

FIVE BUILDS. ONE BUYING FRAMEWORK.

An AI computer, from desk to business hub

A complete purchasing pack for local AI and shared business assistants. Each build includes hardware requirements, a cost worksheet, model limits and ordinary vendor links.

BUILD	EQUIPMENT	LOCAL MODEL TARGET
Essential desk One primary user	\$2,500 Worksheet: \$2,420	8B-14B local models
Balanced desk One primary user	\$5,100 Worksheet: \$5,080	30B-32B local models
Local AI studio One power user; small-team pilot possible	\$8,600 Worksheet: \$8,630	70B quantized models
Shared business hub A small or medium business team	\$28K-\$35K Worksheet: \$30,500	30B-32B default; 70B with testing
Enterprise AI hub Enterprise; central IT and facilities required	\$600K-\$750K Worksheet: \$650,000	Very large low-precision models; benchmark first

Start with the daily work

The Balanced desk is the default recommendation for one business owner who wants useful local AI capability and a high-quality 34-inch workspace. Choose a shared hub when multiple employees need a centrally managed assistant and document service.

Three distinctions that prevent expensive mistakes

Model size is not answer quality. A well-configured smaller model can be more useful than a much larger model with long delays. Use representative business tasks to compare quality and latency.

Memory fit is not simultaneous capacity. Weights, context, retrieval and concurrent requests all consume resources. A 96GB GPU is separate from host system RAM; Apple unified memory is a different architecture.

Free weights do not create free operations. Model licenses differ. Local inference can avoid token charges, while power, cooling, backup, software integration, support and people still have costs.

Using this pack

Each build covers hardware and use cases, capability and deployment, procurement, and source references. CSV shopping lists accompany the individual PDFs. Vendor links are clickable. All figures are USD planning allowances, not final purchase quotes.

Ordinary merchant links only. Affiliate tracking has not been activated.

A BUSINESS NETWORK, NOT JUST A BIG COMPUTER

Make the hub useful and trustworthy

Employees keep using their existing computers. The shared system hosts an authenticated web application, the language model, and a permission-aware document search service.

EMPLOYEE BROWSER	HTTPS + LOGIN	MODEL + DOCUMENT SEARCH	APPROVED BUSINESS FILES
------------------	---------------	-------------------------	-------------------------

Deployment requirements

- 1. Choose the workload.** Name the questions, approved files, tools and outcomes the service must support. Start with document Q&A; and drafting before unattended changes.
- 2. Protect access before retrieval.** Authenticate each person. Filter documents by permissions before they are retrieved; keep the raw model endpoint private. Restrict tools to the authority that user actually has.
- 3. Queue and measure.** Pilot with the exact checkpoint, quantization, context length and concurrent requests. Agree on latency and answer-quality targets. Account count is not active-request capacity.
- 4. Plan operation and recovery.** Assign an owner for patches, monitoring, backups, restore tests and access changes. One machine is a single point of failure; high availability needs additional hardware and design.
- 5. Validate facilities and support.** Use wired networking for the hub. Size UPS and circuits from measured requirements. Enterprise racks require professional power, cooling, installation and an explicit support agreement.

Keep these three budgets separate

Equipment: the worksheet total. **Setup:** deployment, migration, integration and training. **Operations:** energy, cooling, backup, support, staffing and any optional cloud tools.

For the shared business hub, setup is an editorial \$3,000-\$10,000 allowance; enterprise integration is \$50,000-\$150,000. Obtain scoped quotes. Enterprise facilities allowances in the equipment worksheet also require site-specific confirmation.

No performance benchmark was run for this guide. Very large models and pre-release runtimes require vendor acceptance testing before purchase.

01 / A STRONG START

Essential desk

A quiet, capable business desk with a real ultrawide monitor. Cloud agents do the heavy lifting; smaller models stay local.

\$2,500

Equipment planning budget

8B-14B local models

One primary user



01 Specify the machine

Processor	Apple M6: 12-core CPU, 12-core GPU
Memory	32GB unified memory; specify at purchase
Storage	512GB SSD; add external model storage if needed
Display	Dell S3425DW: 3440 x 1440, up to 120Hz, VA, height adjustable
Connection	USB-C display cable; confirm included cable and USB adapters
Software	macOS; current Ollama or LM Studio; compatible agent application

Where it earns its place

- + Email drafting and document summaries
- + Business research with cloud agents
- + Local drafting, classification and short-document Q&A

Network role

Can serve a local model on a private network after proper configuration, but this budget is for one primary user. Model examples describe memory capacity, not measured response speed. All users share the same compute.

