A Review of Foundation Models and Efficient Architectures in Machine Learning (2023–2024)

Uday Vaidya
MCA , Department of Computer Application,
DAV Institute of Engineering & Technology, Jalandhar, Punjab
Corresponding Author: udayvaidya13@gmail.com

Date of Submission: 17-04-2026 Mentor/Guide: Ms. Ankita Chadha

ABSTRACT: *The field of machine learning has witnessed unprecedented growth from 2023 to 2024, with foundational breakthroughs spanning large language models, computer vision, and efficient neural architectures. This review paper critically examines five of the most impactful research contributions of this period: Llama 2 (Meta AI, 2023), the Segment Anything Model (SAM, Meta AI, 2023), Mistral 7B (Mistral AI, 2023), Gemma (Google DeepMind, 2024), and Visual AutoRegressive Modeling (VAR, NeurIPS 2024 Best Paper). For each work, we analyze the motivation, technical methodology, experimental evaluation, and broader impact on the research and practitioner communities. Cross-cutting trends identified include the democratization of high-capability open-source models, the rise of Reinforcement Learning from Human Feedback (RLHF) as a standard alignment technique, the growing primacy of architectural efficiency over raw parameter scale, and the maturation of foundation-model thinking across both language and vision domains. This survey is intended as a structured synthesis for researchers, graduate students, and industry practitioners seeking to understand and build upon the current state of the art in machine learning.*

KEYWORDS: Large Language Models, Segment Anything Model, Reinforcement Learning from Human Feedback (RLHF), Open-Source AI, Autoregressive Image Generation, Vision Transformers, Foundation Models, Grouped-Query Attention, Sliding Window Attention, Efficient Inference.

I. INTRODUCTION

The years 2023 and 2024 are very important in the history of machine learning. After ChatGPT was made available to the public in late 2022, the research community entered a time of quick, widespread innovation. Foundational models, new ways of training, and improvements to architecture that were made just for this purpose came out at a speed that was almost unheard of in the field's history. Open-source consortia and well-funded proprietary laboratories both worked hard to make AI systems that were capable, safe, and computationally efficient. This led to a lot of important scientific literature.

There were a lot of competing pressures during this time. People wanted to push the limits of what AI could do, they needed to make powerful models available to more researchers, and they were starting to realise that safety and alignment should be top design concerns, not afterthoughts. These tensions influenced the research agenda in constructive manners, yielding outputs that both enhanced benchmark performance and broadened the discourse regarding the construction and governance of AI systems.

Five papers from this time stand out because they made important scientific contributions, had a wide range of effects on the community, and used a new method. Llama 2 (Touvron et al., 2023) made high-quality large language models available to everyone by releasing them for free and using a strict RLHF-based alignment pipeline. The Segment Anything Model, or SAM (Kirillov et al., 2023), changed the way image segmentation is done by making it a promptable, general-purpose ability and releasing a dataset that was larger than any other. Mistral 7B (Jiang et al., 2023) showed, in a surprising way, that making smart choices about the architecture of a model could make a 7-billion-parameter model work better than models that are almost twice as big. Gemma (Google DeepMind Team, 2024) added Gemini-class features to the open-source ecosystem and made a strong promise to build responsible deployment infrastructure. Finally, Visual AutoRegressive Modelling, or VAR (Tian et al., 2024), fundamentally reconceived autoregressive image generation as a multi-scale coarse-to-fine process, earning the NeurIPS 2024 Best Paper Award.The review is structured in this way: Section II looks at each of the five chosen papers in terms of their motivation, methodology, and empirical results. Section III offers a systematic comparative analysis across shared dimensions. Section IV talks about the bigger effects, limits, and open research questions that come up from these works. Section V ends the review with directions for future investigation.

II. LITERATURE REVIEW

Paper 1: Llama 2 — Open Foundation and Fine-Tuned Chat Models (Meta AI, 2023)

Touvron et al. (2023) introduced Llama 2, a family of autoregressive language models spanning 7 billion, 13 billion, and 70 billion parameters. All variants were pretrained on approximately two trillion tokens drawn from a curated mixture of publicly available internet text, with a data blend and preprocessing pipeline described in greater detail than most contemporaneous works. The pretraining recipe built on lessons from the original

© 2025 Uday Vaidya, DAV Institute of Engineering & Technology, Jalandhar

LLaMA paper but introduced architectural improvements including a longer context window (4,096 tokens), grouped-query attention for the larger variants to reduce inference-time memory bandwidth demands, and improved normalization strategies.

