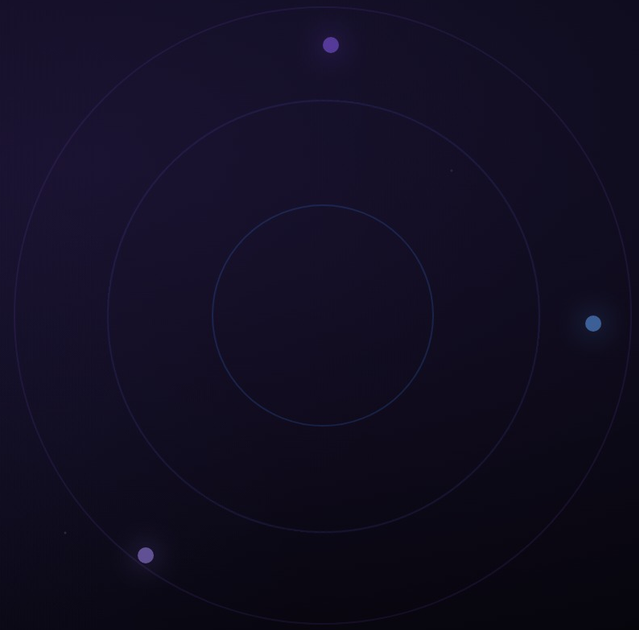
PRODUCT OVERVIEW · ADOBE EDITION · APRIL 2026

VIVID

See how your AI performs —
before real users do.



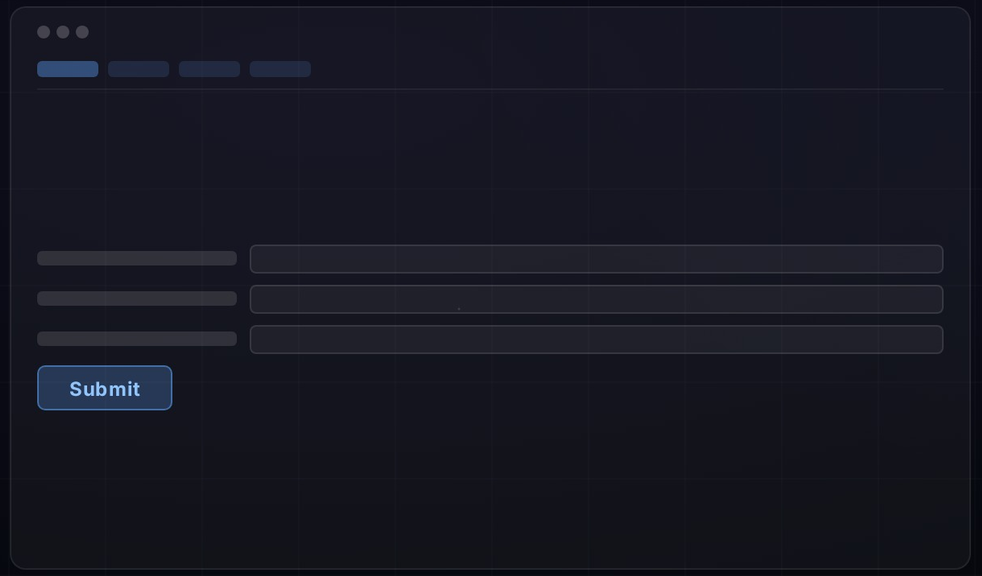
THE SHIFT

From apps you click to agents you talk to.

The software primitive is changing in real time. Testing the new surface with the old tools leaves you flying blind — **because the user is no longer choosing from a menu; they're *collaborating* with a system that has its own state.**