CAPABILITY & DEPLOYMENT

What this build can do

Choose a model and workload first. Memory fit, useful speed and simultaneous demand are different purchasing questions.

Representative model targets

MODEL	FIT / CONDITIONS
Qwen3 8B / 14B	Good fit Approximately 5.2GB / 9.3GB weights, plus context and apps.
Qwen3 32B	Constrained 20GB weights; short context and limited multitasking only.
Llama 3.3 70B	Not recommended 43GB weights exceed comfortable memory capacity.

Can support

- + Run cloud AI agents and ordinary business applications.
- + Run compatible downloadable models with no per-token inference bill.
- + Search approved local documents when a retrieval workflow has been configured.
- + Use 8B-14B quantized models while leaving space for normal work.

Cannot do

- + Automatically become a reliable business agent just by installing a model.
- + Replace tested workflows, backups, human review or account permissions.
- + Make paid cloud services free, or guarantee the quality of every answer.
- + Comfortably serve 70B models or several demanding local requests at once.
- + Treat a 32B model that barely fits as a comfortable long-document workstation.

Concurrent work: pilot before scaling

One local generation at a time

Separate setup and operating budgets

Setup / integration. Workflow installation, migration and training are separate; obtain a scoped quote.

Ongoing operation. Local inference has no token bill. Allow \$50-\$150/month if adding paid cloud AI and off-site backup; electricity and internet additional.

Free model weights can remove per-token model charges. They do not remove hardware, power, administration or paid service costs. Review the exact model license before business use.

PROCUREMENT WORKSHEET

Build the purchase order

Copy these requirements into a vendor request. Amounts are budget allowances unless a dated listing is explicitly identified; no cart or binding quote has been prepared.

QTY	ITEM / CONFIGURATION	LINE TOTAL
1	Mac mini, configured 1 unit; M6 / 32GB / 512GB. Rounded base + upgrade estimate.	\$1,500
1	Dell S3425DW 1 monitor; stand included. Allowance exceeds observed sale price.	\$420
1 set	Logitech Pebble 2 Combo for Mac 1 keyboard/mouse set; Bluetooth; observed \$59.99. Line total is a planning allowance.	\$80
1 each	Logitech Brio 100 + Logitech H390 USB-C headset 1 1080p webcam with privacy shutter; observed \$39.99; USB-A adapter available below. 1 headset; select USB-C variant 981-001575, not USB-A; observed \$24.99 regular. Line total is a planning allowance.	\$100
1	WD My Passport 2TB HDD 1 dedicated backup HDD; observed \$174.99. Includes cable + USB-C adapter; format for Time Machine. Line total is a planning allowance.	\$180
1	CyberPower ST625U 1 UPS; observed MSRP \$84.95. Verify combined watts and required shutdown time before purchase. Line total is a planning allowance.	\$100
1 each	Monoprice Cat6 FLEXboot 7ft gray + Anker USB-C to USB-A adapter 1 Ethernet cable; \$10 planning allowance, not verified live price; choose length for desk location. 1 pack; observed regular \$12.99; adapters for USB-A peripherals. Use supplied monitor and drive cables first. Line total is a planning allowance.	\$40
EQUIPMENT TOTAL		\$2,420

Before ordering: confirm macOS works with your required business software. Prices exclude tax, shipping, furniture, warranty upgrades and setup labor. Worksheet total: \$2,420; the \$2,500 headline is rounded.

Before placing the order

Confirm required business software works on macOS; keep the specified memory. Check monitor refresh/cable compatibility, backup capacity and UPS watts. For a local-model pilot, record the exact model/quantization, working context, time to first token and a representative document task.

Excludes tax, shipping, furniture, warranty/support upgrades, paid software and setup labor unless explicitly included above. A single system provides no automatic failover.

ORDER LINKS & SOURCES

Find the specified equipment

Links identify the recommended products. Check the selected variant, current price and delivery estimate before ordering. Group allowances on the previous page remain the purchasing budget.

Where to buy or request a quote

[Configure at Apple](#)

Choose M6 / 32GB / 512GB; do not buy the base 16GB configuration.

[Monitor at Dell](#)

Check current sale price and included cable.

[Logitech Pebble 2 Combo for Mac · 920-012202](#)

1 keyboard/mouse set; Bluetooth; observed \$59.99.

[Logitech Brio 100 · 960-001616](#)

1 1080p webcam with privacy shutter; observed \$39.99; USB-A adapter available below.

