AI Subscription Proposal for On Call Agent Development and AdBrew Intelligence Evaluation

Prepared for Engineering Leadership
13 September 2026

Decision requested Approve one $100 monthly ChatGPT Pro seat for the engineer developing and evaluating the on-call agent and the AdBrew Intelligence eval stack. Keep production agent inference under a separate operational budget. If procurement prefers API access instead of a developer subscription, begin with a $200 monthly development and evaluation cap because the measured testing workload already exceeds $100 at API-equivalent rates.

Why the usage must be separated

The production on call agent and the work used to improve it are different workloads. Production usage serves Jira and Slack requests. Development usage includes code changes, debugging, benchmark replays, reviewer tests, quality grading, and regression checks. Combining them hides the cost of improving reliability and makes a busy benchmark week look like a production spike.

Usage pool	What it covers	How it should be funded
Production agent	Live investigations, deep RCA, reviewer and summary passes, duplicate checks, and manual on call conversations	Separate service budget using the current production providers or provider APIs
Development and evaluation	Coding, code review, benchmark setup, replaying historical incidents, grading outputs, and building the AdBrew Intelligence eval stack	One named ChatGPT Pro developer seat, with API credits only when unattended automation is required

Measured on call agent usage

The following figures come from provider execution records on brewing 2 for the rolling 30 days ending 13 September 2026. Cost is the providers' reported API equivalent amount. Subscription billing may differ, but the figures show the relative size of each workload.

Workload	Execution artifacts	New input tokens	Cached input tokens	Output tokens	API equivalent
Production	413	40.0M	296.1M	3.64M	$223.13
Benchmarks	613	90.5M	476.0M	7.52M	$157.06
Audits and grading	114	21.8M	111.3M	1.80M	$18.71
Development and evaluation total	727	112.3M	587.3M	9.32M	$175.77

These are execution artifacts rather than ticket counts. A single ticket may create an investigation, deep RCA, reviewer, summary, and duplicate-check artifact. This is the correct unit for token and cost planning because each phase makes a separate model call.

Typical model work during evaluation

Phase	Median new input	Median cached input	Median output	Median API equivalent
Investigation or classifier	176K	877K	15.6K	$0.18
Deep RCA	225K	1.35M	21.9K	$0.23
Reviewer	21.6K	597K	7.4K	$0.19
Compact summary	214K	0	0.8K	$0.15

A full benchmark ticket can use several of these phases. The cost also varies with incident complexity, tool usage, retries, model choice, and how much of the prompt is served from cache.

AdBrew Intelligence already has the telemetry base

AdBrew Intelligence already records the operational data needed to build an eval system. Each analysis run stores its model, reasoning effort, prompt-template hash, new and cached tokens, reasoning tokens, cost, duration, tool calls, failed tool calls, schema repair, finalization reason, run status, and number of insights produced. CloudWatch metrics also cover invocation counts and latency, model-call counts and errors, and tool-call counts and latency.

The missing layer is quality evaluation. We still need a versioned test set, expected behaviors, automated graders, human-reference comparisons, release-to-release scorecards, and a dashboard that joins quality with cost and latency. The on call agent benchmark ledger and reviewer flow provide a practical starting pattern, but the AdBrew Intelligence criteria must measure recommendation correctness, evidence use, duplication, usefulness, and whether an insight is safe to show or apply.

Development use cases funded by the seat

• Implement and review changes across the on call controller, provider adapters, evidence tools, benchmark runner, reviewer, dashboard, and deployment scripts.

• Replay historical incidents without posting to Jira or Slack, compare candidate and production releases, and inspect regressions before deployment.

• Build the AdBrew Intelligence evaluation dataset, graders, release comparisons, cost reports, and failure analysis using its existing analysis run telemetry.

• Run occasional second-provider checks for difficult architecture or correctness decisions without making those providers mandatory for ordinary development.

Recommended budget

Item	Monthly amount	Recommendation
ChatGPT Pro developer seat	$100	Approve now. Use for human-supervised coding, investigation, review, benchmark orchestration, and eval development.
Production model runtime	Separate	Keep outside the developer seat. The last 30 days recorded $223.13 of provider-reported API equivalent production usage.
API-only development alternative	$200 cap	Use only if a subscription is not approved or if scheduled unattended evals must run immediately. Review after one month.

OpenAI currently offers a $100 Pro tier with the same core Pro capabilities and a higher allowance than Plus. OpenAI also states that Codex commonly costs about $100 to $200 per developer per month, depending on the model, parallel work, automations, and speed. This makes the $100 tier a reasonable starting point for one developer, with a usage review after 30 days.

How we will control spend

• Tag production, benchmark, audit, and AdBrew Intelligence eval usage separately.

• Track new input, cached input, output, model, cost, duration, and result status for every run.

• Use smaller models for classification and routine grading, and reserve stronger models for disputed or high-risk cases.

• Use no-posting benchmarks and fixed sample sets for release checks so repeated tests remain comparable.

• Review actual usage after 30 days and change the plan or API cap only from measured demand.

Expected first month output

During the first month, the seat should support continued on call reliability work and the first usable AdBrew Intelligence eval loop. The concrete outputs are a versioned eval dataset, repeatable candidate-versus-production runs, automated quality and safety grades, cost and latency reporting, and a release report that links each score to the underlying run evidence.

Approval requested

Approve one $100 monthly ChatGPT Pro seat for the engineer responsible for on-call agent and AdBrew Intelligence eval development. Keep live agent inference under its existing operational budget. If the team chooses API-only access, approve an initial $200 monthly development and evaluation limit and review it after the first complete month of tagged usage.

Sources

• OpenAI ChatGPT Pro tiers
• OpenAI ChatGPT and Codex rate card
• OpenAI API model pricing

Internal measurement sources include the on call agent provider event ledger on brewing 2 and the AdBrew Intelligence analysis run, model callback, and CloudWatch telemetry code in the AdBrew repository.