IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Local LLM Deployment on 24GB GPUs: Models & Optimizations

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2025-05-27 · large language models · local llms · gpu inference · llm deployment · vram optimization · quantization · gguf · inference frameworks · ai

Original article: https://intuitionlabs.ai/articles/local-llm-deployment-24gb-gpu-optimization

In this article

1. Running [Large Language Models (LLMs)](https://intuitionlabs.ai/articles/anthropic-claude-4-llm-evolution) Locally on a 24GB GPU (RTX 4090/3090)
2. Hardware Considerations: 24 GB VRAM and Model Size
3. Best LLM Models for 24GB GPUs (2025-2026)
4. Local Inference Frameworks (2025-2026)
5. RTX 5090: The New Consumer Champion (32GB)
6. Optimization Best Practices
7. Conclusion

[Revised October 10, 2026]

Running Large Language Models (LLMs) Locally on a 24GB GPU (RTX 4090/3090)

Running large language models (LLMs) on local hardware has become increasingly feasible, and many developers now run models on their own machines to keep data private and reduce cloud costs. This report explores the top LLMs that can be deployed on a single high-end GPU with 24 GB VRAM (NVIDIA RTX 4090, RTX 3090, or AMD RX 7900 XTX), focusing on open-source models with notes on a few closed models, and covers their architectures, VRAM requirements, speeds, context lengths, and use cases. We also discuss popular local inference frameworks (like **llama.cpp**, **vLLM**, **LM Studio**, **Ollama**) and optimization techniques (quantization, RoPE scaling, GGUF format) to maximize performance.

Note: The NVIDIA RTX 5090, released in January 2025, features 32GB VRAM (not 24GB), making it capable of running even larger models. This guide focuses on the 24GB VRAM class (RTX 4090/3090/7900 XTX) which remains the most common high-end consumer configuration.

Update (October 2026): The strongest open models for a single 24GB card are now newer than most of those covered below. Alibaba's Qwen3.8-27B is a dense 27B vision-language model under the Apache 2.0 license with a native context of 262,144 tokens (huggingface.co/Qwen), and Ollama lists its default Q4_K_M build at 18GB (ollama.com). Google's Gemma 4 family, also Apache 2.0, includes a 31B dense model and a 26B mixture-of-experts model with 3.8B active parameters, both with a 256K-token context (huggingface.co/google); their Q4_K_M builds are 20GB and 18GB (ollama.com). For coding, Mistral's 24B Devstral Small 2 is 15GB in Q4_K_M with a 384K-token context (ollama.com), and Qwen3.6-27B is 17GB (ollama.com). The Qwen3.6-35B-A3B mixture-of-experts model is 24GB in Q4_K_M, which leaves no room for context on a 24GB card. GPU prices have also risen sharply since 2025; current figures are in the hardware section below.

Hardware Considerations: 24 GB VRAM and Model Size

GPU VRAM and Model Parameters: The capacity of your GPU's VRAM primarily determines which models you can run. LLMs are often categorized by parameter count (e.g. 7B, 13B, 70B for 7 billion, 13 billion, 70 billion parameters). VRAM usage scales roughly linearly with model size and precision: for example, a 7B model in half-precision (FP16) may require ~14 GB VRAM, whereas a 13B model is ~26 GB (too large to fit 24 GB without compression). A 24GB GPU can handle *smaller models at full precision* or *larger models with quantization (compression)* apxml.com. The RTX 40/50 series GPUs also offer high memory bandwidth and tensor core performance, which improve throughput for lower precision inference.

NVIDIA's own product page confirms the GeForce RTX 4090 is built on the Ada Lovelace architecture with 16,384 CUDA cores, 24 GB of GDDR6X memory on a 384-bit bus, and a 450 W board power rating (nvidia.com). That combination of core count and memory bandwidth is what lets a single 24 GB card sustain useful throughput on 30B-class models once they are quantized down to roughly 4 bits per weight.