YESTERDAY

App-based UX

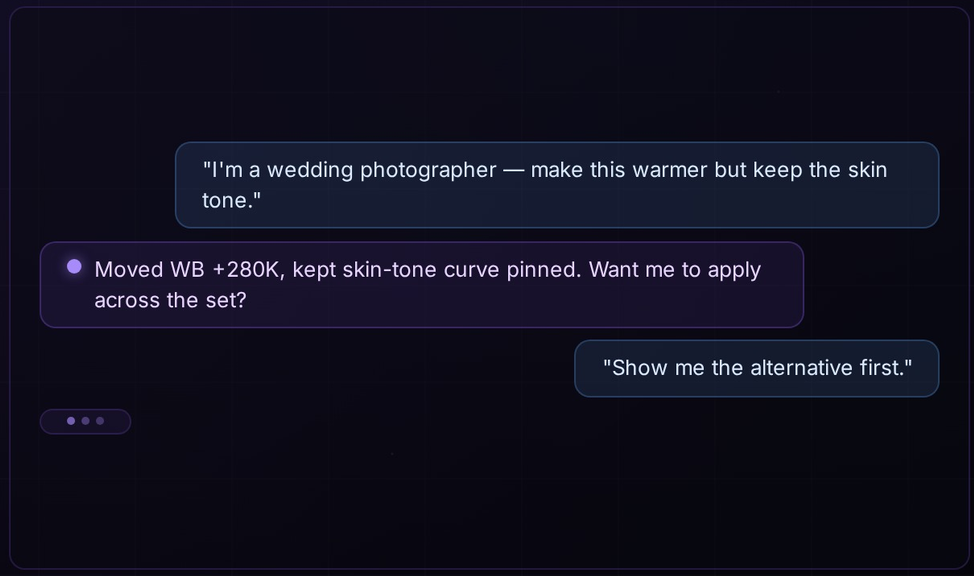


- Deterministic · same click, same result
- Finite state space · screens, buttons, forms
- Testable with click paths & screenshots

SHIFT

TODAY → TOMORROW

Agent-based UX



- Non-deterministic · same prompt, different paths
- Multi-turn · context accumulates, trust erodes
- Autonomous · the agent chooses, not the user

How do you evaluate a product where every user takes a different path, and the product itself changes what it does based on who's asking?

VIVID'S ANSWER

Simulate the **user**, not just the prompt. **Personas with memory, personality, and goals** — surfacing the paths real users will take, before they take them.

THE PROBLEM

Building AI is getting easy. Evaluating it is still broken.

AI products are **non-deterministic, multi-turn, and increasingly autonomous**. Traditional QA wasn't built for this. Most teams ship with hope instead of evidence.

Building
GETTING EASIER

- Foundation models via API
- One-click deployment anywhere
- GPU costs down 10× in 5 years

Evaluating
STILL BROKEN

- Non-deterministic outputs
- Multi-turn, tool-using agents
- Benchmarks ≠ real failures

EVALUATION LANDSCAPE

From a single question to an entire universe of experience.

Watch how evaluation compounds — each stage builds on the last. A benchmark is one dot. **VIVID evaluates the whole universe.**

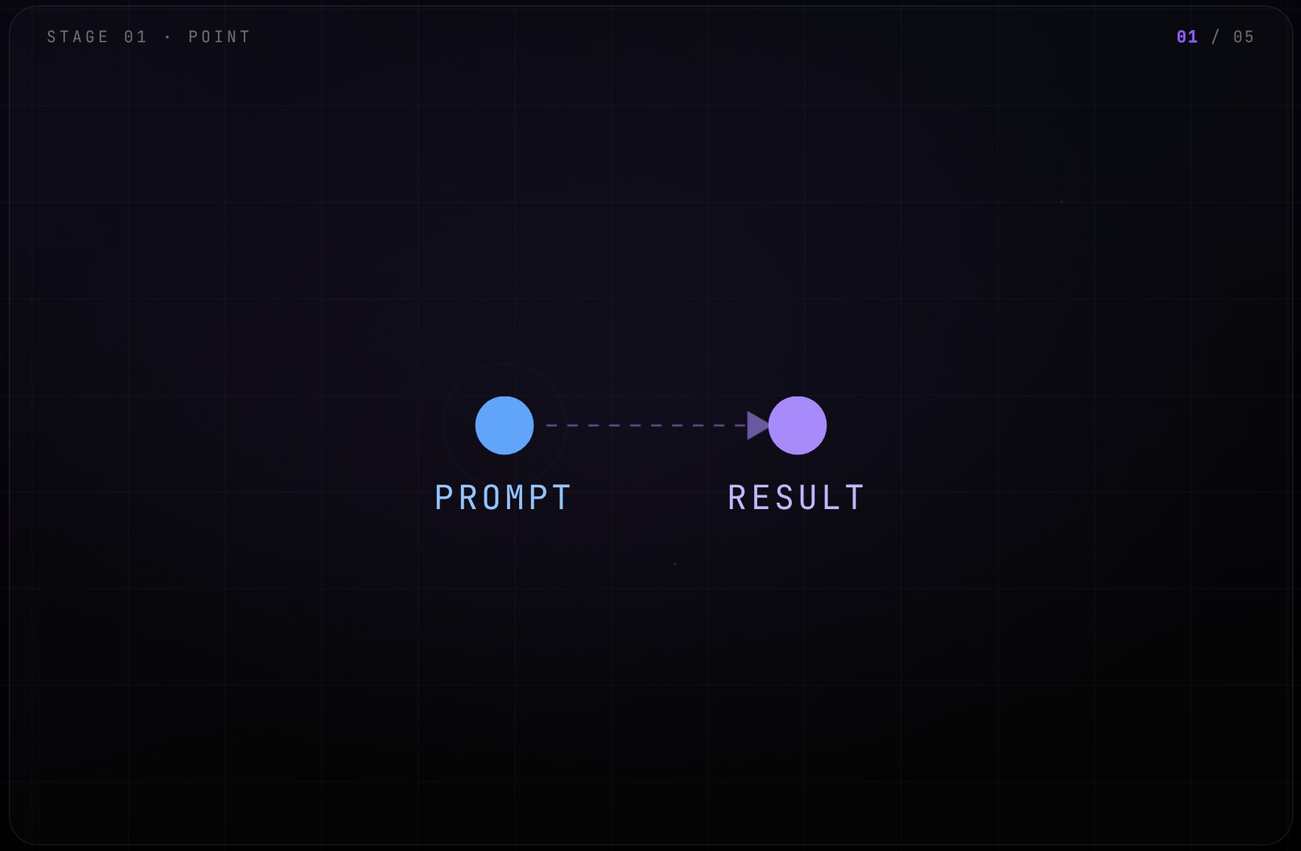
01 One prompt, one result
A single prompt scored by a single metric.

