Optimizing Local Handwriting Recognition: A Guide to Ollama Models on the NVIDIA RTX 4090

1. Introduction to Handwriting Recognition with Local LLMs on Consumer GPUs

The accurate digitization of handwritten text has long presented a complex challenge. Handwritten Text Recognition (HTR) is inherently more intricate than Optical Character Recognition (OCR) for printed materials due to the immense variability in individual handwriting styles, the presence of ligatures, and inconsistencies even within a single person's script.[1] However, the advent of Large Language Models (LLMs), particularly multimodal vision-language models (VLMs), is ushering in a new era for HTR. These advanced models can transcend simple character mapping by leveraging contextual understanding and sophisticated visual feature extraction, leading to more robust and accurate recognition of handwritten content.[2]

This technological advancement is being further democratized by platforms like Ollama and the increasing power of consumer-grade Graphics Processing Units (GPUs). Ollama simplifies the local deployment and management of LLMs, making these powerful tools accessible without necessitating cloud-based services or specialized infrastructure.[4] This local approach is particularly beneficial for users requiring data privacy, offline operational capabilities, or seeking to avoid recurring API costs. Concurrently, high-end consumer GPUs, such as the NVIDIA RTX 4090 with its substantial 24GB of VRAM, are now capable of running these sophisticated LLMs, especially when techniques like model quantization are employed.[6] This convergence brings potent HTR capabilities within the reach of individual researchers, developers, and enthusiasts. The focus of this analysis is on "normal handwriting," implying contemporary, reasonably legible scripts, rather than highly specialized historical document analysis, although some models discussed may possess broader capabilities.

Ollama's recently introduced multimodal engine is a cornerstone of this development, designed to enhance the reliability, accuracy, and support for various data modalities, including vision. This engine is foundational for effective image-based tasks such as OCR and HTR.[3] Key improvements, such as model modularity—which isolates model-specific logic—and enhanced accuracy through more effective handling of image embeddings and batch processing, are crucial for achieving consistent and high-quality HTR performance.[3]

The confluence of powerful open-source multimodal LLMs [3], user-friendly deployment tools like Ollama [5], and capable consumer hardware like the RTX 4090 [6] has cultivated a new ecosystem for tackling advanced AI tasks such as HTR. Historically, high-performance HTR was often confined to proprietary systems or required significant research infrastructure. Open-source VLMs provide the sophisticated analytical capabilities, Ollama offers the framework to make these capabilities locally accessible and manageable, and powerful GPUs like the RTX 4090 supply the necessary computational power. This combination means that tasks previously beyond the reach of many individuals are now feasible, fostering innovation and experimentation in diverse areas, including personal document digitization, accessibility tools, and custom HTR applications. This shift effectively democratizes access to advanced HTR technology, moving it from specialized laboratories to individual desktops and potentially unlocking a new wave of applications.

2. Key Considerations for Selecting Ollama Models for Handwriting OCR on an RTX 4090

Choosing an appropriate Ollama model for handwriting recognition on an NVIDIA RTX 4090 involves navigating several critical factors, primarily dictated by the GPU's capabilities and the specific demands of HTR.

VRAM Constraint (24GB on RTX 4090): The Dominant Factor

The NVIDIA RTX 4090 is equipped with 24GB of GDDR6X VRAM.[6] While this is a substantial amount for a consumer-grade GPU, it remains a finite resource that imposes a hard limit on the size of models that can be run effectively. The number of parameters in an LLM directly correlates with its VRAM requirements. General estimations suggest that small models (2-10 billion parameters) typically require 6-16GB of VRAM, medium-sized models (10-20 billion parameters) need 16-24GB, and large models (20-70 billion parameters) demand 24-48GB.[6] Consequently, models in the 10-30 billion parameter range are prime candidates for an RTX 4090, particularly when model quantization techniques are applied. For instance, an RTX 4090 is recommended for models such as DeepSeek R1 32B (requiring approximately 20GB), Qwen 1.5 32B (18GB), Gemma 2 27B (16GB), LLaVA 34B (20GB), and Code Llama 34B (19GB).[6] This provides a concrete baseline for parameter sizes generally compatible with the hardware.

The Critical Role of Vision Capabilities

Handwriting recognition is fundamentally an image-to-text task. Therefore, the selected models must possess robust vision encoders and effective multimodal fusion capabilities to interpret visual information and translate it into textual data.[2] Ollama's model library clearly tags models with "vision" capabilities, and these are the primary candidates for HTR.[4] Furthermore, Ollama's new multimodal engine is engineered to improve the handling of vision models. This includes enhancements in image caching and the accurate processing of image embeddings, both of which are vital for consistent and high-quality OCR/HTR output.[3]

Model Size, Quantization, and the Performance-VRAM Trade-off

Larger parameter models generally offer the potential for higher accuracy and more nuanced understanding due to their increased capacity. However, they also demand more VRAM. This is where **quantization** becomes a critical technique for running larger models on hardware with limited VRAM like the RTX 4090. Quantization reduces the precision of the model's weights (e.g., from 16-bit floating point (fp16) to 4-bit or 8-bit integers), thereby significantly shrinking the model's size and its VRAM footprint.[10]

For example, a Qwen 2.5 32B model might require around 20GB at full precision.[6] However, a Q4_K_L quantized version of this model could fit into considerably less VRAM, making it more manageable on a 24GB card while leaving overhead for context processing and other system needs. User reports indicate a Q4_K_L quantized Qwen 2.5 32B model utilized approximately 23.72GB of VRAM when loaded.[11] Similarly, the Mistral Small 3.1 model, with 24 billion parameters, is explicitly stated to be runnable on a single RTX 4090.[12] Quantized GGUF variants of Mistral Small 3.1 are optimized for GPUs with as little as 16GB VRAM, with a Q4_K_L version being approximately 13GB in size.[13] The Gemma 3 27B Quantization Aware Training (QAT) variant is also noted to run on a 24GB VRAM GPU, although achieving optimal performance with very large context windows might necessitate careful memory management.[15] It is important to note that some sources might provide VRAM estimates for unquantized models or for multi-GPU setups, which can differ significantly from single RTX 4090 scenarios with quantization.[16]

The impact of quantization on accuracy is a key consideration. While reducing VRAM usage, quantization can potentially degrade model performance. However, modern quantization techniques such as GPTQ (Generative Pre-trained Transformer Quantization), AWQ (Activation-aware Weight Quantization), GGUF (Georgi Gerganov Universal Format), and QAT (Quantization Aware Training), coupled with careful hyperparameter tuning, aim to minimize this accuracy loss.[10] A study by RedHat focusing on Llama 3.1 demonstrated that well-quantized models could recover over 99% of the average score achieved by their unquantized baselines on many general language benchmarks.[10] However, for more specialized and nuanced tasks like OCR of historical documents, another study found that a 4-bit quantized Llama-3.1-70B model (q4) performed slightly worse in Character Error Rate (CER) improvement (38.7%) compared to its 16-bit floating point (fp16) counterpart (42.6% CER improvement), although both still offered substantial improvement over the baseline. The VRAM saving was significant (43.6GB for q4 versus 132.1GB for fp16, the latter clearly not for a single consumer GPU).[18] This suggests a potential trade-off, particularly for tasks requiring fine-grained distinctions.

