

# PRIVATE AI SERVERS

## Evaluate dedicated AI capacity against the work, data boundaries, performance, and total cost.

Client Education Series

An exclusive publication of Sandbox Technologies, Inc.


Sandbox Technologies
3868 W Carson St, Suite 316
Torrance, CA 90503
www.sandboxtech.com
Tel. 424.207.5130


SANDBOX TECHNOLOGIES

Small Business Solutions
Enterprise GrowthPath®
EGP Secure®
EGP Cloud™
ConstructIT®

# Private AI Servers

Evaluate dedicated AI capacity against the work, data boundaries, performance, and total cost.

## What private hosting provides

A private AI server runs a model in an environment reserved for your organization. It may be located in your office, a private data center, or a controlled cloud environment. The design determines which people, documents, applications, and outside services it can access.

Private hosting gives the organization more control over those boundaries. It does not establish privacy or compliance by itself. Permissions, network connections, logs, backups, software updates, and any external integrations must be included in the design and checked against the requirements governing the information.

## Match capability to the actual task

Frontier models are those at or near the leading edge of available AI capability at a given time. A privately hosted model may perform well on selected work without matching a leading hosted service across every task. Compare the configured model with the tools you would otherwise use, using representative work and review criteria. Models available for private hosting can be useful for document retrieval, drafting, classification, and defined workflows. Evaluate them with representative work, including difficult cases and exceptions. A model that performs well on one task may be a poor choice for another. Review where the model comes from, its license and permitted business uses, the software it depends on, and who will deploy, update, and maintain it. Capability is only one part of selecting a model for business use.

Hardware affects response time, the amount of model and context data that can fit in memory, and the number of people who can work at once. A model and its software also impose limits on context length. It does not by itself establish accuracy. Where a model supports reasoning, allowing more processing before an answer may improve results on difficult tasks, with longer waits and less capacity for simultaneous requests. Testing must include the configured model and operating mode. Generating more reasoning tokens requires additional computation and generally increases response time and energy use for that request. Longer sequences may also increase working-memory requirements, depending on the implementation. Reasoning does not make the stored model parameters larger. Measure the actual tradeoff between answer quality, waiting time, memory, and simultaneous requests.

SANDBOX TECHNOLOGIES

Copyright 2026, Sandbox Technologies, Inc.

## Compare deployment and operating costs

Include hardware or hosting, storage, security, backups, maintenance, power, and support. The number of users alone is not enough to size a system. Document volume, model size, response expectations, and concurrent workloads also matter. Compare an on-premises purchase with private hosting and managed services. The better choice depends on workload, operating responsibilities, information requirements, and total cost over time.

A private model can work alongside an outside service when the information boundaries are explicit. For example, an outside model could help design a filing procedure using invented sample data, while a private process applies the procedure to confidential files. Check that actual files and sensitive details cannot leave through prompts, logs, or integrations. Enforce the intended boundary through network and access controls, then test it. A script described as private does not by itself establish that information stays inside the environment. Permissions, review points, and recovery from errors remain necessary.

We're here to help and have structured services to address each of these areas individually or as part of a coordinated plan. To discuss what deserves attention first, contact your Sandbox Technologies Account Manager or Consulting CIO.

# Historical Benchmark Illustration

This comparison from the earlier guide is retained as a historical illustration. It is not a comparison with the newest models or a measurement of a proposed client server. Qwen3-32B was released April 29, 2025. The GPT-4o row is its November 2024 update. These are results reported by the model developer in its technical report, not an independent comparison or a test of a client installation.

| Model and mode | General knowledge MMLU-Redux | Advanced science GPQA Diamond | Coding LiveCodeBench v5 |
|---|---|---|---|
| Qwen3-32B non-thinking | 85.7 | 54.6 | 31.3 |
| Qwen3-32B reasoning enabled | 90.9 | 68.4 | 65.7 |
| GPT-4o November 2024 | 87.0 | 46.0 | 32.7 |

The tests cover broad multiple-choice knowledge, expert science questions, and programming problems checked by running the code. Scores describe those evaluations, not the percentage of everyday work a model will complete correctly. The two Qwen rows use the same model with different processing modes.

Source: Qwen3 Technical Report, Tables 12-14. https://arxiv.org/html/2505.09388v1#S4



Copyright 2026, Sandbox Technologies, Inc.