The defining contribution of Llama 2, however, was not scale or pretraining alone, but the rigorous application of Reinforcement Learning from Human Feedback (RLHF) to produce safety- and helpfulness-aligned chat variants designated Llama 2-Chat. The alignment pipeline proceeded in two stages. First, supervised fine-tuning on a curated set of high-quality demonstrations established a baseline conversational behavior. Second, an iterative RLHF loop refined the model's outputs using two separate reward models—one optimizing for helpfulness and one optimizing for safety—both trained on over one million human preference annotations collected through an in-house annotation pipeline. This dual-reward formulation allowed the authors to explicitly trade off helpfulness against harmlessness, a distinction that had not been operationalized at this scale in any prior publicly documented system.

The authors also introduced Ghost Attention (GAtt), a fine-tuning technique designed to address a common failure mode in multi-turn dialogue: the tendency of RLHF-trained models to lose track of system-level instructions as a conversation lengthens. GAtt conditions the model's attention on the system prompt at every turn during training, effectively teaching the model to treat system instructions as persistent context rather than as initial context that can be overwritten by subsequent dialogue. This technique improved both the coherence and the instruction-following reliability of Llama 2-Chat in long conversations.

Empirically, Llama 2-Chat 70B achieved performance broadly comparable to GPT-3.5 on benchmarks including MMLU, TruthfulQA, and HumanEval, while the smaller Llama 2 variants outperformed the corresponding LLaMA 1 models on nearly every evaluated task. Human preference studies showed Llama 2-Chat 70B on par with or preferred over several closed proprietary models.

The paper's transparent safety reporting—including quantitative red-teaming results, model cards, and explicit discussion of known failure modes—set a new benchmark for responsible model release and influenced nearly every subsequent open-source LLM publication.

Paper 2: Segment Anything Model — SAM (Meta AI, 2023)

Kirillov et al. (2023) introduced the Segment Anything Model (SAM), a foundation model for image segmentation built around the central concept of promptability. Whereas prior segmentation models were trained and evaluated on fixed task definitions—semantic segmentation, instance segmentation, panoptic segmentation—SAM was designed to produce a valid and plausible segmentation mask in response to any arbitrary user-supplied prompt. Supported prompt types include spatial points (foreground and background), bounding boxes, masks from a previous iteration, and free-form text, which can be combined in any combination. This flexible prompt interface made SAM applicable to an enormous range of downstream tasks without the need for task-specific fine-tuning.

The construction of the training dataset, SA-1B, was itself a major scientific contribution. The authors developed a three-stage iterative data engine in which human annotators, assisted by progressively more capable versions of the model under development, annotated images with segmentation masks. Beginning with a fully manual annotation stage, the pipeline transitioned to a semi-automatic stage in which SAM proposed masks that annotators refined, and finally to a fully automatic stage in which SAM generated and filtered its own annotations. The resulting SA-1B dataset contains over 11 million diverse images and approximately 1.1 billion high-quality segmentation masks—orders of magnitude larger than any prior publicly released segmentation benchmark.

SAM's architecture comprises three components. A large image encoder, based on a Vision Transformer pretrained with Masked Autoencoding (MAE), processes the input image into a dense embedding only once per image, enabling prompt-based queries to be answered cheaply thereafter. A lightweight prompt encoder maps each supported prompt type into a vector representation compatible with the image embedding. A fast mask decoder, consisting of a small transformer-based network, predicts segmentation masks from the combined image and prompt embeddings in approximately 50 milliseconds. To handle ambiguity—where a single prompt may be consistent with multiple valid segmentations—the decoder predicts three candidate masks and ranks them by an estimated confidence score.