The necessity of quantization for running larger, potentially more accurate, vision models on the 24GB VRAM of an RTX 4090 is clear.[6] While general benchmarks indicate minimal accuracy loss with sophisticated quantization methods [10], HTR is a task that demands high precision, as small errors in character recognition can significantly alter meaning. The observed slight performance dip in OCR error correction with quantization for a large model [18], and findings that LLMs can perform weaker on non-English HTR [1], suggest that quantization might subtly affect a model's ability to discern fine character features, especially in less common scripts or varied handwriting styles. Therefore, while users should leverage quantization to run more capable models, they must be prepared to empirically test different quantized versions (e.g., Q4_K_M versus Q5_K_M, if available) and potentially accept a minor accuracy trade-off for the benefit of running a larger base model. For critical HTR applications, particularly those involving diverse handwriting styles or multiple languages, thorough testing on representative samples is paramount.

Evaluating HTR Performance: Beyond General Vision Benchmarks

While general vision benchmarks such as MMBench or Visual Question Answering (VQA) datasets provide an indication of a model's overall visual understanding, they are not specifically tailored to evaluate HTR performance. For a more accurate assessment, it is crucial to look for results on benchmarks like DocVQA (Document Visual Question Answering), OCRBench, or specific HTR datasets if available.[19] User experiences and anecdotal evidence also become highly valuable, especially for "normal" handwriting, which may not be comprehensively represented in all formal academic benchmarks.[23] The existence of community projects like the Ollama-OCR package [26], which curates and provides tools for using VLMs for OCR, is itself a testament to the community's efforts in this area. The models supported by such packages often represent good starting points for evaluation.

It's becoming apparent that some models, like Qwen 2.5 VL, demonstrate strong OCR performance even without being exclusively designed for it, in some cases outperforming models specifically trained for OCR.[20] Conversely, general-purpose VLMs like LLaVA might occasionally produce erroneous OCR output.[26] The fact that companies like Mistral have released dedicated "Mistral OCR" APIs underscores the specialized nature of high-quality OCR.[27] Early VLMs aimed for broad visual understanding, with OCR/HTR being one of many potential applications. However, the intricacies of recognizing text within images, especially varied handwriting, often necessitate specific architectural considerations or targeted fine-tuning. Models like Qwen 2.5 VL, which show strong performance on OCR-related benchmarks [20], have likely incorporated training data or architectural elements that inherently benefit text recognition. The development of dedicated tools like Ollama-OCR [26] and specialized APIs suggests that while general VLM capabilities offer a foundation, further specialization—whether through fine-tuning, specific model design, or curated application—often yields superior OCR/HTR results. Consequently, users should prioritize models with demonstrated OCR/HTR benchmark performance or those explicitly highlighted in OCR-focused tools and discussions over generic VLMs, even if the latter are larger or newer. The "best" general vision model is not invariably the "best" for HTR, indicating a trend towards models either being implicitly proficient at OCR due to their comprehensive training and architecture or being explicitly optimized for it.

3. Promising Ollama Vision Models for Handwriting Recognition (RTX 4090 Compatible)

Identifying the optimal Ollama model for handwriting recognition on an RTX 4090 requires a careful assessment of VRAM compatibility, vision capabilities, and specific evidence of OCR/HTR performance. The following table summarizes key characteristics of promising candidates, followed by a detailed analysis of each.

Table 1: Comparative Summary of Promising Ollama Vision Models for Handwriting OCR on RTX 4090

Model Name	**Base Parameter Size**	**Common Quantized Sizes (GB) for RTX 4090**	**Estimated VRAM (GB) on RTX 4090 (Typical Quant)**	**Key Strengths for Handwriting OCR**	**Known Limitations/ Concerns for HTR**	**Relevant Snippets**
Minicpm-V 2.6 / Minicpm-O 2.6 (Vision)	8B	fp16 (15-16GB), Q-formats (e.g., Q4 ~4-5GB)	~5-16GB	SOTA OCRBench performance, surpasses proprietary models; good user HTR reports; efficient; handles varied aspect ratios/high-res images; multilingual.	Omni (O) version's audio features not used by Ollama for vision; ensure vision part is equivalent to V.	[21]
Qwen 2.5 VL	32B (or 72B quant.)	Q4_K_L for 32B (~20-23GB)	~20-23GB (for 32B)	Excellent OCR benchmarks (matches GPT-4o); user confirmed good for handwriting; strong structured data/JSON output; multilingual; HTML bounding boxes.	72B variant likely too large even quantized for stable single RTX 4090 use if full context needed; 32B is safer.	[11]

Llama 3.2 Vision	11B	Base model size (~8-10GB)	~8-10GB	Good for document-centric OCR (charts, DocVQA); ~80% HTR accuracy (anecdotal); supported by Ollama-OCR; VRAM efficient.	English-only for combined image+text tasks.	[19]
Mistral Small 3.1	24B	Q4_K_L (~13GB)	~13-16GB	Broad multilingual OCR reported by users; ~85% HTR accuracy (anecdotal); large context window; runs on RTX 4090.	Lower accuracy on specific diacritics; resource-intensive vs traditional OCR; potential past Ollama performance issues needing workarounds.	[12]
Gemma 3	12B or 27B (quant.)	12B QAT (~8.9GB), 27B QAT (~18GB)	~9-18GB	Strong general multimodal benchmarks; large context; multilingual.	Poor OCR benchmark (JSON extraction); ~70% HTR (anecdotal, enhanced 12B). Less favored for primary HTR.	[15]
LLaVA (1.6 or similar)	13B or 34B (quant.)	13B (~8GB), 34B (~20GB)	~8-20GB	General visual understanding; some	Inconsistent OCR/HTR; "wrong output	[6]

				success with simple English HTR.	sometimes"; often defaults to image description; poor on non-English HTR.	

