
GPT-6 ASTRA & AI CAPABILITIES

Choose the work. Check the result.

GPT-6 Astra in Practice: A Decision Guide for Growth Teams

About this edition

Public research, practical methods and clearly labelled examples. Prepared with AI assistance; this edition reports no original Ranketize experiment results.

Read the current web edition and follow its source links for changing platform guidance.

<https://ranketize.com/resources/gpt-6-astra-growth-guide>

Contents

1. Evaluate the working system
2. Make the AGI discussion operational
3. Connect capability evidence to field productivity
4. Six task briefs worth testing
5. Run a pilot that reveals failure
6. Measure human intervention and cost
7. An illustrative decision, with visible assumptions
8. Sources and edition boundaries

Using the worksheets

Copy the tables from the web edition or download its Markdown text. Record your own observations in the blank fields; any figures in worked examples are illustrative.

Growth teams should evaluate GPT-6 Astra against a specific handoff: a piece of work that someone else can check, accept and use. A polished answer is an intermediate result. The useful outcome might be a defensible market brief, a reconciled analysis or a tested website change.

This first edition combines public documentation and independent research with a proposed evaluation method. Sources were checked on **9 September 2026**. Ranketize has not run the 18 pilot tasks proposed below. The workflows, thresholds and worked example are practical recommendations and illustrations, not measured Astra results.

1. Evaluate the working system

OpenAI documents GPT-6 Astra for reasoning, coding, research, computer use and document creation. Its API model page lists text and image input, text output, structured outputs and access to supported tools. These are product capabilities; they do not establish the accuracy of a particular business deliverable. [OpenAI model documentation](#)

The accompanying guide describes asynchronous tool calling and instructions that can steer work while it is running. In an API application, the application still executes tools and manages pending work. A product must implement these capabilities for its users to benefit from them. [OpenAI model guidance](#)

For a marketing director, the relevant unit of evaluation is therefore the complete arrangement: model, instructions, source material, available tools and review process. Two teams can select the same model and produce different results because one provides approved product facts and the other provides an ambiguous website address.

Before the first task, record the product used, displayed model name, date, available tools and relevant settings. An API result, a Codex task and a consumer chat should remain separate records. Attach the input files and the final accepted deliverable. When a model, source pack or tool changes, that record lets you identify what actually changed.

Start with a bottleneck that already exists. If sales cannot use the current competitor brief because its claims lack evidence, generating five more briefs does not resolve the bottleneck. The experiment should test whether one brief becomes both faster to produce and easier to trust.

2. Make the AGI discussion operational

AGI is an unsettled category. Morris and colleagues propose a framework that separates depth of performance from breadth of generality, then considers deployment autonomy separately. Their paper is a position and classification framework, not a consensus standard or an evaluation of Astra. It also explains why a capable system can be deployed with different degrees of human control. [Levels of AGI, version 5](#)

This distinction is useful in a growth team. Reading a research paper, editing a spreadsheet and changing a webpage demonstrate breadth across activities. They do not by themselves establish dependable performance across the work those activities contain. Nor does the ability to make a change settle who should authorize it.

Translate broad claims into three operating questions:

Dimension	Question to answer with evidence
Performance	Does the deliverable meet the same standard as accepted human work?
Generality	Does that standard hold across different inputs, exceptions and task types?
Autonomy	Which actions can proceed without review, and how does the system handle uncertainty?

These questions adapt the framework for workflow design; they are not an AGI certification. A successful pilot can justify a bounded delegation decision. It cannot settle whether a model has reached general human competence.

Make autonomy explicit in the brief. Permission to inspect a draft page is different from permission to edit it; editing permission is different from permission to publish. A team can expand the first two as evidence improves while retaining a human publishing decision. That is a practical allocation of responsibility, regardless of the label attached to the model.

3. Connect capability evidence to field productivity

Benchmark results and workplace outcomes answer different questions. METR's task-completion horizon estimates the human-expert task duration associated with a chosen model success probability. Its 50% and 80% horizons measure difficulty on a task distribution, rather than how long an agent can safely run unattended. The suite is primarily software engineering, machine learning and cybersecurity, with relatively well-specified tasks. It is not a measure of marketing-team autonomy. [METR time-horizon methodology, updated 8 May 2026](#)

Field studies show why the operating context matters. METR's July 2025 randomized study assigned 246 real issues from 16 experienced open-source developers to conditions allowing or disallowing AI. Tasks with AI took 19% longer in that setting. The tools were from early 2025; the result is historical evidence about that population and workflow, not an Astra estimate. [METR's 2025 study](#)