GPU Options for 24GB VRAM:

GPU	VRAM	Memory Bandwidth	Performance (8B model)	Price
NVIDIA RTX 4090	24GB GDDR6X	1,008 GB/s	~128 tok/s	$1,599 MSRP; ~$3,100 (Oct 2026)
NVIDIA RTX 3090	24GB GDDR6X	936 GB/s	~112 tok/s	~$1,370 used (Oct 2026)
AMD RX 7900 XTX	24GB GDDR6	960 GB/s	~100 tok/s	$999 MSRP; ~$1,000 (Oct 2026)

The RTX 4090 remains the most popular choice for serious local LLM deployment. For budget-conscious users, the RTX 3090 offers strong value in the used market, delivering about 87% of the RTX 4090's 8B throughput in the same benchmark (112 versus 128 tokens per second) at less than half the RTX 4090's October 2026 market price localllm.in.

GPU availability shifted considerably between this article's original publication and late 2026. Supply-chain reports in October 2024 said NVIDIA would stop producing the RTX 4090 ahead of the RTX 5090 launch (vast.ai), and by December 2024, with production halted, the remaining retail units were listed at roughly $2,500 to $3,700 (wccftech.com). Prices have kept rising through 2026. In October 2026, the average of the three lowest-priced eBay listings each day was about $3,093 for an RTX 4090 (gpupoet.com), $1,369 for an RTX 3090 (gpupoet.com) and $1,002 for an RX 7900 XTX (gpupoet.com). The RTX 3090's used market has held up better by comparison: tracked listings in July 2026 averaged around $1,254, with a typical range of $995 to $1,504, which keeps its cost per gigabyte of VRAM well below the RTX 4090's (bestvaluegpu.com). Reviewers continue to describe a used RTX 3090 as the value king for local AI workloads even after two subsequent GeForce generations, precisely because 24 GB of VRAM remains the practical threshold for running 30B-class quantized models comfortably (xda-developers.com). For buyers who do not need a display-class gaming card, NVIDIA's RTX 4000 SFF Ada offers 20 GB of ECC GDDR6 memory in a 70 W, half-height workstation form factor that draws power directly from the PCIe slot with no external power connector, trading some VRAM headroom and CUDA core count for a much smaller power and space footprint (nvidia.com).

Quantization: Quantization reduces memory usage by using lower precision for model weights (and sometimes activations). Common formats include 8-bit (INT8) and 4-bit (INT4) weight compression. For instance, 4-bit quantization cuts memory roughly to one-quarter: a 7B model that needs ~14 GB in FP16 might use only ~4–5 GB in 4-bit form apxml.com. Popular quantization methods are **GPTQ** (post-training quantization for GPU), **bitsandbytes** (8-bit loader), and the **GGUF** 4-bit quant formats used by llama.cpp apxml.com. These enable fitting larger models on 24 GB, with some quality trade-off.

GGUF Format (2025-2026 Standard): The GGUF format has become the de facto standard for local LLM deployment. It offers modular file structure, centralized metadata management, and excellent cross-platform support. llama.cpp now supports 1.5-bit through 8-bit integer quantization, with formats like Q4_K_M providing the best balance of quality and efficiency for most users github.com/ggml-org/llama.cpp. For aggressive quantization (IQ3, IQ2), using an importance matrix (`--imatrix`) significantly improves quality.