[Logitech H390 USB-C headset · 981-001575](#)

1 headset; select USB-C variant 981-001575, not USB-A; observed \$24.99 regular.

[WD My Passport 2TB HDD · WDBWML0020BGY-WESN](#)

1 dedicated backup HDD; observed \$174.99. Includes cable + USB-C adapter; format for Time Machine.

[CyberPower ST625U · 625VA / 360W](#)

1 UPS; observed MSRP \$84.95. Verify combined watts and required shutdown time before purchase.

[Monoprice Cat6 FLEXboot 7ft gray · 9798](#)

1 Ethernet cable; \$10 planning allowance, not verified live price; choose length for desk location.

[Anker USB-C to USB-A adapter · 2-pack](#)

1 pack; observed regular \$12.99; adapters for USB-A peripherals. Use supplied monitor and drive cables first.

Source references

Apple configurations	Display specification
Representative local model sizes	70B sizing
Context memory	

Ordinary merchant links only. Affiliate tracking is not activated. Prices checked October 16, 2026.

02 / OUR EVERYDAY RECOMMENDATION

Balanced desk

The most balanced choice for a business owner who wants a polished workspace and useful room to grow into local AI.

\$5,100

Equipment planning budget

30B-32B local models

One primary user



01 Specify the machine

Processor	Apple M5 Pro: 18-core CPU, 20-core GPU
Memory	64GB unified memory; do not substitute 24GB
Storage	1TB internal SSD
Display	Dell U3425WE: 3440 x 1440, 120Hz, IPS Black, Thunderbolt hub
Connection	Thunderbolt / USB-C display; Ethernet preferred for stationary use
Software	macOS; current Ollama or LM Studio; cloud agent + local model option

Where it earns its place

- + Business planning, proposals and research
- + Local document Q&A with citations to selected files
- + Coding assistants and lightweight automation
- + A daily desk that can also experiment with private AI

Network role

A small private-network pilot is possible. Use a request queue; do not promise a team-wide service level from this personal desktop. Model examples describe memory capacity, not measured response speed. All users share the same compute.

CAPABILITY & DEPLOYMENT

What this build can do

Choose a model and workload first. Memory fit, useful speed and simultaneous demand are different purchasing questions.

Representative model targets

MODEL	FIT / CONDITIONS
Qwen3 8B / 14B	Good fit Useful for quick tasks and modest simultaneous work after testing.
Qwen3 32B	Recommended size 20GB weights leave more room for context, retrieval and apps.
Llama 3.3 70B	Conditional 43GB weights can fit with careful runtime/context limits; not the everyday sizing target.

Can support

- + Run cloud AI agents and ordinary business applications.
- + Run compatible downloadable models with no per-token inference bill.
- + Search approved local documents when a retrieval workflow has been configured.
- + Make 30B-32B quantized models a practical local target.
- + Keep more normal applications open alongside local inference.

Cannot do

- + Automatically become a reliable business agent just by installing a model.
- + Replace tested workflows, backups, human review or account permissions.
- + Make paid cloud services free, or guarantee the quality of every answer.
- + Promise fast 70B responses with long context or concurrent employees.
- + Run the largest downloadable models on a 64GB memory pool.

Concurrent work: pilot before scaling

Start with one; queue heavier jobs

Separate setup and operating budgets

Setup / integration. Agent workflows, file organization, migration and training need a separate scope and quote.

Ongoing operation. Plan \$50-\$150/month for optional paid cloud AI and off-site backup, plus electricity/internet. API-heavy unattended work can cost more.

Free model weights can remove per-token model charges. They do not remove hardware, power, administration or paid service costs. Review the exact model license before business use.

PROCUREMENT WORKSHEET

Build the purchase order

Copy these requirements into a vendor request. Amounts are budget allowances unless a dated listing is explicitly identified; no cart or binding quote has been prepared.