Detailed Model Analysis:

3.1 Qwen 2.5 VL (Visual Language) Series (e.g., 32B, or quantized 72B)

- **Overview**: The Qwen series, developed by Alibaba, includes powerful multimodal models, with the Qwen 2.5 VL versions specifically tailored for vision-language tasks. These models are available in a wide range of parameter sizes, from 0.5B up to 72B and even larger experimental versions.[4]
- **VRAM & RTX 4090 Compatibility**:
  - The Qwen 2.5 32B (non-VL coder variant) is listed as requiring approximately 20GB of VRAM, making it suitable for the RTX 4090.[6]
  - User reports confirm that a quantized Qwen 2.5 32B model (specifically, a Q4_K_L quant) can run on an RTX 4090, utilizing around 23.72GB of VRAM with a 64k context.[11] This suggests that a 32B VL model should also be feasible within the 24GB VRAM limit.
  - While the Qwen 2.5 VL 72B variant has demonstrated very strong OCR performance [20], its base model size (approximately 47GB [6]) is too large for a single RTX 4090. Aggressive quantization might be attempted, but the 32B variant offers a more reliable and stable option for single-GPU operation.
- **Handwriting/OCR Performance & Insights**:
  - **Strong OCR Benchmarks**: Qwen 2.5 VL models, including both 72B and 32B variants, have shown remarkable performance in OCR benchmarks. They achieved approximately 75% accuracy in JSON extraction tasks from documents, a level comparable to GPT-4o, and notably outperformed specialized models like mistral-ocr (72.2%) and Gemma-3 (42.9%) in these specific evaluations.[20] This is particularly impressive as these Qwen models were not exclusively designed for OCR.
  - **User Confirmation for Handwriting**: In a community discussion regarding Qwen 2.5 VL's capabilities, a user explicitly confirmed its effectiveness with handwritten documents.[23]
  - **Strengths**: Key advantages for complex document OCR include enhanced structured data processing, improved JSON output generation, a large context window (supporting up to 128K tokens), and multilingual support across 29

    languages.[20] The capability to provide bounding boxes in HTML output is a valuable feature for visual verification of OCR results.[28]
  - **Prompts for OCR**: For optimal results, specific prompts tailored to the document type are recommended (e.g., "Extract all invoice details..." for invoices).[20] Using lower temperature settings (e.g., 0.0-0.3) is also advised to enhance accuracy.[20]
  - **Ollama Availability**: Qwen models, including qwen2.5vl, are available in the Ollama library.[3]
  - The strong performance of Qwen 2.5 VL, a general multimodal model, in specialized OCR tasks, even surpassing some dedicated OCR models [20], is noteworthy. This suggests that its underlying architecture, extensive pretraining (the Qwen 2.5 series was pretrained on up to 18 trillion tokens [20]), or the specific nature of its visual understanding capabilities are highly conducive to text recognition. Its proficiency with structured data [20] likely contributes to this, as recognizing text often involves understanding its layout and structure within a document. This challenges the notion that only narrowly specialized models can excel at tasks like OCR and positions Qwen 2.5 VL as a versatile yet powerful candidate for HTR.

3.2 Llama 3.2 Vision (e.g., 11B variant)

- **Overview**: Meta's Llama 3.2 series introduces vision-enabled models, available in 11B and 90B parameter sizes, capable of processing both text and images.[4]
- **VRAM & RTX 4090 Compatibility**:
  - The Llama 3.2 Vision 11B model is reported to require approximately 8GB of VRAM [31], making it comfortably runnable on an RTX 4090 with ample VRAM to spare.
  - The 90B variant, requiring 64GB of VRAM [31], is too large for a single RTX 4090.
  - Ollama lists llama3.2-vision with both 11B and 90B size options.[4]
- **Handwriting/OCR Performance & Insights**:
  - **Optimized for Visual Tasks**: These models are designed for a range of visual recognition tasks, including image reasoning, captioning, and answering questions about images.[19]
  - **Benchmark Performance**: The Llama 3.2 Vision models demonstrate strong performance in chart and diagram understanding, achieving high scores on benchmarks like AI2 Diagram (92.3) and DocVQA (90.1), where they outperform models like Claude 3 Haiku.[19] This indicates a good capability with structured visual information that often contains embedded text.
  - **Ollama-OCR Support**: Llama 3.2 Vision is listed as a supported model within the Ollama-OCR package, where it is described as an "advanced model with high accuracy for complex documents".[26] This is a significant endorsement of its OCR capabilities from a community tool focused on this task.
  - **User Experience**: One user mentioned employing Llama 3.2 Vision for OCR tasks via the Ollama-OCR tool.[32] Another user, in a Reddit discussion about Ollama

    vision models, reported achieving approximately 80% accuracy for handwriting recognition with llama3.2-vision:11b.[25]
  - **Language Support**: The Llama 3.2 Vision models support multiple languages for text-only tasks. However, a critical limitation for HTR is that for combined image and text tasks, they currently only support English.[31]
  - Llama 3.2 Vision's proficiency in document-level understanding, particularly with charts and diagrams [19], and its characterization as suitable for "complex documents" by the Ollama-OCR project [26], suggest its strengths lie in recognizing text within structured or semi-structured layouts. While an anecdotal handwriting accuracy of ~80% [25] is respectable, it might not be state-of-the-art compared to models potentially more focused on diverse, unstructured handwriting. The English-only limitation for image-plus-text tasks is a significant drawback for users requiring multilingual HTR capabilities. Thus, Llama 3.2 Vision 11B is a strong candidate for English handwriting, especially if it appears within broader document contexts, but other models may be more appropriate for purely unstructured or multilingual handwriting.

3.3 Mistral Small 3.1 (24B)

- **Overview**: Mistral Small 3.1 is an open-source model from Mistral AI with 24 billion parameters, designed to support both text and image inputs.[4]
- **VRAM & RTX 4090 Compatibility**:
  - The model is explicitly stated to be capable of running on a single RTX 4090.[12]
  - Quantized GGUF variants, such as Q4_K_L, are optimized to run on GPUs with around 16GB of VRAM, with a corresponding file size of approximately 13GB.[13] This configuration leaves ample VRAM headroom on an RTX 4090.
  - It is important to note that some users have reported performance issues with earlier versions of Ollama (e.g., 0.6.5) when running this model, including slow inference speeds, high CPU utilization, and underutilization of VRAM. These issues were potentially attributed to VRAM estimation problems within Ollama or the size of the model's projector graph. Setting the num_gpu parameter to its maximum value or a high number of layers (e.g., 41 for a q3_k_s quant on a 16GB VRAM system) was suggested as a workaround.[24] Users should verify performance with current Ollama versions.
- **Handwriting/OCR Performance & Insights**:
  - **Multimodal Understanding**: Mistral Small 3.1 incorporates "state-of-the-art vision understanding" and supports an extensive 128k token context window.[4]
  - **Benchmark Performance**: The model has been shown to outperform comparable models like Gemma 3 and GPT-4o Mini in benchmarks such as GPQA and in multilingual tasks, particularly for European and East Asian languages.[12]
  - **User Experience for OCR**:
    - One user expressed satisfaction with its performance in OCRing text for smaller, less common languages.[24]

