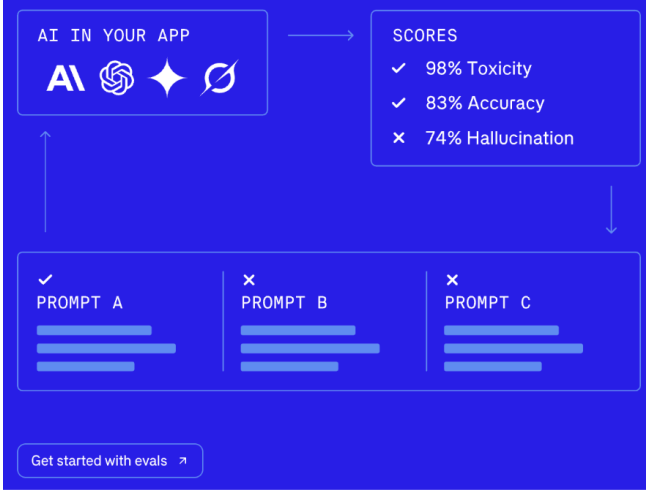
braintrust

Generation Digital

WHY RUN EVALS?

Agents fail in unpredictable ways



How do you know your AI feature works?

Evals test your AI with real data and score the results. You can determine whether changes improve or hurt performance.

Are bad responses reaching users?

Production monitoring tracks live model responses and alerts you when quality drops or incorrect outputs increase.

Can your team improve quality without guesswork?

Side-by-side diffs allow you to compare the scores of different prompts and models, and see exactly why one version performs better than another.

Intuitive mental model

All evals are composed of a dataset, task, and scorers. This framework gives teams a shared understanding for testing and improving AI applications systematically.

Cross-functional collaboration

Engineers write code-based tests. Product managers prototype in the UI. Everyone can review results and debug issues together in real time.

Built for scale

Reliable, fast infrastructure handles high-volume production traffic and complex testing workflows.

ITERATE

Refine prompt and eval ideas fast with playgrounds

Try playground ↗

Fast prompt engineering

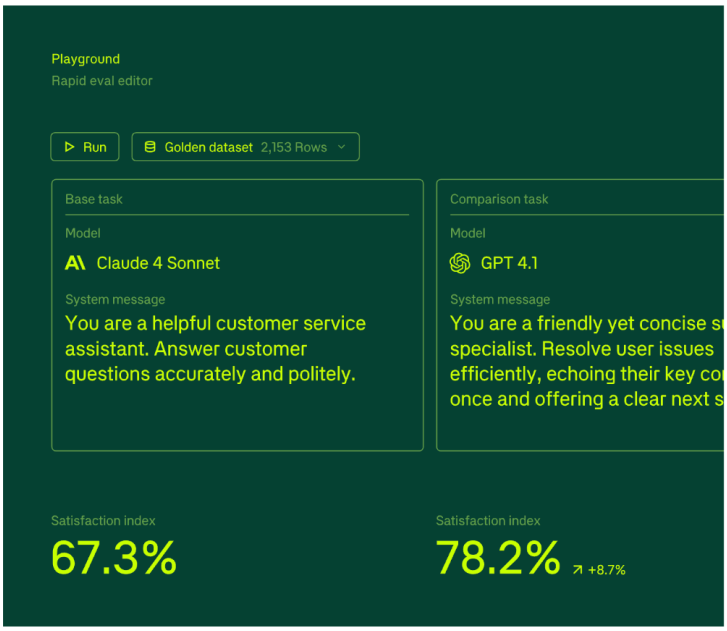
Tune prompts, swap models, edit scorers, and run evaluations directly in the browser. Compare traces side-by-side to see exactly what changed.

Batch testing

Run your prompts against hundreds or thousands of real or synthetic examples to understand performance across scenarios.

AI-assisted workflows

Automate writing and optimizing prompts, scorers, and datasets with Loop, our built-in agent.



EVAL

Run comprehensive tests on every prompt change to measure accuracy, consistency, and safety

Quantifiable progress

Measure changes against your own benchmarks to make data-driven decisions.

Quality and safety gates

Prevent quality regressions and unsafe outputs from reaching users.

Automated and human scoring

Run automated tests on every change, then layer human feedback to capture the nuance machines miss.

SHIP

Track production AI applications with real-time monitoring and online scoring

Live performance monitoring

Track latency, cost, and custom quality metrics as real traffic flows through your application.

Automations and alerts

Configure alerts that trigger when quality thresholds are crossed or safety rails trip.

Scalable log ingestion

Ingest and store all application logs with Brainstore, purpose-built for searching and analyzing AI interactions at enterprise scale.

AI-POWERED WORKFLOWS

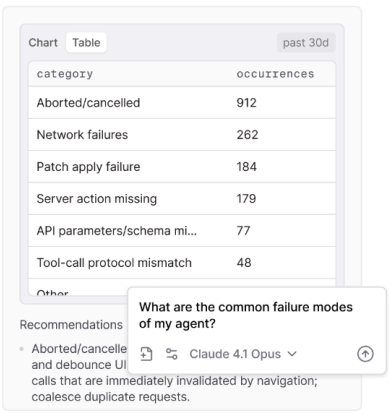
Loop

Loop is an agent that automates the most time-intensive parts of AI development. Available everywhere you work: Logs, Playgrounds, Experiments, and more.

Try Loop ↗

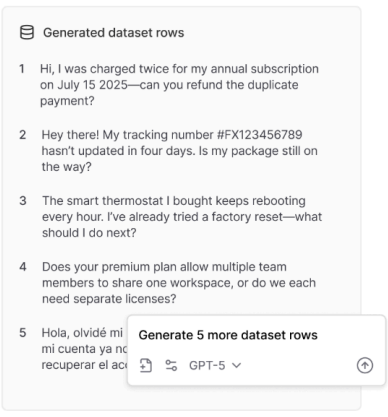
Surface insights

Get instant analysis of common use cases, patterns, and failure modes in your production data.



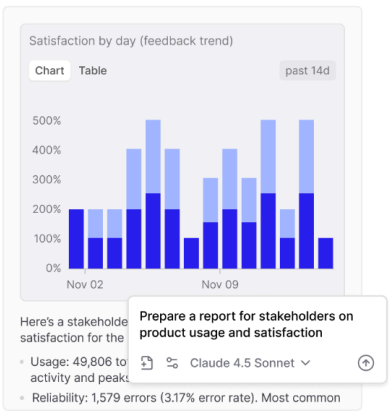
Optimize evals

Create and iterate on prompts, datasets, and scorers with the full context of your project.



Collaborate across teams

Work together through an intuitive interface that makes AI development accessible to everyone.



BUILT FOR ENTERPRISE

Security and compliance at scale

Braintrust meets the demanding security, performance, and collaboration requirements of large organizations building AI applications at scale.

Granular permissions

Role-based access control with org-level permissions and project isolation to meet your security and compliance requirements.

SOC 2 Type II

Third-party security certification with comprehensive security controls.

Trust center ↗

Hybrid deployment

Self-hosting options to maintain full control over your AI data and meet strict compliance requirements.