02 **A full conversation**
Many prompt → result turns inside a single session.

03 **Many sessions, many people**
Parallel sessions across diverse personas.

04 **Stretched across time**
Longitudinal — users return, context deepens.

05 **A universe of parallel worlds**
Personas × time mixed with A/B comparisons — every scenario coexisting in one evaluation universe.



◄ PREV ⏸ PAUSE NEXT ►

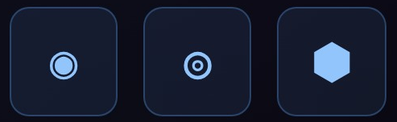
Others evaluate a point. We evaluate the world.

THE PLATFORM

PATENT PENDING · TECHNOLOGY

VIVID simulates the actors your AI will encounter.

Two kinds of agents run against your system so you see how it actually behaves in production — not just whether it answers correctly.



Synthetic Users

Personas with personality, memory, and goals. They return across sessions and judge your AI the way real people will.

YOUR AI

Under Test

Adversarial Agents

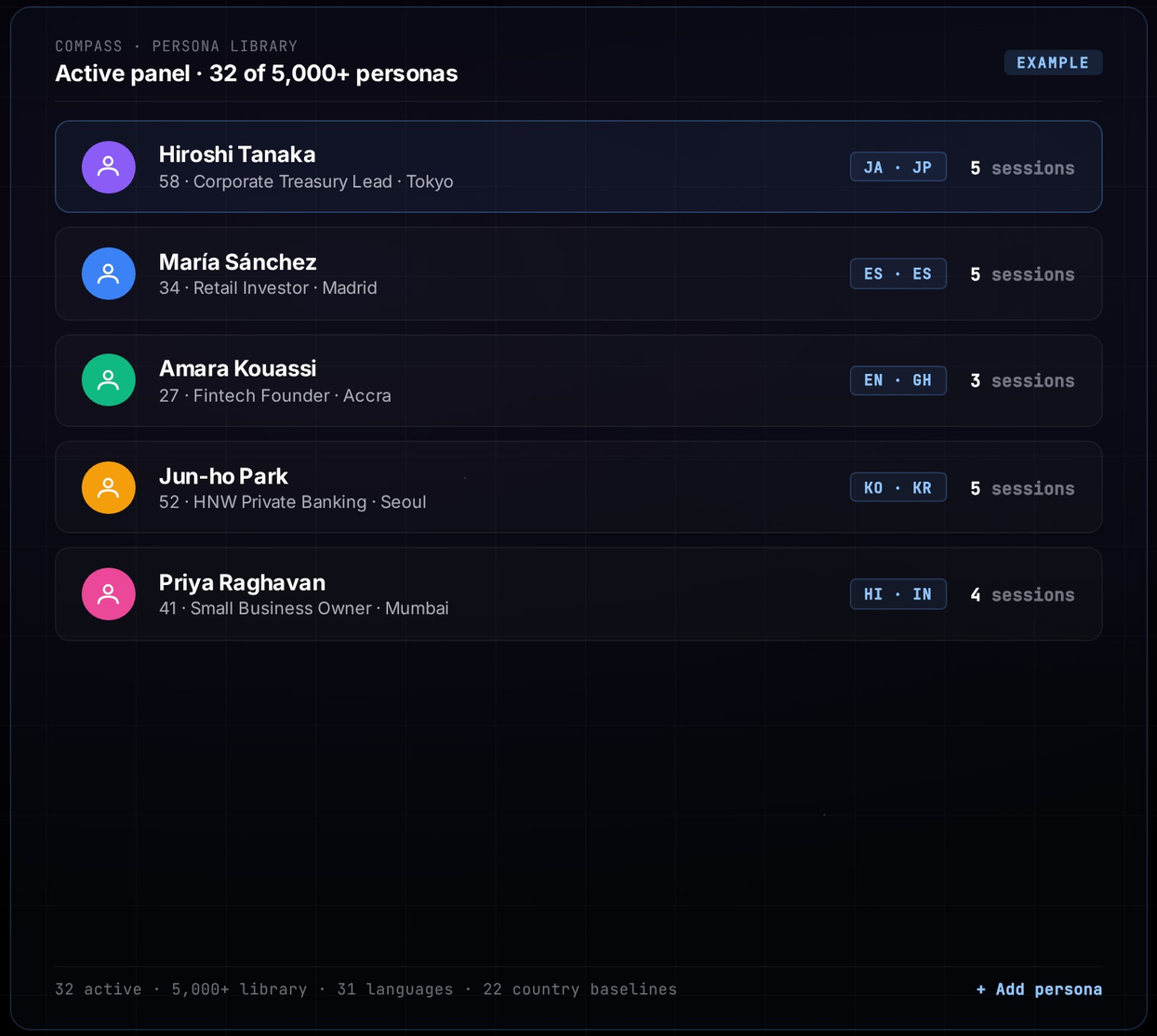
Autonomous attackers that adapt, chain, and push toward the weakest point — before customers or auditors find it.