        - Another user provided a detailed comparison with a traditional 2-stage OCR process for Vietnamese documents. They noted that Mistral Small 3.1 was significantly easier to set up but was less accurate with accented Latin characters and was considerably slower and more resource-intensive. This user felt its recognition capabilities were comparable to built-in browser OCR engines.[24] This suggests its strength may lie in broad language applicability rather than deep accuracy for specific, complex scripts without further fine-tuning.
        - A user in an Ollama vision model discussion reported achieving approximately 85% accuracy for handwriting recognition with mistral-small3.1.[25]
        - Another user mentioned successfully using mistral-small3.1 as an "OCR assistant".[25]
    - **Limitations**: The detailed user comparison [24] indicates potential accuracy challenges with specific diacritics or accents and significantly higher resource consumption compared to traditional OCR methods for their particular use case. The previously reported performance issues on some setups [24] also warrant consideration.
    - Mistral Small 3.1 is often praised for its straightforward setup with Ollama and its ability to handle a variety of languages for OCR tasks.[24] However, it may struggle with fine-grained character details, such as accents, and can be resource-intensive compared to more specialized OCR tools.[24] The reported 85% handwriting accuracy [25] is good, but may not be top-tier if other models offer higher precision. The potential performance and VRAM utilization issues with certain Ollama versions should be verified. Consequently, Mistral Small 3.1 appears to be a suitable candidate for users who need a quick and easy setup for OCR/HTR across various languages (especially European and East Asian) where absolute precision on every character is not paramount, or for general "OCR assistance." For high-stakes, high-accuracy HTR on specific complex scripts, or if computational efficiency is a critical factor, it might be outperformed by other models.

3.4 Gemma 3 (e.g., 12B, or quantized 27B)

- **Overview**: Gemma 3 is Google's family of open models, built from the same research and technology as the Gemini models. Gemma 3 is multimodal, handling both text and image inputs to generate text outputs, and is available in 1B, 4B, 12B, and 27B parameter sizes.[4]
- **VRAM & RTX 4090 Compatibility**:
    - The older Gemma 2 27B model is listed as requiring 16GB of VRAM, making it suitable for an RTX 4090.[6]
    - For Gemma 3, a 12B image-to-text model was estimated by one source to need 31.2GB of VRAM, which would be too high for a single RTX 4090 without

    significant quantization.[16] However, the orieg/gemma3-tools:12b-it-qat variant available on Ollama is listed at 8.9GB [15], indicating that well-quantized and VRAM-friendly versions exist.
  - The Gemma 3 27B-it-qat (quantization-aware trained) variant is available on Ollama with a size of 18GB.[15] This version is stated to run on a 24GB VRAM GPU, although very large context windows might pose a challenge.
  - A user successfully ran a gemma-3-27b-it q4_k_m GGUF quantized model on a Mac Mini M4 with 24GB of shared RAM, with the model utilizing approximately 16GB of RAM.[37] This suggests that a similarly quantized 27B Gemma 3 model should be manageable on an RTX 4090.
- **Handwriting/OCR Performance & Insights**:
  - **Features**: Gemma 3 models boast a large 128K token context window (except for the 1B variant), support for over 140 languages, vision understanding capabilities via a tailored SigLIP vision encoder, and a Pan & Scan feature for handling high-resolution images.[36]
  - **Benchmark Performance**: The Gemma 3 27B IT variant is reported to match the performance of Gemini-1.5-Pro on many general multimodal benchmarks.[37] However, in a specific OCR benchmark focused on JSON extraction from documents, Gemma-3 (27B) scored only 42.9%. This was significantly lower than the performance of Qwen 2.5 VL and mistral-ocr in the same benchmark [20], which raises a major concern for its suitability for OCR tasks.
  - **User Experience**: One user reported approximately 70% handwriting accuracy when using ebdm/gemma3-enhanced:12b.[25]
  - Despite Gemma 3's impressive general multimodal capabilities and strong performance on various benchmarks [37], its notably poor performance on a key OCR-specific benchmark [20] is a significant drawback. Gemma 3 models are built from advanced Gemini research and possess features like large context windows and extensive multilingual support [36], which would suggest strong general text and image understanding. However, the low OCR benchmark score (42.9% on JSON extraction) indicates that these strengths do not automatically translate to high-fidelity text extraction from images. This discrepancy implies that the model's vision encoder (SigLIP) or its training data might not be optimally tuned for the fine-grained detail required for accurate OCR. Therefore, Gemma 3, despite its overall prowess and suitability for an RTX 4090 (especially its quantized 27B variant), appears to be a weaker candidate for primary OCR/HTR tasks compared to models like Qwen 2.5 VL or even Mistral Small 3.1, based on current OCR-specific benchmark data. The anecdotal ~70% handwriting accuracy [25] is also lower than that reported for other models.

3.5 LLaVA (Large Language and Vision Assistant) (e.g., 13B, or quantized 34B)

- **Overview**: LLaVA is a well-known open-source multimodal model that typically combines a vision encoder (often a CLIP ViT-L/14) with an LLM like Vicuna.[2] LLaVA 1.6 is an updated iteration of this model.

- **VRAM & RTX 4090 Compatibility**:
  - The LLaVA 13B model requires approximately 8.0GB of VRAM [6], making it easily accommodate an RTX 4090.
  - The LLaVA 34B model requires around 20GB of VRAM [6], which is also suitable for an RTX 4090. An RTX 4090 benchmark study included LLaVA 34B in its assessments.[7]
  - Ollama's library lists LLaVA with 7B, 13B, and 34B parameter options.[4]
- **Handwriting/OCR Performance & Insights**:
  - **General Vision**: LLaVA is generally proficient at visual and language understanding tasks, such as image captioning and visual question answering.[2]
  - **Ollama-OCR Support**: LLaVA is included in the list of models supported by the Ollama-OCR package. However, this support comes with a significant caveat: "LLaVa model can generate wrong output sometimes".[26] This warning raises concerns about its reliability for OCR tasks.
  - **Handwriting/OCR User Experience**:
    - An arXiv paper studying the application of LLMs to HTR mentions LLaVA in the context of Large Vision Language Models (LVLMs) inheriting typographic vulnerabilities from their vision encoders.[38] While not a direct performance review, its inclusion implies its use in such research. The paper generally concludes that LLMs show strong performance on modern English HTR but are weaker on other languages and lack self-correction capabilities.[1]
    - A YouTube video demonstrating LLaVA (likely version 1.5 or 1.6 via Ollama) showed it successfully recognizing the stylized handwritten phrase "Take it easy" and the planner text "Every day is a fresh start." However, it failed to transcribe handwritten Tamil text, although it correctly identified the script as Tamil.[39] This highlights potential language limitations for HTR.
    - A user in an Ollama vision model discussion stated that llava:7b offers "no OCR, but good image description".[25] This is a strong negative indicator for its OCR utility, at least for the 7B version or based on that user's experience.
    - Another user reported that when testing Ollama-OCR with LLaVA, the model tended to describe the image rather than performing OCR.[32]
  - LLaVA is a capable general-purpose VLM [2], but its performance in HTR and OCR appears inconsistent. The explicit warning about potentially incorrect output [26] is a significant concern for tasks requiring high accuracy. Its failure to transcribe Tamil handwriting [39] points to limitations in multilingual HTR, aligning with broader findings about the challenges LLMs face with non-English HTR.[1] User reports suggesting it is better suited for image description than OCR [25] indicate that it might not prioritize text extraction unless very specifically prompted, or that its