SAM demonstrated strong zero-shot generalization across 23 diverse segmentation datasets spanning natural images, medical images, aerial imagery, and video frames. In many settings, SAM matched or exceeded the performance of fully supervised specialist models. The model and the SA-1B dataset were released publicly, catalyzing a broad wave of downstream research and applications in domains including surgical robotics, autonomous driving, and remote sensing.

Paper 3: Mistral 7B (Mistral AI, 2023)

Jiang et al. (2023) introduced Mistral 7B, a 7-billion-parameter autoregressive language model whose central thesis was that architectural ingenuity can substitute for—and in many cases surpass—raw parameter count as a lever of model quality. The paper challenged the prevailing assumption that model capability scales primarily with parameter count, demonstrating that a 7B model could outperform a 13B LLaMA 2 model and match or exceed a 34B LLaMA 1 model on most evaluated benchmarks through the application of two targeted attention mechanisms.

The first mechanism, Grouped-Query Attention (GQA), restructures the standard multi-head attention computation by partitioning query heads into groups, each of which shares a single key-value head. This reduces the size of the key-value cache during autoregressive generation by a factor proportional to the number of groups, directly reducing memory bandwidth requirements at inference

© 2025 Uday Vaidya, DAV Institute of Engineering & Technology, Jalandhar

time and enabling faster token generation on standard GPU hardware. GQA

had been proposed in earlier work but Mistral 7B demonstrated its practical value in a competitive production-scale setting.

The second mechanism, Sliding Window Attention (SWA), replaces the standard full-context attention mask with a rolling window that limits each token's attention to a fixed-size neighborhood of recent tokens. By stacking multiple SWA layers, information from tokens well beyond the immediate window can still influence later layers through the chain of overlapping attention windows—a property the authors call the "effective receptive field." SWA reduces the computational cost of attention from quadratic to linear in sequence length for long inputs, enabling Mistral 7B to handle sequences of up to 8,192 tokens without prohibitive memory cost.

Mistral 7B was evaluated on a broad suite including commonsense reasoning benchmarks (HellaSwag, WinoGrande, PIQA, ARC Easy and Challenge), mathematical reasoning (GSM8K), and code generation (HumanEval). The model outperformed LLaMA 2 13B on every evaluated benchmark and matched or exceeded LLaMA 1 34B on most reasoning tasks. An instruction-following variant, Mistral 7B-Instruct, outperformed LLaMA 2 13B-Chat in human preference studies. The model was released publicly under an Apache 2.0 license, making it one of the most permissively licensed competitive models of its era and the dominant community base for fine-tuning work through 2024.

Paper 4: Gemma — Open Models Based on Gemini Research (Google DeepMind, 2024)

The Gemma team at Google DeepMind (2024) introduced Gemma, a family of lightweight open language models in 2-billion and 7-billion parameter configurations. Unlike many open model releases that adapted or distilled from preexisting public checkpoints, Gemma was built from the same foundational research infrastructure—data pipelines, tokenization, architectural choices, and training recipes—that underpinned the proprietary Gemini Ultra, Pro, and Nano series. This lineage gave Gemma access to a quality of pretraining data curation and training stability techniques that had not previously been available in an openly released model.

The 7B model was pretrained on approximately six trillion tokens of primarily English-language text drawn from web documents, code repositories, and mathematics corpora. The model architecture adopted multi-query attention, rotary positional embeddings (RoPE), and gated linear units with the GeGLU activation function—choices aligned with the Gemini family design. Both 2B and 7B base models were released alongside instruction-tuned variants, the latter fine-tuned using a combination of supervised instruction tuning and RLHF. Context length was set at 8,192 tokens for both variants.

In evaluations across academic benchmarks, Gemma 7B outperformed Llama 2 7B and 13B on the majority of tasks examined, including MMLU, GSM8K, HumanEval, and BIG-Bench Hard. The Gemma 2B variant demonstrated competitive results for its scale, outperforming several contemporaneous 7B-class models on specific mathematical and coding tasks. Notably, both variants were evaluated with a strong emphasis on safety, undergoing red-teaming for harmful content generation and responsible use evaluations prior to release.