QTY	ITEM / CONFIGURATION	LINE TOTAL
1	Mac mini, configured 1 unit; M5 Pro 18/20 / 64GB / 1TB. Rounded configurator arithmetic.	\$3,200
1	Dell U3425WE 1 monitor with supplied stand and Thunderbolt hub.	\$1,000
1 set	Logitech Wave Keys + Lift Combo for Mac 1 ergonomic keyboard/mouse set; observed \$119.99; choose for hand fit. Line total is a planning allowance.	\$150
1 each	Logitech Brio 500 + Logitech H390 USB-C headset 1 1080p USB-C webcam with privacy shutter; observed \$129.99. 1 headset; select USB-C variant 981-001575, not USB-A; observed \$24.99 regular. Line total is a planning allowance.	\$180
1	WD My Passport 4TB HDD 1 dedicated backup HDD; \$254.99-\$298.99 observed across page/index; use \$300 allowance. Includes USB-C adapter; format for Time Machine. Line total is a planning allowance.	\$300
1	CyberPower CP1000PFCLCD 1 pure sine wave UPS; observed MSRP \$199.95. Size shutdown runtime against measured computer, display and router load. Line total is a planning allowance.	\$210
1 each	Monoprice Cat6 FLEXboot 7ft gray + Anker USB-C to USB-A adapter 1 Ethernet cable; \$10 planning allowance, not verified live price; choose length for desk location. 1 pack; observed regular \$12.99; adapters for USB-A peripherals. Use supplied monitor and drive cables first. Line total is a planning allowance.	\$40
EQUIPMENT TOTAL		\$5,080

Equipment worksheet totals \$5,080 before tax. Keep the 64GB memory. Choosing the S3425DW monitor saves about \$580: equipment target about \$4,500. Confirm business software compatibility with macOS.

Before placing the order

Confirm required business software works on macOS; keep the specified memory. Check monitor refresh/cable compatibility, backup capacity and UPS watts. For a local-model pilot, record the exact model/quantization, working context, time to first token and a representative document task. Excludes tax, shipping, furniture, warranty/support upgrades, paid software and setup labor unless explicitly included above. A single system provides no automatic failover.

ORDER LINKS & SOURCES

Find the specified equipment

Links identify the recommended products. Check the selected variant, current price and delivery estimate before ordering. Group allowances on the previous page remain the purchasing budget.

Where to buy or request a quote

[Configure at Apple](#)

Verify 18-core CPU / 20-core GPU, 64GB, 1TB in final cart.

[Monitor at Dell](#)

U3425WE; this is 1440 pixels tall, not a 4K display.

[Logitech Wave Keys + Lift Combo for Mac · 920-013628](#)

1 ergonomic keyboard/mouse set; observed \$119.99; choose for hand fit.

[Logitech Brio 500 · 960-001493](#)

1 1080p USB-C webcam with privacy shutter; observed \$129.99.

[Logitech H390 USB-C headset · 981-001575](#)

1 headset; select USB-C variant 981-001575, not USB-A; observed \$24.99 regular.

[WD My Passport 4TB HDD · WDBRMD0040BGY-WESN](#)

1 dedicated backup HDD; \$254.99-\$298.99 observed across page/index; use \$300 allowance. Includes USB-C adapter; format for Time Machine.

[CyberPower CP1000PFCLCD · 1000VA / 600W](#)

1 pure sine wave UPS; observed MSRP \$199.95. Size shutdown runtime against measured computer, display and router load.

[Monoprice Cat6 FLEXboot 7ft gray · 9798](#)

1 Ethernet cable; \$10 planning allowance, not verified live price; choose length for desk location.

[Anker USB-C to USB-A adapter · 2-pack](#)

1 pack; observed regular \$12.99; adapters for USB-A peripherals. Use supplied monitor and drive cables first.

Source references

Apple configurations	Display specification
Representative local model sizes	70B sizing
Context memory	

Ordinary merchant links only. Affiliate tracking is not activated. Prices checked October 16, 2026.

03 / MORE ROOM FOR BIGGER MODELS

Local AI studio

A premium 40-inch workspace with the memory to explore 70B models and keep a substantial local document workflow.

\$8,600

Equipment planning budget

70B quantized models

One power user; small-team pilot possible



01 Specify the machine

Processor	Apple M5 Max: 18-core CPU, 40-core GPU
Memory	128GB unified memory
Storage	1TB SSD; model archive on separate external storage if needed
Display	Dell U4025QW: 5120 x 2160, up to 120Hz, IPS Black
Connection	Thunderbolt display; wired Ethernet for shared access
Software	macOS; MLX-capable runtime or Ollama; authenticated browser frontend for any sharing

Where it earns its place

- + Larger local language models and private knowledge search
- + Complex document work on a sharp 40-inch display
- + Creative production alongside AI experimentation
- + Lower-cost shared-office pilot when queued requests are acceptable

Network role

Can be a lower-cost shared hub for light usage. All users share its compute and one failure domain; it is not a high-availability server. Model examples describe memory capacity, not measured response speed. All users share the same compute.

