様のシステムでの実績ではありません。