Accompanying the model weights, Google DeepMind released a Responsible Generative AI Toolkit that included safety classifiers, model debugging tools, and a comprehensive usage policy. This level of deployment infrastructure—model weights paired with safety evaluation tools and explicit governance documentation—represented a meaningful advance in operationalizing AI safety research for open-source practitioners and significantly raised the bar for what responsible open model release could look like.

Paper 5: Visual AutoRegressive Modeling — VAR (NeurIPS 2024 Best Paper)

Tian et al. (2024) proposed Visual AutoRegressive Modeling (VAR), a fundamentally new approach to autoregressive image generation that challenges the dominant paradigm of raster-scan token prediction. Prior autoregressive image models, from the original VQ-VAE to VQGAN and MaskGIT, treated image generation as sequential prediction of discrete tokens arranged in a fixed spatial order, typically row by row from top-left to bottom-right. While this

formulation allowed the application of language-model-style architectures to image generation, it suffered from a mismatch between the sequential prediction objective and the inherently two-dimensional spatial structure of images, limiting both sample quality and scaling efficiency.

VAR reformulates image generation as a sequence of next-scale predictions rather than next-token predictions. Beginning from a single 1×1 representation of the image —essentially a global semantic embedding—the model autoregressively predicts progressively higher-resolution token maps: 2×2, 4×4, 8×8, and so on up to the final full-resolution representation. Each step predicts an entire token map in parallel given all previous, coarser-resolution maps. This coarse-to-fine structure means the model must learn to commit to global semantic and structural decisions early and then progressively refine spatial detail, which mirrors how human visual attention and artistic composition processes work.

Architecturally, VAR uses a GPT-2-style transformer with modifications to support variable-length sequences of 2D token maps. A multi-scale VQ tokenizer compresses images into hierarchical token pyramids at multiple resolutions. The training objective is standard cross-entropy prediction over the token vocabulary, applied to each scale level in sequence. No diffusion process, denoising network, or iterative refinement is required at inference time—generation is a single forward pass through the scale sequence, producing all tokens at each resolution level in parallel before proceeding to the next.

VAR demonstrated that autoregressive image generation models exhibit scaling laws closely analogous to those observed in large language models—a result that had not

© 2025 Uday Vaidya, DAV Institute of Engineering & Technology, Jalandhar

previously been established for visual generation. Performance on class-conditional ImageNet 256×256 generation improved predictably with model parameter count and training compute across five model sizes (from 310M to 2B parameters), following a power law in FID score. The largest evaluated model, VAR-d30, achieved an FID of 1.73 on ImageNet 256×256, surpassing state-of-the-art diffusion transformers (DiT, SiT) while requiring approximately 20 times fewer function evaluations and being 200 times faster at inference on comparable hardware. The NeurIPS 2024 Best Paper award reflected broad recognition of VAR's significance in bridging language and vision modeling paradigms.

III. COMPARATIVE ANALYSIS

Table 1: Comparative Summary of Selected Machine Learning Papers

Paper /	Year	Key	Benchma	Key
Llama 2	2023	RLHF + SFT	MMLU, HumanEv	Open-source LLM with
SAM (Segment Anything)	2023	Promptab le segmentat	SA-1B dataset	Universal image segmentatio
Mistral 7B	2023	Grouped-Query Attention	HellaSwa g, WinoGra	Efficient 7B model outperformi
Gemma (Google)	2024	Gemini-derived pretrainin	MATH, GSM8K, BIG-	Open models for responsible
VAR (Visual AutoRegressi	2024	Coarse-to-fine	ImageNet , GenEval	NeurIPS 2024 Best

Table 1 (shown above) provides a structured summary of the five reviewed papers across the dimensions of publication year, core technical innovation, primary evaluation benchmarks, and principal scientific contribution. Analyzing these papers together reveals several important cross-cutting themes that characterized the 2023–2024 machine learning research landscape.

Open-source democratization emerged as a defining trend of the period. All five reviewed papers released model weights, datasets, or both under publicly accessible licenses. Llama 2 and Mistral 7B adopted permissive Apache-compatible licenses; SAM released both model weights and the SA-1B dataset; Gemma provided model weights alongside a safety toolkit; VAR released model weights and code. This collective commitment to openness created a rich ecosystem of derivative research and practical applications that would have been impossible under a closed-model paradigm.