CAPABILITY & DEPLOYMENT

What this build can do

Choose a model and workload first. Memory fit, useful speed and simultaneous demand are different purchasing questions.

Representative model targets

MODEL	FIT / CONDITIONS
Qwen3 32B	Comfortable capacity Speed still depends on runtime and input length.
Llama 3.3 70B	Good capacity target 43GB default quantized weights plus context; pilot latency.
Qwen3 235B / full DeepSeek R1	Does not fit 142GB / 404GB default weight files before runtime overhead.

Can support

- + Run cloud AI agents and ordinary business applications.
- + Run compatible downloadable models with no per-token inference bill.
- + Search approved local documents when a retrieval workflow has been configured.
- + Accommodate a 70B model at common 4-bit quantization with useful memory headroom.
- + Host a small-team pilot with a queue and approved document collections.

Cannot do

- + Automatically become a reliable business agent just by installing a model.
- + Replace tested workflows, backups, human review or account permissions.
- + Make paid cloud services free, or guarantee the quality of every answer.
- + Run all frontier-scale open models: Qwen3 235B at 142GB already exceeds this memory.
- + Guarantee many simultaneous long-context 70B requests.
- + Provide NVIDIA CUDA compatibility; check specialized software first.

Concurrent work: pilot before scaling

Pilot 1-3 modest requests; no guarantee

Separate setup and operating budgets

Setup / integration. Small-team sharing needs authentication, retrieval permissions, a request queue and testing; quote separately.

Ongoing operation. Local inference avoids model API charges. Shared services add administration, backup and support. Paid cloud tools remain optional and separately billed.

Free model weights can remove per-token model charges. They do not remove hardware, power, administration or paid service costs. Review the exact model license before business use.

PROCUREMENT WORKSHEET

Build the purchase order

Copy these requirements into a vendor request. Amounts are budget allowances unless a dated listing is explicitly identified; no cart or binding quote has been prepared.

QTY	ITEM / CONFIGURATION	LINE TOTAL
1	Mac Studio, configured 1 unit; M5 Max 18/40 / 128GB / 1TB. Rounded configurator arithmetic.	\$5,400
1	Dell U4025QW 1 monitor with stand, speakers and hub; verify desk depth.	\$2,250
1 set	Logitech Wave Keys + Lift Combo for Mac 1 ergonomic keyboard/mouse set; observed \$119.99; choose for hand fit. Line total is a planning allowance.	\$150
1 each	Logitech Brio 500 + Logitech H390 USB-C headset 1 1080p USB-C webcam with privacy shutter; observed \$129.99. 1 headset; select USB-C variant 981-001575, not USB-A; observed \$24.99 regular. Line total is a planning allowance.	\$180
1	WD My Passport 6TB HDD 1 dedicated backup HDD; \$334.99-\$391.99 across page/capacity selector; use \$400 allowance. Includes USB-C adapter; format for Time Machine. Line total is a planning allowance.	\$400
1	CyberPower CP1000PFCLCD 1 pure sine wave UPS; observed MSRP \$199.95. Size shutdown runtime against measured computer, display and router load. Line total is a planning allowance.	\$210
1 each	Monoprice Cat6 FLEXboot 7ft gray + Anker USB-C to USB-A adapter 1 Ethernet cable; \$10 planning allowance, not verified live price; choose length for desk location. 1 pack; observed regular \$12.99; adapters for USB-A peripherals. Use supplied monitor and drive cables first. Line total is a planning allowance.	\$40
EQUIPMENT TOTAL		\$8,630

This is the largest personal tier here, not a promise to run every open model. Buy memory for your model and context before upgrading internal storage.

Before placing the order

Confirm required business software works on macOS; keep the specified memory. Check monitor refresh/cable compatibility, backup capacity and UPS watts. For a local-model pilot, record the exact model/quantization, working context, time to first token and a representative document task. Excludes tax, shipping, furniture, warranty/support upgrades, paid software and setup labor unless explicitly included above. A single system provides no automatic failover.

ORDER LINKS & SOURCES

Find the specified equipment

Links identify the recommended products. Check the selected variant, current price and delivery estimate before ordering. Group allowances on the previous page remain the purchasing budget.

Where to buy or request a quote

[Configure at Apple](#)

Choose M5 Max 18-core CPU / 40-core GPU, 128GB, 1TB.

[Monitor at Dell](#)

U4025QW: confirm the final connection supports the desired refresh rate.

[Logitech Wave Keys + Lift Combo for Mac · 920-013628](#)

1 ergonomic keyboard/mouse set; observed \$119.99; choose for hand fit.