inherent OCR capabilities are weaker. While the 34B version might offer improvements over smaller variants, the fundamental architecture and training focus might still present these limitations. Therefore, LLaVA seems less promising for reliable and accurate transcription of normal handwriting, especially for multilingual content, compared to more OCR-focused models.

3.6 Minicpm-V / Minicpm-O Series (e.g., 2.6 8B variant)

- **Overview**: The MiniCPM-V and MiniCPM-O series are multimodal models developed with a focus on efficiency and strong visual capabilities, including OCR. MiniCPM-V 2.6, an 8B parameter model, is built using SigLip-400M and Qwen2-7B as foundational components.[40] The MiniCPM-O 2.6 is an 8B "omni" model, also based on Qwen2.5-7B, which extends capabilities to include audio and speech processing in addition to vision and text.[21]
- **VRAM & RTX 4090 Compatibility**:
  - These models are relatively small; for instance, older MiniCPM-V versions were around 2.8B parameters, while the 2.6 versions are 8B.[40] An 8B parameter model, especially when quantized, will run very comfortably on an RTX 4090, leaving substantial VRAM headroom.
  - MiniCPM-V 2.6 (8B fp16 version) is available on Ollama.[40] GGUF quantized versions are also available for MiniCPM models.[21]
- **Handwriting/OCR Performance & Insights**:
  - **Strong OCR Benchmarks**: MiniCPM-V 2.6 (8B) is reported to achieve state-of-the-art performance on OCRBench, surpassing several proprietary models including GPT-4o, GPT-4V, and Gemini 1.5 Pro in this specific benchmark.[40] Similarly, MiniCPM-O 2.6 (8B) claims state-of-the-art performance on OCRBench for models under 25B parameters, outperforming GPT-4o-202405.[21] These are very strong indicators of its proficiency in OCR tasks.
  - **Specific Features**: These models can process images with any aspect ratio and resolutions up to 1.8 million pixels (e.g., 1344x1344). They are designed for trustworthy behavior, exhibiting low hallucination rates, and support multilingual capabilities including English, Chinese, German, French, Italian, and Korean, among others.[40] A key efficiency feature is their superior token density, meaning they use fewer visual tokens to represent large images, which improves inference speed and reduces memory usage.[21]
  - **User Experience for OCR/Handwriting**:
    - A user testing various Ollama models for OCR tasks found that minicpm-v provided better results, as it attempted to format the output and distinguish between different parts of the document, unlike LLaVA or Llama models which sometimes merely described the image.[32]
    - Another user shared positive results from using MiniCPM-V 2.6 (run via KoboldCPP, though the base model is the same as what Ollama can run) on

        a handwritten note, providing both the input image and the successful OCR output.[44] This is direct evidence of good handwriting recognition performance.
        - A Reddit comment highlighted that the "previous only vision (minicpm 2.6) was a great model, current OMNI vision is even more powerful, and for many task like OCR/other vision tasks, it almost matches the bigger gpt4o".[41]
        - The Ollama-OCR package also lists Minicpm-v as one of its supported models.[26]
    - Minicpm models, particularly the 2.6 versions (both V and O, focusing on the vision component for HTR), offer a compelling combination of high efficiency (due to their smaller size and high token density) and top-tier OCR benchmark results, where they reportedly surpass much larger proprietary models.[21] Smaller models generally translate to faster inference times and lower VRAM consumption, which on an RTX 4090 means more resources are available for other concurrent tasks or for processing larger batches of images. The strong performance on OCRBench [21] is a direct testament to their text recognition capabilities. Positive user experiences with both general OCR structure [32] and specific handwriting samples [44] further corroborate these benchmark claims. Features such as handling varied image aspect ratios, high-resolution inputs, and multilingual support are all advantageous for diverse handwriting recognition scenarios. Consequently, MiniCPM-V 2.6 or MiniCPM-O 2.6 (if its vision capabilities are identical or superior for OCR and readily available via Ollama) emerges as a very strong candidate, offering an excellent balance of high accuracy for HTR and efficient resource utilization.

3.7 Other Potential Models (Briefly)

- **Moondream**: This model is described as a small vision language model designed for efficiency, particularly on edge devices.[26] While its efficiency is a plus, its HTR capabilities for "normal handwriting" compared to larger, more specialized models are less documented in the available information. Stated limitations include potential inaccuracies and difficulties with nuanced instructions.[45] A user noted that the latest April 2025 release of Moondream seemed promising for small model OCR but expressed concerns about its Ollama compatibility.[25]
- **Granite3.2-vision**: Characterized as a compact and efficient model specifically designed for visual document understanding, including elements like tables and charts.[26] One user on Reddit described it as a "beast" for OCR when used with Ollama version 0.5.13.[32] Its focus on document understanding is promising for recognizing text within structured contexts, which could include handwriting in forms or annotated documents.
- **BakLLaVA**: The provided information does not offer explicit details on BakLLaVA's HTR performance or its VRAM requirements on an RTX 4090. A Reddit thread discussing

Ollama-OCR mentioned LLaVA and Llama 3.2 Vision but did not single out BakLLaVA for HTR.[32] General search results for BakLLaVA tend to be more generic or relate to broader issues like confidence values in LLM-based OCR.[9]

- **DeepSeek-VL2-27B (with a 4B vision encoder component)**: This model is mentioned in MiniCPM-O's comparative benchmark table, where it shows strong scores on OCRBench (809) and other visual benchmarks.[42] If a 27B DeepSeek-VL model is available on Ollama and can run on an RTX 4090 (noting that the DeepSeek R1 32B text-only model requires 20GB [6]), it could be a viable option. However, direct evidence regarding its HTR performance is less prominent in the current set of materials.