A newer quantization format worth tracking is EXL3, shipped by the ExLlamaV3 project as a streamlined derivative of the QTIP method developed at Cornell RelaxML. It computes Hessians on the fly and uses a fused Viterbi kernel so a model can be converted to a target bitrate in one pass, and a 3-bit-per-weight EXL3 quantization with a 4096-token cache can run in well under 16 GB of VRAM (github.com/turboderp-org/exllamav3). Independent benchmarking of llama.cpp's own quantization ladder on Llama-3.1-8B-Instruct supports the ordering most local users already assume in practice: Q4_K_M and Q5_K_M sit close enough to full precision on the accuracy/size curve that the loss is rarely noticeable in general chat or coding use, while formats below roughly 4 bits per weight show measurably larger degradation without an importance matrix (arxiv.org). The llama.cpp project's own quantize tool documentation remains the authoritative reference for what each suffix (_S, _M, _L) means and which base precisions (F32, BF16, FP16) can be converted into which GGUF quant (github.com/ggml-org/llama.cpp).

Model Size Guidelines for 24GB:

Model Size	VRAM (Q4 Quant)	VRAM (FP16)	Suitable for 24GB?
7-8B	~4-5 GB	~14 GB	✅ Excellent
13-14B	~8-9 GB	~26 GB	✅ Good
32-34B	~19-20 GB	~64 GB	✅ Tight fit
70B	~35 GB	~140 GB	⚠️ Requires offloading

Example: A 70B model in FP16 needs ~140 GB of memory, but in 4-bit (INT4) it's about 35 GB. This still exceeds 24 GB, but with aggressive quantization or layer offloading to system RAM, single-GPU operation becomes possible with performance tradeoffs.

Inference Speed: Speed is typically measured in tokens generated per second (tok/s). It depends on model size, quantization level, and the efficiency of the software. A larger model means more computation per token, so throughput drops as model size increases. On an RTX 4090-class GPU, a 7B model might generate on the order of ~100–140 tokens/s, whereas a 30B+ model might do ~30–40 tokens/s under similar conditions news.ycombinator.com. For example, using the optimized *exllama* GPU backend, users reported ~140 tok/s for a 7B model and ~40 tok/s for a 33B model on a 24 GB GPU.

RTX 4090 Benchmark Results (Ollama, Q4 quantization):

Model	Parameters	Eval Speed	GPU Utilization
LLaMA 3.1	8B	95.51 tok/s	92-96%
LLaMA 2	13B	70.90 tok/s	92-96%
Qwen 2.5	14B	63.92 tok/s	92-96%
DeepSeek-R1	32B	34.22 tok/s	92-96%

databasemart.com

This illustrates the **inverse scaling** of speed with size. The RTX 4090 excels in hosting lightweight and mid-range LLMs, with evaluation speeds consistently utilizing 92-96% of GPU capacity. Performance drops significantly for 40B+ models, where the 24GB VRAM becomes a limiting factor.

Best LLM Models for 24GB GPUs (2025-2026)

The local LLM landscape has evolved significantly. Here are the top models optimized for 24GB VRAM:

Llama 3.1/3.2 (Meta)

Meta's Llama 3.1 and 3.2 remain among the most popular open-source LLMs. Llama 3.1 is available in 8B, 70B, and 405B sizes, and Llama 3.2 added 1B and 3B text models and 11B and 90B vision models. The 8B variant is ideal for 24GB GPUs running at full precision, while the 70B can run with 4-bit quantization and layer offloading. Features include improved reasoning, multilingual support, and extended context windows up to 128K tokens meta.ai. Meta's own model card for the 8B instruct variant confirms pretraining on roughly 15 trillion tokens with a December 2023 data cutoff, and 39.3 million cumulative GPU hours of training compute on H100-class hardware (huggingface.co/meta-llama). For enterprise deployments evaluating larger open-weight releases beyond single-card limits, see the Llama 4 Hardware Requirements and VRAM Guide for multi-node cluster configurations and memory bandwidth benchmarks.

Mistral Small 3 (24B)

Released in early 2025, Mistral Small 3 represents the sweet spot for 24GB GPUs. It has a 32K-token context window (huggingface.co/mistralai) and fits comfortably with Q4 quantization (~14-15GB). Excellent for general-purpose tasks and coding mistral.ai.