[Logitech Brio 500 · 960-001493](#)

1 1080p USB-C webcam with privacy shutter; observed \$129.99.

[Logitech H390 USB-C headset · 981-001575](#)

1 headset; select USB-C variant 981-001575, not USB-A; observed \$24.99 regular.

[WD My Passport 6TB HDD · WDBY3J0060BGY-WESN](#)

1 dedicated backup HDD; \$334.99-\$391.99 across page/capacity selector; use \$400 allowance. Includes USB-C adapter; format for Time Machine.

[CyberPower CP1000PFCLCD · 1000VA / 600W](#)

1 pure sine wave UPS; observed MSRP \$199.95. Size shutdown runtime against measured computer, display and router load.

[Monoprice Cat6 FLEXboot 7ft gray · 9798](#)

1 Ethernet cable; \$10 planning allowance, not verified live price; choose length for desk location.

[Anker USB-C to USB-A adapter · 2-pack](#)

1 pack; observed regular \$12.99; adapters for USB-A peripherals. Use supplied monitor and drive cables first.

Source references

Apple configurations	Display specification
Representative local model sizes	70B sizing
Context memory	

Ordinary merchant links only. Affiliate tracking is not activated. Prices checked October 16, 2026.

04 / A TEAM RESOURCE

Shared business hub

A dedicated GPU workstation that employees reach through a secure browser portal, with shared knowledge and controlled access.

\$28K-\$35K

Equipment planning budget

30B-32B default; 70B with testing

A small or medium business team



01 Specify the machine

Platform	Thinkmate GPX-WS 740RTP or integrator-validated equivalent
CPU	AMD Threadripper 9960X, 24 cores
GPU	1 x NVIDIA RTX PRO 6000 Blackwell Max-Q, 96GB GDDR7 ECC, 300W GPU
RAM	128GB DDR5 ECC RDIMM (4 x 32GB)
Storage	1TB NVMe boot plus 4TB NVMe model/data SSD
Network	10GbE wired server link; ordinary employee laptops use a browser
OS	Vendor-supported Ubuntu LTS with validated Blackwell driver/runtime
Power	Integrator-validated chassis/cooling/PSU; platform lists 1650W PSU, which is capacity, not continuous consumption
Software	vLLM or SGLang model service behind authenticated HTTPS application; document retrieval with permission filters; backups and monitoring
Admin display	Dell U3425WE 34-inch admin monitor

Where it earns its place

- + Permission-aware internal knowledge assistant
- + SOP, policy and project-document questions with source citations
- + Shared proposal drafting and research
- + Scheduled, approved internal workflows

Network role

Employees use existing computers via HTTPS. Authenticate each user, filter documents by permission before retrieval, keep the raw model endpoint private, and queue requests. A single hub has no automatic failover.

Model examples describe memory capacity, not measured response speed. All users share the same compute.

CAPABILITY & DEPLOYMENT

What this build can do

Choose a model and workload first. Memory fit, useful speed and simultaneous demand are different purchasing questions.

Representative model targets

MODEL	FIT / CONDITIONS
Qwen3 32B	Starting target Begin with 2-4 simultaneous requests and 8K-16K context; unmeasured pilot target.
Llama 3.3 70B	Conditional shared use About 43GB weights at common 4-bit; context and concurrency compete for remaining VRAM.
Full 671B / trillion-parameter models	Not a GPU-memory fit 96GB GPU VRAM is separate from the 128GB host RAM.

Can support

- + Serve multiple employee browser sessions on the private business network
- + Run a quantized 30B-32B model as a default shared assistant
- + Run a 70B model at about 4-bit precision with context and concurrency tested together
- + Search approved SOPs, manuals and project documents and cite retrieved sources
- + Draft summaries, proposals and internal answers; run approved, restricted tools

Cannot do

- + Run full 671B or trillion-parameter models entirely in 96GB GPU memory
- + Guarantee a fixed number of fast simultaneous users without a workload pilot
- + Treat 128GB system RAM as another 128GB of GPU VRAM
- + Provide automatic document permissions or failover merely by owning the hardware

Concurrent work: pilot before scaling

Start with 2-4 simultaneous requests, 8K-16K working context and a 30B-32B quantized model; this is a test target, not measured capacity. Test 70B separately.

Separate setup and operating budgets

Setup / integration. Allow \$3,000-\$10,000 for deployment/integration as an editorial planning estimate, separate from equipment. Scope and quote before purchase.