METR's February 2026 follow-up complicates any simple slowdown story. Developers' willingness to participate, task selection and time measurement with concurrent agents introduced problems. The researchers describe the newer data as an unreliable signal of the current productivity effect and weak evidence about its size. A preliminary point estimate should not become a confident current speedup claim. [METR's 2026 update](#)

Different work has produced different findings. Brynjolfsson, Li and Raymond's revised study covers 5,172 customer-support agents and reports a 15% average increase in issues resolved per hour after AI assistance, with effects varying by experience and skill. This concerns a different tool, workforce and outcome. Its percentage cannot be combined with METR's figures to predict an Astra return. [Generative AI at Work, version 2](#)

Our interpretation is that useful evaluation needs both capability and operating evidence. Capability evidence helps select tasks worth attempting. Local measurement establishes whether context preparation, verification and correction consume the apparent benefit. Record changes in quality as well as time: a faster brief that introduces an unsupported competitor claim has failed an essential requirement.

4. Six task briefs worth testing

Each brief below specifies a deliverable, evidence and acceptance criteria. Use approved business materials and a workspace where the task owner can inspect changes. Choose tasks that correspond to real recurring work; do not create busywork solely because a model can perform it.

1. A decision-ready competitor brief

Brief: Compare three named alternatives for one buyer segment and decision. Supply your approved product facts, the buyer's requirements and dated competitor pages. Request a concise comparison with a source beside each externally verifiable claim, plus unresolved questions.

Accept when: Every comparison uses the same criteria; current and historical facts are separated; unavailable information remains unknown; conclusions follow from the buyer's requirements. Reject invented pricing, inferred product limitations presented as facts, and unsupported claims of superiority. The owner checks the claims that could change the buying decision.

2. A reconciled acquisition analysis

Brief: Use approved exports to explain a defined change in qualified enquiries between two periods. Provide definitions, timezone, exclusions and the business question. Request calculations, a reconciliation to input totals and competing explanations.

Accept when: Totals reconcile or discrepancies are quantified; denominators accompany rates; duplicates and missing values receive explicit treatment; observed patterns are distinguished from causal explanations. A second person reproduces the headline calculation. Recommendations must specify what further observation would distinguish the competing explanations.

3. A source audit for an important page

Brief: Review one draft commercial or educational page against supplied sources. Produce a claim ledger with the exact sentence, supporting passage, date, limitation and proposed correction. Keep stylistic preferences separate from evidence problems.

Accept when: Material claims are accounted for, citations support their actual wording, and numbers retain their original population and denominator. Outdated claims are replaced only when newer evidence supports the replacement. The editor can trace each change back to the source without reconstructing the whole research process.

4. A synthesis of customer objections

Brief: Analyze an approved, appropriately prepared set of interview notes or call transcripts for one segment. Request objections, triggers, purchasing criteria and contradictions, with references to the underlying records. State how many records exist and how they were selected.

Accept when: Themes link to evidence; dissenting examples remain visible; frequency refers only to this sample; quotations are accurate and permitted for the intended use. The result must not describe a small

convenience sample as a market-wide survey. The owner checks whether the suggested follow-up questions address real gaps.

5. A website improvement ready for review

Brief: Implement one bounded improvement in an isolated branch or preview. Supply the intended behavior, existing components, accessibility expectations and routes affected. Request the change, a short explanation and evidence from relevant checks.

Accept when: The target behavior works at supported screen sizes; links and keyboard interaction work; required checks pass; unrelated behavior is preserved. Content remains accurate. The reviewer can inspect a working preview and understand any unresolved limitations. Publishing belongs to the separately defined approval step.

6. A weekly decision memo

Brief: Combine approved analytics, campaign notes and customer evidence into a one-page memo. Ask for material changes, possible explanations and at most three proposed decisions. Every decision should have an owner, evidence, uncertainty and next observation.

Accept when: Numbers trace to source records, important contradictory signals remain visible, and proposed actions fit actual constraints. “Do more content” fails because it leaves scope and rationale unspecified. A useful recommendation identifies the buyer question, missing evidence, intended page and observation that would justify continuing.

5. Run a pilot that reveals failure

The proposed first pilot contains **18 task instances: three for each of the six briefs**. It has not been run. Its purpose is to expose workflow problems and decide where further testing is worthwhile, not to estimate a universal productivity effect.

For each brief, choose a routine instance, an instance with missing or conflicting information, and one with a meaningful exception. Set acceptance criteria before seeing outputs. Include the normal human process as the comparison, with its existing tools and final review, so the pilot measures a realistic alternative.

Where feasible, allocate comparable tasks to AI-assisted and existing workflows without selecting the easiest tasks for AI. Avoid having the same person immediately repeat an identical task in the second condition: remembered answers contaminate the comparison. If tasks cannot be matched credibly, report descriptive cases and their limitations.