Qwen 2.5 (Alibaba)

Qwen 2.5 models, particularly the Coder variants, have achieved impressive benchmarks on code generation, reasoning, and debugging. The 14B and 32B versions are well-suited for 24GB GPUs, with the 32B requiring Q4 quantization qwen.ai.

DeepSeek R1 Distilled Models

The full DeepSeek R1 (671B) requires datacenter hardware, but the distilled variants are excellent for local deployment:

Model	VRAM Required	Performance
DeepSeek-R1-Distill-Qwen-32B	~20GB (Q4)	Excellent reasoning
DeepSeek-R1-Distill-Qwen-14B	~9GB (Q4)	Fits 12GB+ GPUs
DeepSeek-R1-Distill-Qwen-7B	~4.7GB (Q4)	RTX 3070+ compatible

deepseek.com; Q4_K_M sizes from ollama.com

Gemma 2 (Google)

Google's Gemma 2 models (9B and 27B) offer competitive performance with efficient architectures designed for consumer hardware. The 27B model fits comfortably on 24GB with Q4 quantization.

Qwen3 (Alibaba)

Qwen3 succeeded the Qwen2.5 line with a range of dense and mixture-of-experts models, and Alibaba has since released Qwen3.5, Qwen3.6 and Qwen3.8 (see the October 2026 update above). The Qwen3-32B dense checkpoint carries 32.8B parameters, a native context window of 32,768 tokens (extendable to 131,072 with YaRN), and ships under the Apache 2.0 license, so it can be redistributed and fine-tuned without the restrictions some earlier open-weight releases carried (huggingface.co/Qwen). Ollama's library lists a Q4_K_M build of Qwen3-32B at roughly 20GB on disk, which leaves enough headroom on a 24GB card for a moderate context window once KV cache overhead is accounted for (ollama.com). Qwen3-32B (or the mixture-of-experts Qwen3-30B-A3B variant, which activates only about 3B parameters per token) remains a solid choice for a 24GB card, ahead of the older Qwen 2.5 generation this article originally covered, though the 27B dense models of the later Qwen3.6 and Qwen3.8 generations are smaller in Q4_K_M (17GB to 18GB) and have a native context of 262,144 tokens (ollama.com).

Gemma 3 (Google)

Google's Gemma 3 line succeeded Gemma 2 and added native multimodal (vision) input at the 4B, 12B, and 27B sizes. Google published official Quantization-Aware Trained (QAT) checkpoints rather than relying solely on third-party post-training quantization, and the company's developer blog reports that the int4 QAT build of Gemma 3 27B needs approximately 14.1GB of VRAM, comfortably inside a 24GB budget with room left for a large context window (developers.googleblog.com). The official Q4_0 unquantized-weights release on Hugging Face documents the same QAT approach and provides the checkpoint used by most downstream GGUF conversions (huggingface.co/google). Because QAT bakes the quantization error into training rather than applying it afterward, Gemma 3 27B tends to hold up better at 4-bit than post-training-quantized models of similar size.

Mistral Small 3.2 (Mistral AI)

Mistral AI iterated on its 24B "Small" line twice since the 3.0 release covered above: Mistral Small 3.1 added long-context and vision support, and Mistral Small 3.2 (released in June 2025) focused on reliability, notably cutting infinite-generation loops on long or repetitive prompts roughly in half relative to 3.1 (huggingface.co/mistralai). Both versions keep the 24B parameter count of the original Small 3 release and raise its context window from 32K to 128K tokens, so the VRAM math for the weights on a 24GB GPU is unchanged. Independent commentary on the release notes that Mistral recommends a low sampling temperature (around 0.15) and pairing the model with vLLM for the best function-calling and tool-use behavior (simonwillison.net).

gpt-oss-20b (OpenAI)

OpenAI's first open-weight release since GPT-2, gpt-oss, shipped in two sizes under an Apache 2.0 license. The smaller gpt-oss-20b is a 21B-parameter mixture-of-experts model with only about 3.6B active parameters per token, and OpenAI's own Hugging Face model card documents native support down to roughly 16GB of memory when quantized, well inside a 24GB budget (huggingface.co/openai). Community VRAM testing puts a Q4_K_M build of gpt-oss-20b at around 12GB, plus a few gigabytes of KV cache at typical context lengths, meaning a 24GB card can run it alongside a browser, IDE, or a second small model without contention (willitrunai.com). The larger 120B sibling (about 117B total, ~5B active) targets 80GB-class datacenter cards and does not fit a 24GB consumer GPU even quantized.