Ongoing operation. Budget staff or managed support separately; electricity example at 0.5kW average, 730 hours and \$0.16/kWh is \$58.40/month before cooling, backup service or labor.

Free model weights can remove per-token model charges. They do not remove hardware, power, administration or paid service costs. Review the exact model license before business use.

PROCUREMENT WORKSHEET

Build the purchase order

Copy these requirements into a vendor request. Amounts are budget allowances unless a dated listing is explicitly identified; no cart or binding quote has been prepared.

QTY	ITEM / CONFIGURATION	LINE TOTAL
1	Configured Linux AI workstation Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$25,000
1	34-inch Dell UltraSharp U3425WE admin monitor Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$1,000
1 set	Keyboard, mouse and admin peripherals Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$200
1 package	Backup NAS plus drives Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$2,000
1	UPS sized to measured load and shutdown needs Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$1,500
1 lot	10GbE switch, cabling and adapters Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$800
EQUIPMENT TOTAL		\$30,500

Puget's live configurator lists RTX PRO 6000 Blackwell Max-Q 96GB at +\$15,184.16. The complete \$25,000 server allowance is an estimate, not a saved vendor configuration. Puget's higher-spec 256GB/9975WX reference system lists \$37,116.87.

Before placing the order

Require a written acceptance test for the exact model/checkpoint, quantization, context and simultaneous requests. Verify user login, document access boundaries, restore from backup, queue behavior and support ownership. For enterprise hardware, obtain facilities sign-off for power, cooling, rack and delivery.

Excludes tax, shipping, furniture, warranty/support upgrades, paid software and setup labor unless explicitly included above. A single system provides no automatic failover.

Where to buy or request a quote

[Thinkmate GPX-WS 740RTP configuration / quote](#)
Request the exact configuration and acceptance benchmark; listing is not a saved order.

[Puget Systems workstation consultation / configuration](#)
Request the exact configuration and acceptance benchmark; listing is not a saved order.

Shared-system lines are requirements for a configured integrator quote. Ask the seller to confirm the final bill of materials; the CSV includes a quote URL for each line.

Source references

Thinkmate GPX-WS 740RTP configuration / quote	Puget Systems workstation consultation / configuration
Ollama: Qwen3 model files	Ollama: Llama 3.3 70B
Ollama: context length	NVIDIA RTX PRO 6000 Blackwell Max-Q

Ordinary merchant links only. Affiliate tracking is not activated. Prices checked October 16, 2026.

05 / FRONTIER-SCALE INFRASTRUCTURE

Enterprise AI hub

A professionally installed GPU rack for very large open-weight models and a company-wide internal AI service.

\$600K-\$750K

Equipment planning budget

Very large low-precision models; benchmark first

Enterprise; central IT and facilities required



01 Specify the machine

Platform	Supermicro SYS-822GS-NB3RT-01-G2, 8U air-cooled rack server
CPU	2 x Intel Xeon 6768P, 64 cores each
GPU	8 x NVIDIA B300 HGX with NVLink; vendor lists 2.3TB HBM3e total
RAM	2TB DDR5-6400 host memory
Storage	Listed server includes only 960GB boot NVMe; quote at least 2 x 7.68TB compatible enterprise NVMe data drives separately
Network	8 x 800GbE/XDR800 accelerator ports; separate application and management networks; no need for 800GbE employee desktops
Power	Professional facility design: approximately 15kW server-class load plus cooling; vendor/facility engineer must size circuits, UPS, feeds and rack
OS	Integrator-supported Linux with validated B300 serving stack
Software	vLLM/SGLang model serving, authenticated internal portal, access-aware retrieval, rate limits, job queue, monitoring and backup
Admin display	Separate admin desk with 34-inch ultrawide

Where it earns its place

- + Permission-aware internal knowledge assistant
- + SOP, policy and project-document questions with source citations
- + Shared proposal drafting and research
- + Scheduled, approved internal workflows

Network role

Employees use existing computers via HTTPS. Authenticate each user, filter documents by permission before retrieval, keep the raw model endpoint private, and queue requests. A single hub has no automatic failover.

Model examples describe memory capacity, not measured response speed. All users share the same compute.

CAPABILITY & DEPLOYMENT

What this build can do

Choose a model and workload first. Memory fit, useful speed and simultaneous demand are different purchasing questions.