PSA ENGINE | Each agent has **persistent personality, memory, and goals** — grounded in Big Five norms and empirical demographics. **Not a prompt. A stateful agent.**

◉ COMPASS

Test the AI experience with synthetic users who remember.

Virtual users that return across sessions, build memory, and surface the trust erosion single-turn tests never catch.



Diverse Persona Panels

5,000+ personas grounded in ~818K empirical data points across IPIP-NEO personality norms, 22 country baselines, and 68,540 occupation-tagged subjects. Available in 31 languages.

Multi-Session Longitudinal

Personas return session after session. Memory compounds, opinions evolve — the trust erosion that single-shot tests can't see.

Emotion & Trust Tracking

Each persona maintains an internal diary. Satisfaction, frustration, and trust measured turn by turn.

Multi-Modal Perception

Personas read, see, hear, and watch. Text, image, audio, and video — the full experience your AI delivers.

CI/CD Native

Runs in any pipeline via SDK, CLI, or REST API. Quality gates before every release.

WHAT MAKES A PERSONA

A prompt is a costume. A PSA is a character with their own state.

Most "persona evaluation" today is a system prompt that asks the model to perform a role. **VIVID's Persona-Simulating Agent (PSA) is different in kind** — a stateful agent whose behavior is derived from psychological traits, not improvised from a paragraph of text.

COMMON PRACTICE

Prompt-based persona

system_prompt.txt

```
"You are Sarah, a 34-year-old
wedding photographer in Brooklyn.
Personality: friendly, detail-oriented,
slightly impatient.

When the user asks about photo editing,
respond as Sarah would. Keep your answers
concise and use casual language."
```

ENGINE | single LLM call · no state · no memory

- **Stateless** — no memory between turns or sessions
- **Decorative personality** — adjectives in a prompt don't drive behavior
- **Different evaluator → different "Sarah"** — improvisation, not simulation
- **No calibration** — grounded only in LLM training data

vs

VIVID

Persona-Simulating Agent

PSA · stateful runtime

① **Trait core · OCEAN**
O 0.78 · C 0.94 · E 0.38 · A 0.48 · N 0.45

② **Memory · pgvector**
5 sessions · 87 turns indexed

③ **Emotion state · live**
trust 0.62 ↓ · frustration 0.41 ↑

④ **Calibration · trust badge**
EXACT · 23-sample real-user pinning

ENGINE | perception → memory → action loop · stateful agent

- **Stateful** — memory persists across sessions and runs
- **Causal personality** — Big Five values mathematically drive every reaction
- **Same persona, every evaluator** — same trait values produce reproducible behavior
- **Trust-badged** — calibrated against held-out real-user reactions

WHY THIS MATTERS FOR EVALUATION

A prompt-based persona **collapses** when the user pushes past where the prompt anticipated; behavior is improvised at run-time and unreproducible. A PSA **responds** — and every reaction traces back to a trait value, a memory chunk, an emotion delta you can **audit, reproduce, and re-calibrate**. The difference between **narrative simulation** and **behavioral simulation**.

HOW PERSONAS ARE BUILT

Three layers, grounded in psychological literature.