4. Comparative Analysis and Performance Insights for Handwriting Recognition

A direct comparison of the discussed models reveals varying strengths and weaknesses for the specific task of handwriting recognition. The summary table provided in Section 3 offers an at-a-glance overview of these characteristics.

Direct Comparison for HTR/OCR Accuracy

Based on OCR benchmarks and user experiences, **Qwen 2.5 VL (32B, quantized)** and **Minicpm-V/O 2.6 (8B)** emerge as top-tier contenders. Both show exceptionally strong performance on OCR-specific benchmarks.[20] Qwen's ability to match GPT-4o in some OCR evaluations [20] and Minicpm's claims of surpassing it [40] are particularly noteworthy. Furthermore, positive user reports confirm their efficacy with handwritten text.[23]

Llama 3.2 Vision (11B) demonstrates promise for document-centric OCR, particularly for English-language content, and has an anecdotal HTR accuracy reported around 80%.[19] However, its limitation to English for combined image-text tasks restricts its applicability for multilingual handwriting.[31]

Mistral Small 3.1 (24B) offers broad language support and ease of use, with user-reported HTR accuracy around 85%.[24] However, it may face challenges with specific character details (like diacritics) and can be more resource-intensive compared to traditional OCR methods.[24]

Gemma 3 (quantized 27B or 12B), despite its strong general multimodal benchmark scores, performed poorly in a key OCR benchmark.[20] This makes it a less favored option for this specific application, although one user reported ~70% HTR accuracy with an enhanced 12B version.[25]

LLaVA (13B/34B) appears inconsistent for OCR/HTR tasks. Warnings about incorrect output [26] and user feedback suggesting it is better suited for image description than precise text extraction [25] diminish its standing for reliable HTR.

Synthesizing User Experiences and Anecdotal Evidence

Handwriting recognition is a nuanced task, and real-world user experiences provide invaluable insights. For instance, a Reddit user (randygeneric) offered a concise comparison of several models: mistral-small3.1 (~85% HTR accuracy), llama3.2-vision:11b (~80% HTR accuracy), ebdm/gemma3-enhanced:12b (~70% HTR accuracy), and llava:7b (no OCR capability noted).[25] This practical feedback complements formal benchmark data.

The successful application of Minicpm-V 2.6 to a user's handwritten note [44] serves as strong positive evidence for its capabilities. Similarly, another user's preference for minicpm-v for extracting structured text from files over llava or llama [32] further supports Minicpm's strengths in OCR-related tasks.

A detailed comparison of Mistral Small 3.1 against traditional OCR methods for Vietnamese text [24] highlighted an important trade-off: LLMs might offer easier setup for diverse document types but can lack the precision and speed of specialized tools for specific, well-supported languages and scripts. This underscores that the "best" model is highly dependent on the specific context of use—the nature of the handwriting, the language, the required accuracy, and available computational resources. No single model excels across all dimensions. For instance, if the primary need is English handwriting within structured documents, Llama 3.2 Vision might be suitable. For multilingual HTR where high accuracy is paramount, Qwen or Minicpm are strong contenders. If ease of setup for a less common language is prioritized, Mistral could be a starting point. This variability necessitates testing with representative samples to determine the optimal fit for a given user's "normal handwriting."

The Role of Prompting and Post-Processing

Effective HTR often involves more than just selecting a model; prompting strategies and post-processing techniques can significantly influence the quality of results. The Ollama-OCR package, for example, incorporates image preprocessing steps (like resizing and normalization) and allows for the use of custom prompts.[26]

For Qwen 2.5 VL, specific prompts tailored to the document type (e.g., invoices, forms, tables) and explicitly requesting structured JSON output are recommended for better OCR outcomes. Adjusting model parameters, such as lowering the temperature setting (e.g., to 0.0-0.3), is also advised to enhance accuracy.[20] User observations that minicpm-v "tries to format the output and distinguish between the different parts of the document" [32] suggest either good default behavior or responsiveness to formatting instructions within prompts. The LLMOCR script utilized with Minicpm-V in one successful handwriting test [44] likely incorporates specific prompting strategies to guide the model. Furthermore, research into OCR post-correction for historical documents has explored prompt-based methods, indicating that prompting is a key modality for interacting with and refining the output of these models.[18]

The broader ecosystem, including tools like Ollama-OCR [26] and active community discussions and benchmarks [23], plays a crucial role in navigating the rapidly evolving landscape of local VLMs for tasks like HTR. The sheer number of available models and their variants makes individual evaluation a time-consuming endeavor.[4] Projects like Ollama-OCR pre-select models known for some OCR utility and provide a framework for their application, acting as an initial filter. Community-driven benchmarks offer independent performance data, while user discussions on platforms like Reddit provide real-world experiences, highlight potential caveats (such as LLaVA's inconsistency [26]), and share practical tips (like performance workarounds for Mistral Small [34]). This collective intelligence helps users identify promising models and potential pitfalls more efficiently than isolated research might allow. Leveraging these community resources is advisable; the models supported by Ollama-OCR offer a good starting list, and monitoring ongoing discussions and emerging benchmarks can guide future model selection and highlight new state-of-the-art contenders or issues with existing ones.

5. Recommendations for Normal Handwriting Recognition on RTX 4090

Based on the analysis of VRAM compatibility, OCR/HTR benchmark performance, user experiences, and specific model features, the following recommendations are provided for selecting an Ollama model for recognizing normal handwriting on an NVIDIA RTX 4090.

Top Tier Recommendations (Balancing Accuracy, Feasibility, and Evidence):

1. **Minicpm-V 2.6 (8B) / Minicpm-O 2.6 (8B) (Vision Component)**
   - **Reasoning**: This model series stands out due to its strong OCR benchmark performance, with reports indicating it surpasses even larger proprietary models in this domain.[21] Positive user experiences, specifically with handwriting recognition [44] and general structured OCR output [32], further bolster its candidacy. Its efficiency, characterized by a small model size and high token density, ensures it runs comfortably on an RTX 4090, leaving ample VRAM for other processes or larger batch sizes. Additionally, it offers multilingual support and robust handling of various image properties.[40]
   - **Ollama Tag**: Look for minicpm-v:2.6-fp16 or other quantized versions if available. It's important to verify if the vision component of minicpm-o is separately available and performs equivalently or better for vision tasks in Ollama, or if minicpm-v is the primary choice for vision-centric applications.
2. **Qwen 2.5 VL (32B, quantized)**
   - **Reasoning**: This model demonstrates excellent OCR benchmark results, achieving performance comparable to GPT-4o in some evaluations.[20] User feedback confirms its effectiveness for handwriting recognition.[23] Its strengths in processing structured data and generating JSON output are highly beneficial for many OCR workflows.[20] With appropriate quantization, it runs effectively on an RTX 4090.[11]
   - **Ollama Tag**: Use qwen2.5vl and ensure selection of a 32B variant or a well-quantized larger variant if benchmarks strongly favor it and it demonstrably fits within the 24GB VRAM. Given that benchmarks suggest the 72B and 32B Qwen VL models have very similar OCR accuracy [23], the 32B version is likely the more practical and stable choice for a single RTX 4090.