Representative model targets

MODEL	FIT / CONDITIONS
DeepSeek-V4-Pro 1.6T	Vendor benchmark required MIT; published preview. Mixed FP4/FP8 support, memory and serving compatibility must be validated.
Kimi-K3 2.8T	Compatibility unverified Custom license. Pre-release recipe requires at least 8 x GB300 and suggests multi-node for production. This quoted B300 HGX configuration is not validated by that recipe; vendor must prove compatibility and throughput.
2.8T at BF16	Does not fit About 5.6TB of weights before runtime overhead; more hardware required.

Can support

- + Host very large open-weight models in supported low-precision formats
- + Pilot DeepSeek-V4-Pro 1.6T (MIT, published preview; mixed FP4/FP8) after vendor validation. Kimi-K3 2.8T is a capacity investigation only: its pre-release serving recipe specifies at least 8 x GB300, not this B300 HGX system; B300 compatibility is unverified.
- + Serve a company-wide application with centrally managed access and scheduled agent jobs
- + Run smaller models with more concurrency or separate model replicas instead of always choosing the largest model

Cannot do

- + Promise every largest model, full precision, maximum context and unlimited simultaneous users on one node
- + Run 2.8T parameters in BF16 entirely in this GPU memory (weights alone about 5.6TB)
- + Deliver a free-to-operate business service; energy, cooling, support and engineering remain material
- + Offer high availability from a single server; redundancy costs extra

Concurrent work: pilot before scaling

Require a vendor acceptance benchmark for the exact model/checkpoint, quantization, context and concurrent request target before ordering. Start pilot measurement at 1, 4, 8 and 16 simultaneous requests; do not publish these as guaranteed capacities.

Separate setup and operating budgets

Setup / integration. Allow \$50,000-\$150,000 for deployment/integration as an editorial planning estimate, separate from equipment. Scope and quote before purchase.

Ongoing operation. At an illustrative 15kW average and \$0.16/kWh, 730 hours costs \$1,752/month for server electricity alone. Cooling, facility/colocation, networking, support contracts and operations staff are additional. Average draw must be measured; this is not a quoted utility bill.

Free model weights can remove per-token model charges. They do not remove hardware, power, administration or paid service costs. Review the exact model license before business use.

PROCUREMENT WORKSHEET

Build the purchase order

Copy these requirements into a vendor request. Amounts are budget allowances unless a dated listing is explicitly identified; no cart or binding quote has been prepared.

QTY	ITEM / CONFIGURATION	LINE TOTAL
1 lot	Supermicro 8 x B300 server, rounded allowance Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$525,000
1 lot	Model NVMe storage, backup and shared storage allowance Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$15,000
1 lot	Business network integration, optics and cabling allowance Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$10,000
1 lot	Rack, PDU, UPS, electrical and cooling integration allowance Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$75,000
1	Separate admin workstation with 34-inch ultrawide monitor Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$2,000
1 lot	Quote contingency Planning allowance; include in integrator quote. Quantity 1 package unless specified in hardware requirements.	\$23,000
EQUIPMENT TOTAL		\$650,000

Earlier research observed \$524,626 for this exact Wiredzone SKU; a later check displayed "Email For Price." Treat it as quote-only. The \$525,000 server and \$650,000 overall allowances are editorial planning references, not current orderable prices or a quoted complete installation. NVIDIA DGX B300 is a different, separately quoted system.

Before placing the order

Require a written acceptance test for the exact model/checkpoint, quantization, context and simultaneous requests. Verify user login, document access boundaries, restore from backup, queue behavior and support ownership. For enterprise hardware, obtain facilities sign-off for power, cooling, rack and delivery.

Excludes tax, shipping, furniture, warranty/support upgrades, paid software and setup labor unless explicitly included above. A single system provides no automatic failover.

Where to buy or request a quote

[Wiredzone exact Supermicro server listing](#)

Request the exact configuration and acceptance benchmark; listing is not a saved order.

[Supermicro product / request quote](#)

Request the exact configuration and acceptance benchmark; listing is not a saved order.

[NVIDIA DGX B300 alternative / partner quote](#)

Request the exact configuration and acceptance benchmark; listing is not a saved order.

Shared-system lines are requirements for a configured integrator quote. Ask the seller to confirm the final bill of materials; the CSV includes a quote URL for each line.

Source references

Wiredzone exact Supermicro server listing	Supermicro product / request quote
NVIDIA DGX B300 alternative / partner quote	DeepSeek-V4-Pro model card
Kimi-K3 model card	Kimi-K3 serving recipe
Kimi-K3 license	

Ordinary merchant links only. Affiliate tracking is not activated. Prices checked October 16, 2026.