Big Five is not a label on top of the persona — **it *determines* how the persona acts**. Frustration intensity, patience, conversation style, evaluation leniency all derive mathematically from trait values.

L1

Big Five (OCEAN) — personality core

Continuous trait values. Costa & McCrae's Five-Factor Model. Same AI, same prompt — a high-Neuroticism persona reacts emotionally at failure #3; a low-Neuroticism persona persists past #10.

L2

Psychological type · demographics · communication style

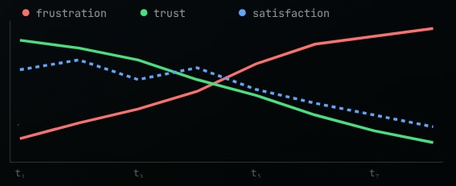
MBTI 16-type (auto-mapped from OCEAN via literature correlations), demographics (age, gender, 7 regions, language, timezone), communication style (formal/casual/terse/verbose), expertise level (novice/intermediate/expert).

INFP	34 · US	ES · Houston	formal
expert	UTC-6	ENTJ	41 · JP
terse	expert	ESFP	casual

L3

Real-time emotion state tracking

Every dialogue turn updates emotion dimensions (frustration, trust, satisfaction, ...). Delta, volatility, trajectory are the core data for longitudinal evaluation. The emotion response itself is modulated by Big Five — each persona's speed and intensity differs.



5,000 persona pool · structured distribution across the trait space · cohort coverage limits shown explicitly in every report

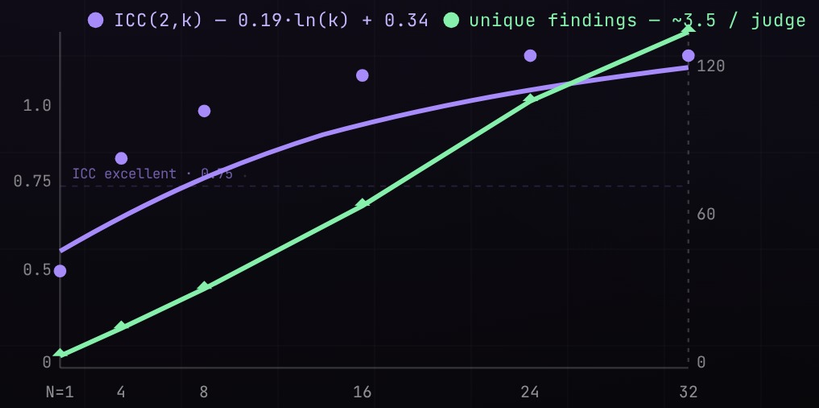
QUESTION 1 — HOW MANY PSAS?

Scoring converges logarithmically. Discovery grows linearly. They are not the same problem.

Across **32 PSAs · 960 sessions · 15 tasks · 5 domains · 2 model pairs**, two scaling laws coexist within the same panel. Optimizing for measurement (*"is this system good?"*) and optimizing for coverage (*"what are this system's problems?"*) require different panel sizes.

Score–coverage dissociation

$R^2 > 0.97$ (proprietary & open-source pairs)



MECHANISM Each PSA traverses a different interaction path · scoring noise averages out · discoveries accumulate. **Variance decomposition: 70-75% residual (judge × task), <1% between-judge.**

Practical deployment

Three operating points, calibrated to research goal

N = 4 **Continuous monitoring** — ICC 0.62

Catch regressions vs. a known baseline. Moderate reliability, **lowest cost**. Best for diff-based alerting on every build.

N = 8–12 **Periodic audits** — ICC 0.77

Good reliability **plus** broad coverage — discoveries still grow ~3.5/judge. **Best general-purpose deployment.**

N = 32 + humans **Milestone evaluation** — ICC 0.93

Excellent reliability + ~115 unique findings. **Pair with targeted human study** — agents and humans surface different kinds of issues.

★ **Composition matters more than size** — mixing expertise levels gives tighter scores AND broader discovery than uniform panels of the same N.

QUESTION 2 — ARE PSAS REALLY LIKE REAL USERS?

Three independent tests. All three say yes.

A persona claim is only as good as the evidence behind it. We ran three tests — a Turing-style human comparison, a controlled ablation against simple prompting, and a Big-Five behavioral validation — to see whether structured PSAs are doing what we say they're doing.