Second Tier Recommendations (Good Performance with Some Caveats):

3. **Llama 3.2 Vision (11B)**
    - **Reasoning**: This model is proficient in document-centric OCR [19], is supported by the Ollama-OCR package [26], and has user-reported HTR accuracy of approximately 80%.[25] It is also very VRAM efficient, requiring only about 8GB.[31]
    - **Caveat**: Its primary limitation is that it supports English-only for combined image and text tasks.[31] If the user's "normal handwriting" is exclusively English, this model is a strong contender.
    - **Ollama Tag**: llama3.2-vision:11b.[4]
4. **Mistral Small 3.1 (24B, quantized)**
    - **Reasoning**: Users have reported good multilingual OCR capabilities [24] and HTR accuracy around 85%.[25] It is confirmed to run on an RTX 4090 [12] and features a large context window.
    - **Caveats**: There are potential concerns regarding lower accuracy on specific character details or diacritics.[24] Additionally, past performance and VRAM utilization issues with some Ollama versions might require workarounds.[24] It can also be more resource-intensive compared to some alternatives.
    - **Ollama Tag**: mistral-small3.1:24b or specific quantized tags such as 24b-instruct-2503-iq4_NL.[13]

Models to Approach with Caution for Primary HTR:

- **Gemma 3**: Despite its strong general multimodal capabilities, its performance in specific OCR benchmarks has been notably poor.[20]
- **LLaVA**: Appears inconsistent for OCR/HTR tasks, with warnings about incorrect output [26] and user feedback suggesting it is often better for general image description rather than precise text extraction.[25]

Practical Advice for Implementation:

- **Start with Top Tier Models**: Begin experimentation with Minicpm-V 2.6 and Qwen 2.5 VL (32B, quantized) as they show the most promise based on current evidence.
- **Select Appropriate Quantization**: For the Qwen 2.5 VL 32B model, use a reliable quantized version (e.g., Q4_K_M or similar GGUF quants available through Ollama). The 8B Minicpm models might perform well even at fp16 or higher-bit quantization levels, given their smaller base size.
- **Test with Representative Data**: The user must test the chosen models on samples of the specific "normal handwriting" they intend to process to ensure optimal performance for their use case.
- **Employ Effective Prompt Engineering**: Experiment with clear and specific prompts that directly ask for transcription. For models like Qwen, requesting JSON output might

be beneficial if structured data extraction is required.

- **Ensure Ollama is Updated**: Use an up-to-date version of Ollama, especially if considering models like Mistral Small 3.1, to mitigate potential past performance bugs. Consult Ollama documentation or GitHub issue trackers for any model-specific advice or known issues.
- **Consider Using the Ollama-OCR Package**: The Ollama-OCR Python package [26] can be a valuable tool as it provides a convenient wrapper for several models and may incorporate best practices for prompting and preprocessing.

6. Conclusion

The task of recognizing normal handwriting using locally run Ollama models on an NVIDIA RTX 4090 is indeed feasible, with several promising candidates emerging from the current landscape of open-source multimodal LLMs. The analysis indicates that **Minicpm-V 2.6 (8B)** and **Qwen 2.5 VL (32B, appropriately quantized)** stand out as particularly strong contenders. Their selection is supported by a combination of excellent OCR/HTR benchmark performance, positive user feedback specifically related to handwriting, and their comfortable VRAM footprint on an RTX 4090, which allows for efficient operation.

Llama 3.2 Vision (11B) presents a solid option for English-language document-based handwriting, offering good VRAM efficiency. Mistral Small 3.1 (24B) provides broad multilingual support, which is a significant advantage for diverse datasets, though it comes with some caveats regarding fine-grained accuracy and potential resource intensity. Conversely, models like Gemma 3 and LLaVA, despite their strengths in other multimodal areas, appear less suited for primary HTR tasks based on current OCR-specific benchmarks and user experiences.

The definition of "normal handwriting" can vary significantly, encompassing different styles, languages, and levels of legibility. Therefore, while these recommendations provide a strong starting point, **empirical testing is paramount**. The user must evaluate the top recommended models on their specific data and use case to determine the absolute best fit that balances accuracy, speed, and ease of use.

The field of local, open-source multimodal LLMs is characterized by rapid advancement.[2] Newer models and improved versions of existing ones will undoubtedly continue to emerge, potentially offering even better HTR performance and efficiency. A noteworthy trend is the increasing ability of general-purpose multimodal models to achieve high performance on specialized tasks like OCR/HTR. Models such as Qwen 2.5 VL and Minicpm-V/O are demonstrating that capabilities previously requiring highly niche, task-specific models can now be effectively delivered by broadly competent VLMs. This is likely due to training on vast and diverse datasets (e.g., Qwen's reported 18 trillion tokens [20]) or the development of highly effective architectures and encoders (e.g., Minicpm's focus on efficiency and OCR [40]). This "baking in" of specialized skills is evident when these models outperform some dedicated OCR tools or large proprietary systems in specific benchmarks.[20] This development lowers the barrier to entry for achieving high performance on specialized tasks like HTR and suggests a future where single, powerful, locally-run VLMs can handle a wider array of complex tasks without significant compromises. Staying engaged with the Ollama community and monitoring relevant benchmarks will be key to leveraging these future improvements and harnessing the evolving power of local AI.