DeepSeek-V3.2 (DeepSeek)

DeepSeek has continued iterating past the R1 distillations covered above. DeepSeek-V3.2, released December 1, 2025 under the MIT license, introduces DeepSeek Sparse Attention (DSA), which reduces attention computation to near-linear complexity for long-context workloads and supports a 160,000-token context window at the full model size (huggingface.co/deepseek-ai). The full V3.2 checkpoint is a 685B-parameter mixture-of-experts model that targets multi-GPU datacenter deployment rather than a single 24GB card, and DeepSeek's release notes list only V3.2 and V3.2-Speciale, with no small distilled derivatives (api-docs.deepseek.com). The same release-notes index shows DeepSeek-V4 (preview April 24, 2026) and DeepSeek-V4.1-Flash (September 10, 2026) following it, and these are also datacenter-scale models. For a 24GB GPU today, the practical DeepSeek options remain the R1 distillations at the 32B, 14B, and 7B sizes rather than any full flagship checkpoint.

Local Inference Frameworks (2025-2026)

llama.cpp

Written in pure C/C++ with no external dependencies, llama.cpp remains a standout for efficiency and portability. Key 2025-2026 updates include:

- **NVFP4 and MXFP4 4-bit floating-point quantization types** alongside the established integer K-quants
- **Up to 35% faster token generation** on mixture-of-experts models on NVIDIA GPUs, from optimizations NVIDIA reported in January 2026 (developer.nvidia.com)
- **Vulkan support** for AMD and Intel GPUs
- **Feature-rich CLI and web UI** under 90MB total size

llama.cpp is ideal for extreme portability, minimal dependencies, and running on consumer-grade hardware github.com/ggml-org/llama.cpp.

vLLM

vLLM is engineered for high-performance, production-grade LLM inference. Key features include:

- **PagedAttention technology** reducing memory fragmentation by 50%+ and increasing throughput 2-4x for concurrent requests
- **vLLM v0.11.0** explicitly supports NVIDIA Blackwell architecture (RTX 5090) with native NVFP4/CUTLASS
- **vLLM-Omni** (November 2025) enables omni-modality model serving including diffusion transformers

For multi-user applications, vLLM delivered 35x+ the request throughput compared to llama.cpp at peak load vllm.ai.

Ollama

Ollama offers effortless installation, ready-to-use models, and a plug-and-play API. Perfect for rapid prototyping without complex setup. Supports mixed CPU/GPU inference for partial layer offloading ollama.ai.

LM Studio

LM Studio's GUI makes it easy for beginners and those who prefer graphical interfaces. Notable for:

- Good performance on lower-spec hardware with Vulkan offloading
- Easy function schema definition for agent workflows
- Interactive model testing and comparison

lmstudio.ai

Recommended Workflow: Use Ollama for rapid prototyping, then migrate to vLLM for production deployment where performance becomes critical medium.com/@rosgluk.

ExLlamaV3 (EXL3)

ExLlamaV3 is a purpose-built inference and quantization library for consumer GPUs, and its EXL3 format (described above) is now a third option alongside GGUF and GPTQ/AWQ for users who want the smallest possible footprint at a given quality target. The project ships as a single-step converter: give it a Hugging Face checkpoint and a target bits-per-weight, and it produces a quantized model without a separate calibration pass for most model sizes (github.com/turboderp-org). It remains more niche than llama.cpp or vLLM, but it is a reasonable option when a model needs to be squeezed to 3 or 3.5 bits per weight to fit a 24GB card alongside a large context window.