Use three judgments: accepted, accepted after correction, and rejected. Also record whether any critical error occurred, such as an unsupported material claim, incorrect denominator or unauthorized action. An attractive final document should not hide the correction history.

Have a person with the relevant expertise evaluate the artifact. The same model can help identify potential problems, but its approval is not independent evidence. Hide the production method from the reviewer where practical and record disagreements.

Preserve failed runs, abandoned attempts and retries. Count the preparation and repair needed to reach the accepted output. Compare task types separately before considering a total; a research workflow and a website change may have different failure patterns.

A reasonable next decision is narrow: continue a particular workflow under specified review conditions, revise its inputs, or stop using it. Broader delegation needs more observations, including repeated runs and failures that matter in production.

6. Measure human intervention and cost

Use one worksheet per task. Time should capture active effort and elapsed delivery time separately: an agent can finish while its owner is unavailable, and a fast draft can wait a day for review.

Field	What to record
Identity	Task, date, product/model, tools, settings, input versions
Preparation	Human minutes collecting context and defining the task
Production	Human active minutes and agent elapsed time, recorded separately
Intervention	Count, reason, active minutes and consequence of each intervention
Review	Minutes checking facts, calculations, behavior and presentation
Correction	Human minutes, retries and final outcome
Direct expense	Actual model, tool and attributable service charges
Acceptance	First-pass decision, final decision, critical errors and reviewer
Comparison	Existing-process cost, quality and elapsed time; comparability limits

Classify interventions by cause: missing context, unclear objective, unsupported evidence, incorrect work, tool failure or required business judgment. The reason matters. More detailed inputs may resolve missing context; repeated incorrect calculations call for a different control.

For internal cost comparison, use:

Task cost = active human hours × the team's chosen hourly cost + attributable model and tool expense.

Keep the hourly-cost assumption visible. Allocate reusable setup separately so a one-off experiment is not confused with steady operation. For multiple tasks, divide all task costs, including failed attempts, by the number accepted. Report acceptance rate alongside that figure. Lower cost with unacceptable quality does not meet the objective.

Use actual invoices or usage records for direct expense. Rates, account access and processing choices can change; a printed specification table will age faster than this worksheet. Freed staff time is capacity, not automatically a cash saving or additional revenue. Record what the team actually did with it before claiming a business return.

7. An illustrative decision, with visible assumptions

The following figures are invented solely to demonstrate the worksheet. They are not a client result, an Astra benchmark or a prediction.

A team compares two ways to produce one source-checked competitor brief. Both deliverables must satisfy the acceptance criteria in chapter four. Its chosen internal labor cost is €60 per active hour.

Component	Existing process	Illustrative AI-assisted process
Preparation	15 minutes	20 minutes
Production: human active time	95 minutes	15 minutes
Review	20 minutes	25 minutes
Correction	10 minutes	15 minutes
Total human active time	140 minutes	75 minutes
Attributable direct tool expense	€0	€8
Calculated task cost	€140	€83

Under those assumptions, the difference is €57 for one accepted task. The calculation excludes any initial setup investment and says nothing about elapsed delivery time, actual revenue or performance on future tasks.

Now test the fragile assumption. If the AI-assisted version requires another hour of verification, its cost becomes €143. The initial apparent advantage disappears. If the output is rejected, its cost still belongs in the pilot total, and replacement work must also be counted.

This sensitivity check directs the next experiment: investigate why verification takes time. Was the evidence difficult to obtain, poorly referenced or incorrectly interpreted? The useful response may be a better source pack or a narrower brief. Simply increasing the number of generated pages would leave the underlying problem intact.

Before expanding, require evidence of accepted work across the relevant exceptions, a manageable review burden and clear ownership when a task fails. Recheck the workflow after material changes to the model, tools or business inputs. The lasting asset is a process for deciding what can be delegated and checking that it continues to work.

8. Sources and edition boundaries

All sources were accessed on 9 September 2026. The official documentation is a changing product reference. Research findings retain their original dates, populations and limitations.

- OpenAI: [GPT-6 Astra model documentation](#) and [model guidance, Astra section](#).
- Morris and colleagues: [Levels of AGI, version 5](#), revised September 2025.
- METR: [Task-completion time horizons](#), updated 8 May 2026.
- METR: [Experienced developer productivity study](#), 10 July 2025, and [experiment-design update](#), 24 February 2026.
- Brynjolfsson, Li and Raymond: [Generative AI at Work, version 2](#), revised 6 November 2024.

The evaluation design and worksheets are Ranketize's proposed practical application of these distinctions. This edition contains no original Astra performance dataset, measured Ranketize productivity gain or determination that AGI has been achieved.