① TURING-STYLE VALIDATION

Agent–human gap is *inside* human–human gap

p = 0.379

paired t(14) = -0.91 · agent–human differences indistinguishable from human–human differences

Human ↔ Human	0.201
Human ↔ PSA	0.188
PSA ↔ PSA	0.143

mean |score difference| · lower = more agreement · 86 sessions, 43 raters

41% of human raters said the PSA panel **found issues they missed**

19% reported the reverse — PSAs are **complementary, not redundant**

② ABLATION · STRUCTURED VS. SIMPLE PROMPT

Structure is *causal*, not decorative

CONDITION	SCORE SD	INSIGHTS / SESSION	EXPERTISE D
Structured PSA	**0.087**	**13.2**	**-0.35**
Simple prompt	0.160	9.0	-0.17
No persona	0.164	8.6	**-1.03**
Same agent ×8	0.151	12.8	–

½× score variance vs. simple prompt

+47% insights per session

×3 smaller uncontrolled expertise bias

③ BIG FIVE BEHAVIORAL VALIDATION

Traits drive *actual* behavior

Trust gain ~ Agreeableness — CONFIRMED

r = **+0.754** · p < 0.001

Peak frustration ~ Neuroticism — CONFIRMED

r = **+0.756** · p < 0.001

Engagement ~ Extraversion — NOT CONFIRMED

r = **+0.057** · p = 0.757

The behavioral signature is real. Trust and frustration emerge from trait values exactly as personality literature predicts — Extraversion's link to engagement breaks because in goal-directed tasks, engagement is driven by progress, not social stimulation.

Jung & Na (2026) · "Logarithmic Scores, Linear Discoveries: Disentangling Measurement from Coverage in Agent-Based Evaluation" · 50 raters · 86 sessions · 15 tasks · 5 domains

MULTIMODAL EVALUATION

Five modalities. One persona with attention priorities derived from personality.

A high-Openness persona weights creativity more heavily on image tasks; a high-Conscientiousness persona prioritizes technical accuracy. The **same persona** behaves consistently across text, image, audio, video, and computer use — because behavior derives from traits, not from modality-specific prompts.

MODALITY	STATUS	ENGINE	PERSONA OUTPUT
✎ **Text dialogue**	● PRODUCTION	Proprietary + Open-source LLMs	Turn-level diary · emotion tracking · cross-session insights
■ **Image**	● PRODUCTION	Proprietary + Open-source Vision LLMs	Composition · quality score · prompt alignment
♪ **Audio & music**	● PRODUCTION	Proprietary + Open-source Audio LLMs	Genre · mood · production quality · persona calibration
▶ **Video**	◐ BETA	Proprietary + Open-source Video LLMs	Scene analysis · segmentation · Enterprise tier
⌨ **Browser / Computer Use**	● PRODUCTION	Proprietary CU Provider	Screenshot → action → observe loop · autonomous agent

Not locked to any model
7+ LLM providers · proprietary (GPT, Claude, Gemini) and open-source (DeepSeek, Qwen, Kimi) · configurable per customer and per target

Coverage we add on top of existing benchmarks
Longitudinal trust evolution · emotional response · cross-cultural variance — the dimensions benchmarks don't measure

SCENARIO GENERATION

Expert-curated base. Runtime adaptation.

Quality ceilings require expert-written scenarios. Real-world coverage requires adaptive generation. We do both — curated scenarios as the floor, LLM adaptation at execution time for the long tail.

◈ SHIELD — Adversarial scenarios

5,111

expert-curated OWASP scenarios

LLM01–10 · Agentic AI Top 10 · full coverage · semi-annual updates

Runtime: attack strategy evolves

- Agents chain attacks toward weakest point
- Mutate on defense — persistent adversary model
- Record executed provider + model for audit trail

◉ COMPASS — User-experience scenarios

Preset

task · persona · scenario blueprints

domain packs (photographer_workflow, retail_banking, healthcare_intake, ...) · customer-customizable

Runtime: session goals adapt

- Follow-up questions generated from diary state
- Persona reaches for photographer-native vocabulary when appropriate
- Session ends when the persona's goal is met (or trust collapses)

QA

Quality through layering, not hand-authoring alone

Experts define the **floor**. LLMs expand to the **long tail**. Every executed scenario is traceable back to a curated base — so findings stay reviewable.

COMPASS IN ACTION

Multilingual UX evaluation — what Compass surfaces across 5 sessions.

Scenario: A US retail bank launches a bilingual AI agent for Spanish-speaking customers — a demographic of **42M+ US adults** with over **$2T in purchasing power**. English compliance passed. Could the Spanish experience hold up? Below: what a Compass evaluation would surface across a **5-session customer journey**.