Version Currency Check (Mid-2026)

All three of the major frameworks referenced in this article have shipped substantial updates since it was first published. vLLM's GitHub release history shows v0.24.0 landing on June 28, 2026, with 408 commits from 200 contributors, continued maturation of DeepSeek-V3.2 support across backends, and Model Runner V2 now selected by default for Llama, Mistral, and Qwen3 dense models; by October 5, 2026 it had reached v0.31.0 (717 commits from 307 contributors), which added a `vllm preload` command that keeps quantized weights resident in GPU memory across engine restarts (github.com/vllm-project). Ollama reached v0.40.2 on October 8, 2026 (github.com/ollama). LM Studio's public changelog documents a steady cadence of point releases through 2026 covering MLX engine performance work for vision-capable models and Flash Attention now

enabled by default on its Vulkan and Metal backends (lmstudio.ai). A May 2026 roundup of local AI runtime updates covering Ollama, vLLM, llama.cpp, MLX, and LM Studio together is a useful single reference for tracking this pace of change without following five separate changelogs (codersera.com). The practical takeaway for a 24GB deployment is the same one this article made originally: pin a specific framework version for any production workload, and re-test after upgrading, since quantization kernels and default settings do change between releases.

RTX 5090: The New Consumer Champion (32GB)

For those considering an upgrade, the NVIDIA RTX 5090 (released January 2025) offers significant improvements:

Specification	RTX 5090	RTX 4090
VRAM	32GB GDDR7	24GB GDDR6X
Memory Bandwidth	1,792 GB/s	1,008 GB/s
CUDA Cores	21,760	16,384
Performance (8B)	~213 tok/s	~128 tok/s
Performance (32B)	~61 tok/s	~34 tok/s
Price (MSRP)	$1,999	$1,599

The RTX 5090 delivers about 67% more tokens per second than the RTX 4090 on an 8B model (213 versus 128), and at its $1,999 MSRP that works out to $9.38 per token/second localllm.in. Its 32GB VRAM gives 32B-class models more room for long contexts, but a 70B model in Q4_K_M is about 43GB, so it still needs partial offloading to system RAM or a 2- to 3-bit quantization (Q2_K is 26GB) (ollama.com). Market prices have also moved far above MSRP: in October 2026, the average of the three lowest-priced eBay listings each day was about $6,136 (gpupoet.com) runpod.io databasemart.com.

Optimization Best Practices

1. **Choose the right quantization:** Q4_K_M provides the best balance for most users. Use Q5_K_M for critical applications.
2. **Use importance matrices:** For aggressive quantization (IQ3, IQ2), always use `--imatrix` for better quality.
3. **Match model to use case:** Coding tasks benefit from specialized models like Qwen Coder or DeepSeek Coder. General chat works well with Llama or Mistral.
4. **Monitor VRAM usage:** Always reserve 20-30% additional VRAM for context windows and overhead.
5. **Consider hybrid inference:** Ollama and llama.cpp support offloading some layers to system RAM, enabling larger models at reduced speed.