Works cited

1. Benchmarking Large Language Models for Handwritten Text Recognition - arXiv, accessed May 16, 2025, http://arxiv.org/pdf/2503.15195
2. Best Multimodal Models Ollama - BytePlus, accessed May 16, 2025, https://www.byteplus.com/en/topic/516146
3. Ollama's new engine for multimodal models, accessed May 16, 2025, https://ollama.com/blog/multimodal-models
4. library - Ollama, accessed May 16, 2025, https://ollama.com/library?ref=noted.lol
5. Read Ollama in Action: Building Safe, Private AI with LLMs, Function Calling and Agents | Leanpub, accessed May 16, 2025, https://leanpub.com/ollama/read
6. Choosing the Right NVIDIA GPU for LLMs on the Ollama Platform - Database Mart, accessed May 16, 2025, https://www.databasemart.com/blog/choosing-the-right-gpu-for-popluar-llms-on-ollama
7. Benchmarking LLMs on NVIDIA RTX 4090 GPU Server with Ollama - Database Mart, accessed May 16, 2025, https://www.databasemart.com/blog/ollama-gpu-benchmark-rtx4090
8. aleibovici/ollama-gpu-calculator - GitHub, accessed May 16, 2025, https://github.com/aleibovici/ollama-gpu-calculator
9. library - Ollama, accessed May 16, 2025, https://ollama.com/library
10. We ran over half a million evaluations on quantized LLMs—here's what we found, accessed May 16, 2025, https://developers.redhat.com/articles/2024/10/17/we-ran-over-half-million-evaluations-quantized-llms
11. hardware requirements to run qwen 2.5 32B? : r/LocalLLaMA - Reddit, accessed May 16, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1gp4g8a/hardware_requirements_to_run_qwen_25_32b/
12. mistral-small3.1 - Ollama, accessed May 16, 2025, https://ollama.com/library/mistral-small3.1
13. LoTUs5494/mistral-small-3.1:24b/system - Ollama, accessed May 16, 2025, https://ollama.com/LoTUs5494/mistral-small-3.1:24b/blobs/6def0561329e
14. Mistral-Small-3.1-24B-Instruct-2503 (GGUF) - Ollama, accessed May 16, 2025, https://ollama.com/LoTUs5494/mistral-small-3.1:24b
15. orieg/gemma3-tools:27b-it-qat - Ollama, accessed May 16, 2025, https://ollama.com/orieg/gemma3-tools:27b-it-qat
16. A Step-by-Step Guide to Install Gemma-3 Locally with Ollama or Transformers - NodeShift, accessed May 16, 2025, https://nodeshift.com/blog/a-step-by-step-guide-to-install-gemma-3-locally-with-ollama-or-transformers
17. Qwen/Qwen2.5-Coder-32B-Instruct · Requesting information about hardware

resources, accessed May 16, 2025, https://huggingface.co/Qwen/Qwen2.5-Coder-32B-Instruct/discussions/28

18. OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches - arXiv, accessed May 16, 2025, https://arxiv.org/html/2502.01205v1
19. Llama 3.2 Guide: How It Works, Use Cases & More - DataCamp, accessed May 16, 2025, https://www.datacamp.com/blog/llama-3-2
20. Qwen-2.5-72b: Best Open Source VLM for OCR? - Apidog, accessed May 16, 2025, https://apidog.com/blog/qwen-2-5-72b-open-source-ocr/
21. MiniCPM-o 2.6: A GPT-4o Level MLLM for Vision, Speech and Multimodal Live Streaming on Your Phone - GitHub, accessed May 16, 2025, https://github.com/OpenBMB/MiniCPM-o
22. QwenLM/Qwen2.5-VL - GitHub, accessed May 16, 2025, https://github.com/QwenLM/Qwen2.5-VL
23. Qwen-2.5-72b is now the best open source OCR model : r/LocalLLaMA - Reddit, accessed May 16, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1jm4agx/qwen2572b_is_now_the_best_open_source_ocr_model/
24. Ollama now supports Mistral Small 3.1 with vision : r/LocalLLaMA - Reddit, accessed May 16, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1juc553/ollama_now_supports_mistral_small_31_with_vision/
25. Vision models that work well with Ollama - Reddit, accessed May 16, 2025, https://www.reddit.com/r/ollama/comments/1kiqggq/vision_models_that_work_well_with_ollama/
26. imanoop7/Ollama-OCR - GitHub, accessed May 16, 2025, https://github.com/imanoop7/Ollama-OCR
27. Mistral OCR, accessed May 16, 2025, https://mistral.ai/news/mistral-ocr
28. Show HN: Qwen-2.5-32B is now the best open source OCR model | Hacker News, accessed May 16, 2025, https://news.ycombinator.com/item?id=43549072
29. Qwen/Qwen2.5-VL-32B-Instruct - Hugging Face, accessed May 16, 2025, https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
30. llama3.2-vision:11b - Ollama, accessed May 16, 2025, https://ollama.com/library/llama3.2-vision:11b
31. Introducing Llama 3.2 Vision: A Comprehensive Look - daily.dev, accessed May 16, 2025, https://app.daily.dev/posts/introducing-llama-3-2-vision-a-comprehensive-look-tgv6dt5jy
32. Ollama-OCR - Reddit, accessed May 16, 2025, https://www.reddit.com/r/ollama/comments/1j3fh7d/ollamaocr/
33. Mistral 3.1 Small Review: Strengths, Benchmarks, and Use Cases - Latenode, accessed May 16, 2025, https://latenode.com/blog/mistral-3-1-small-review
34. How to fix slow inference speed of mistral-small 3.1 when using Ollama - Reddit, accessed May 16, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1judvfg/how_to_fix_slow_inference_speed_of_mistralsmall/

35. Vision - IQ3_XXS - Mistral Small 3.1 24b · Issue #10393 - GitHub, accessed May 16, 2025, https://github.com/ollama/ollama/issues/10393
36. Gemma 3 model card | Google AI for Developers - Gemini API, accessed May 16, 2025, https://ai.google.dev/gemma/docs/core/model_card_3
37. Gemma 3 Release - a google Collection : r/LocalLLaMA - Reddit, accessed May 16, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1j9dkvh/gemma_3_release_a_google_collection/
38. SCAM: A Real-World Typographic Robustness Evaluation for Multimodal Foundation Models - arXiv, accessed May 16, 2025, https://arxiv.org/html/2504.04893v1
39. Llava 34B Released! Exceeding Gemini Pro in Performance Benchmarks? - YouTube, accessed May 16, 2025, https://www.youtube.com/watch?v=QKGUIHBbGyo
40. minicpm-v:8b-2.6-fp16 - Ollama, accessed May 16, 2025, https://ollama.com/library/minicpm-v:8b-2.6-fp16
41. MiniCPM-o 2.6: An 8B size, GPT-4o level Omni Model runs on device - Reddit, accessed May 16, 2025, https://www.reddit.com/r/LocalLLaMA/comments/1i11961/minicpmo_26_an_8b_size_gpt4o_level_omni_model/
42. openbmb/MiniCPM-o-2_6 - Hugging Face, accessed May 16, 2025, https://huggingface.co/openbmb/MiniCPM-o-2_6
43. Qwen2.5-Omni Flagship vs MiniCPM-V Powerhouse: Complete Analysis of Parameters, Hardware, Resources, and Advantages - Tech Explorer, accessed May 16, 2025, https://stable-learn.com/en/qwen-omni-vsminicpm-v/
44. I'm absolutely amazed at how capable the new 1B model is, considering it's just - Hacker News, accessed May 16, 2025, https://news.ycombinator.com/item?id=41652377
45. moondream - Ollama, accessed May 16, 2025, https://ollama.com/library/moondream