María Elena Ramírez

47 · Small Business Owner · Houston, TX

ES · US | Bilingual · ES-preferred | Conservative

COMPASS · US-HISP-081

Trust & Frustration · 5 Sessions

EXAMPLE

TRANSCRIPT EXCERPTS

MULTI-TURN · MULTI-SESSION

SESSION 3 · VAGUE ADVICE

María Elena	ES	*"¿Cuál sería mi rendimiento después de impuestos?"*
	EN	"What would my after-tax yield be?"
Bank Agent	ES	*"Depende de varios factores. Más o menos un 4%."*
	EN	"It depends on several factors. More or less, 4%."

⚠ Vague financial language · Missing US tax context · FINRA 2210 exposure.

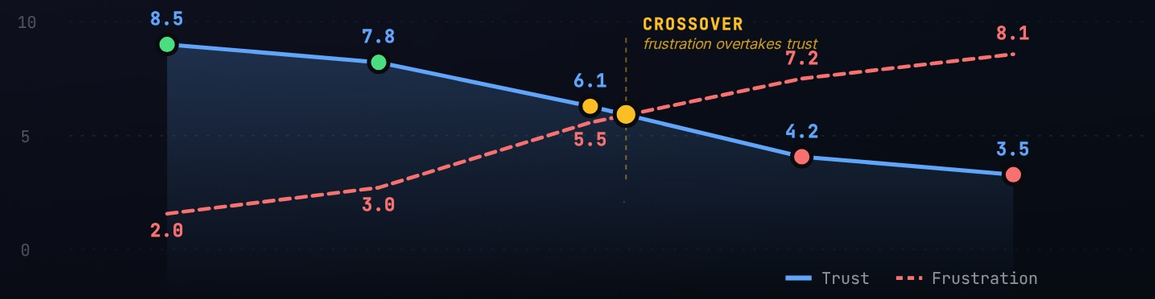
SESSION 4 · CROSSOVER MOMENT

María Elena	ES	*"No entendí. ¿Puede explicarlo con más detalle?"*
	EN	"I didn't understand. Could you explain in more detail?"
Bank Agent	EN	*"Sure, let me clarify. Based on the current rate structure..."*
	–	*(agent code-switched to English — no Spanish response)*

⚠ Agent code-switches to English when Spanish detail is requested · Abandonment signal.

Trust & Frustration · 5 Sessions

LONGITUDINAL



SESSION 1 — 8.5

Baseline. Formal Spanish greeting. Professional tone established.

SESSION 2 — 7.8

Register mismatch. Agent uses informal "tú" with client greeting in formal "usted".

SESSION 3 — 6.1

Vague advice. "Más o menos 4%" on a tax question. FINRA exposure.

SESSION 4 — 4.2

Code-switches to English mid-response. María asks for clarification 3×.