Conclusion

Running LLMs locally on 24GB GPUs has never been more accessible. The RTX 4090 and RTX 3090 provide excellent performance for models up to 32B parameters with proper quantization. For those seeking more headroom, the RTX 5090's 32GB VRAM adds room for 32B-class models with long contexts, while 70B models still need low-bit quantization or partial offloading. With frameworks like llama.cpp, vLLM, and Ollama continuously improving (NVIDIA reported up to 35% faster token generation on mixture-of-experts models in llama.cpp in January 2026), local AI deployment remains a compelling alternative to cloud-based solutions for privacy, cost, and performance-conscious users.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://intuitionlabs.ai/articles/anthropic-claude-4-llm-evolution
2. https://intuitionlabs.ai/articles/meta-prompting-llm-self-optimization
3. https://intuitionlabs.ai/articles/conversational-ai-ui-comparison-2025
4. https://huggingface.co/Qwen/Qwen3.8-27B
5. https://ollama.com/library/qwen3.8
6. https://huggingface.co/google/gemma-4-31B-it
7. https://ollama.com/library/gemma4
8. https://ollama.com/library/devstral-small-2
9. https://ollama.com/library/qwen3.6
10. https://apxml.com/posts/best-local-llm-rtx-40-gpu
11. https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4090/
12. https://localllm.in/blog/best-gpus-llm-inference-2025
13. https://vast.ai/article/nvidia-rtx-4090-to-be-discounted-sooner-rather-than-later
14. https://wccftech.com/nvidia-geforce-rtx-4090-massive-price-hike-prior-to-rtx-5090-launch-over-2500-usd/
15. https://gpupoet.com/gpu/learn/price/october-2026/nvidia-geforce-rtx-4090
16. https://gpupoet.com/gpu/learn/price/october-2026/nvidia-geforce-rtx-3090
17. https://gpupoet.com/gpu/learn/price/october-2026/amd-radeon-rx-7900-xtx
18. https://bestvaluegpu.com/history/new-and-used-rtx-3090-price-history-and-specs/
19. https://www.xda-developers.com/used-rtx-3090-value-king-local-ai/
20. https://www.nvidia.com/en-us/products/workstations/rtx-4000-sff/
21. https://github.com/ggml-org/llama.cpp
22. https://github.com/turboderp-org/exllamav3
23. https://arxiv.org/html/2601.14277v1
24. https://github.com/ggml-org/llama.cpp/blob/master/tools/quantize/README.md
25. https://news.ycombinator.com/item?id=36628784
26. https://www.databasemart.com/blog/ollama-gpu-benchmark-rtx4090

27. https://ai.meta.com/llama/
28. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
29. https://gpusmith.com/articles/en/llama-4-hardware-requirements
30. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501
31. https://mistral.ai/
32. https://qwen.ai/
33. https://www.deepseek.com/
34. https://ollama.com/library/deepseek-r1
35. https://huggingface.co/Qwen/Qwen3-32B
36. https://ollama.com/library/qwen3:32b
37. https://developers.googleblog.com/en/gemma-3-quantized-aware-trained-state-of-the-art-ai-to-consumer-gpus/
38. https://huggingface.co/google/gemma-3-27b-it-qat-q4_0-unquantized
39. https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506
40. https://simonwillison.net/2025/Jun/20/mistral-small-32/
41. https://huggingface.co/openai/gpt-oss-20b
42. https://willitrunai.com/blog/gpt-oss-20b-vram-requirements
43. https://huggingface.co/deepseek-ai/DeepSeek-V3.2
44. https://api-docs.deepseek.com/news/news251201/
45. https://developer.nvidia.com/blog/open-source-ai-tool-upgrades-speed-up-llm-and-diffusion-models-on-nvidia-rtx-pcs/
46. https://vllm.ai/
47. https://ollama.ai/
48. https://lmstudio.ai/
49. https://medium.com/@rosgluk/local-llm-hosting-complete-2025-guide-ollama-vllm-localai-jan-lm-studio-more-f98136ce7e4a
50. https://github.com/vllm-project/vllm/releases
51. https://github.com/ollama/ollama/releases
52. https://lmstudio.ai/changelog
53. https://codersera.com/blog/local-ai-runtimes-may-2026-update/
54. https://ollama.com/library/llama3.3
55. https://gpupoet.com/gpu/learn/price/october-2026/nvidia-geforce-rtx-5090
56. https://www.runpod.io/blog/rtx-5090-llm-benchmarks
57. https://www.databasemart.com/blog/ollama-gpu-benchmark-rtx5090

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.