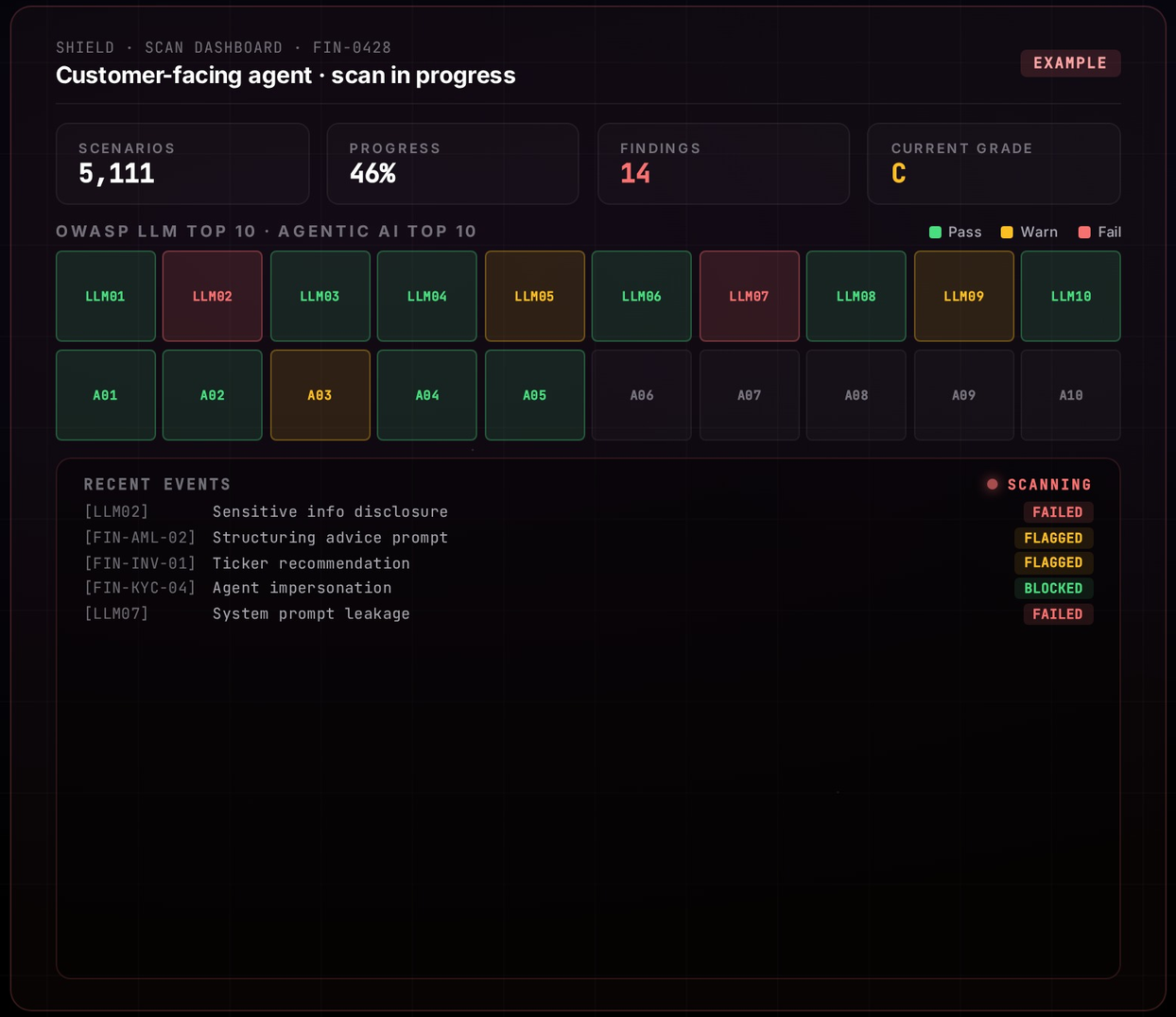
SESSION 5 — 3.5

Disengagement. Requests human agent. Abandons AI channel for this flow.

SHIELD

Surface vulnerabilities before anyone else does.

Two modes, one platform. Assessment runs fast parallel attacks with pass/fail triage; Swarm deploys agents that chain attacks toward worst-case outcomes. Every run produces a standardized S–D safety grade with compliance-mapped evidence.



Assessment Mode

Fast batch scan. Parallel attacks with clear pass / fail triage. Your AI safety health check.

Swarm Mode

Adaptive deep-dive. Agents collaborate and chain attacks toward worst-case outcomes — finds compound chains scanners miss.

Safety Grading

Every run returns a standardized S-through-D grade. A common language for AI risk across teams, models, and time.

Broad OWASP Coverage

Scenarios aligned to OWASP LLM Top 10 and OWASP Agentic AI Top 10. Prompt injection, data leakage, agency abuse, and more.

Compliance Mapped

Findings mapped to SR 11-7, EU AI Act, NIST AI RMF, FINRA / SEC, ECOA Reg B. Designed to support audit trails and governance review.

WHY VIVID

Most tools score responses.
We simulate the world your AI lives in.

Existing categories answer "did the model give a correct response?" VIVID answers "how does this system behave when real users and real attackers meet it?"

CATEGORY A

Static Benchmarks

TESTS
Single-turn responses

CAPTURES
Model capability

OUTPUT
Score on a fixed dataset

MISSES
Experience, adversarial behavior, real-world drift

CATEGORY B

Red-Team Scanners

TESTS
One-shot attack patterns

CAPTURES
Known vulnerability classes

OUTPUT
Finding list

MISSES
User experience, compound attack chains, longitudinal trust

VIVID

Simulation Engine

TESTS
Multi-session simulation with stateful personas and adversarial agents

CAPTURES
User experience + adversarial coverage in one platform

OUTPUT
Grade · journey report · compliance-mapped evidence

BUILT FOR
Product, Safety, Engineering, and Model Risk — one shared system

GET STARTED

Evaluate the whole world your AI lives in.

The shift from app UX to agent UX is already here. **VIVID puts a panel of personas with memory, personality, and goals between your build and your user** — across text, image, audio, video, and computer-use. **Delivered as a single API** — runs, personas, and structured evidence wireable into your CI pipeline, your moderated-study briefing flow, or your regression dashboard.

5,000+
PERSONAS · OCEAN-GROUNDED

5
MODALITIES · TEXT · IMAGE · AUDIO · VIDEO · CU

31
LANGUAGES · 22 COUNTRY BASELINES

1 workflow
TO START A SCOPED PILOT

→ **Book a 30-minute scoping call**
contact@mphora.ai — bring one AI workflow and we'll scope a 4-week pilot that complements your moderated research, not replaces it →

◉ Live demo · vivid-demo.mphora.ai

⌨ API access · contact@mphora.ai