The centrality of alignment and safety reflects a maturation of the field's values. Llama 2's dual-reward RLHF pipeline and Gemma's responsible AI toolkit both represent substantial investments in making deployed models not just capable but behaviorally reliable. This trend stands in contrast to earlier generations of open language model releases, which typically offered base models without alignment work, leaving safety tuning to downstream users.

Architectural efficiency as a first-class research objective is most clearly exemplified by Mistral 7B, where GQA and SWA together enabled a model to outperform one of twice its size. However, the same philosophy appears in SAM's design of a single-image-encoding, multi-prompt-decoding architecture that amortizes the expensive encoder computation across all subsequent queries, and in

VAR's parallel within-scale prediction that achieves fast generation without iterative refinement.

Foundation model universality continued to advance, with SAM establishing promptable segmentation as a task-agnostic capability and VAR suggesting that visual generation models may be amenable to the same scaling recipes that produced transformative gains in language modeling. Together, these trends point toward a future in which a small number of broadly capable, efficiently deployed, and responsibly governed models serve as universal platforms for downstream AI applications.

IV. DISCUSSION

The five papers reviewed here collectively illuminate both the remarkable achievements and the persistent open challenges of machine learning research in 2023–2024. Each work makes a significant scientific contribution, yet each also points toward limitations and unanswered questions that motivate continued investigation.

Llama 2 and Gemma together demonstrate that safety-tuned open language models are feasible at scale, but both papers explicitly acknowledge that RLHF-based alignment does not provide formal safety guarantees. Jailbreaking attacks—techniques that elicit harmful outputs from aligned models by manipulating prompt structure—remain effective against even carefully RLHF-trained systems. This highlights the need for robustness research beyond preference optimization, including adversarial training, formal verification of behavioral constraints, and interpretability-based approaches to understanding and controlling model behavior. The dual-reward formulation in Llama 2 is a step in the right direction, but the fundamental brittleness of learned reward models remains an open problem.

SAM's strong zero-shot performance on natural images is impressive, but its evaluation revealed consistent performance gaps on medical imaging and remote sensing tasks, domains with substantial distribution shift relative to the SA-1B training data. Foundation models trained on internet-scale natural image data may systematically underperform on specialized domains, motivating work on domain-adaptive pre training, continual learning for visual foundation models, and the construction of large-scale annotated datasets in underrepresented image domains. The SA-1B data engine itself—a human-AI collaborative

© 2025 Uday Vaidya, DAV Institute of Engineering & Technology, Jalandhar

annotation pipeline—represents a promising template for efficiently scaling annotation in new domains.

Mistral 7B and Gemma raise fundamental questions about the practical value of very large language models for most real-world deployment scenarios. If a 7-billion-parameter model running on consumer hardware can match the capability of a 70-billion-parameter model on the majority of everyday tasks, the economic case for deploying large models weakens significantly. This has profound implications for edge inference, privacy-preserving local computation, and AI accessibility in resource-constrained settings—including healthcare and education applications in developing regions where cloud API access is cost-prohibitive. The remaining gap between 7B and 70B class models on complex multi-step reasoning and knowledge-intensive tasks, however, motivates continued work on knowledge distillation, parameter-efficient fine-tuning, and speculative decoding.

VAR opens a genuinely new research frontier by establishing that autoregressive visual generation models exhibit language-model-like scaling laws. If this result generalizes—and early follow-up work suggests it may—systematic investment in compute and data quality could yield predictable improvements in image generation quality, similar to the empirical scaling laws that guided the development of GPT-3, PaLM, and Chinchilla. Open questions include whether VAR's coarse-to-fine paradigm extends to video generation (where temporal consistency over sequences of token maps is required), 3D shape synthesis, and audio generation. The relationship between the multi-scale VQ tokenizer quality and downstream generation performance is also underexplored and likely to be a productive area for future work.

A broader challenge cutting across all five papers is the evaluation methodology. Each work reports performance on standard academic benchmarks, but the relationship between benchmark scores and real-world utility is often unclear and increasingly contested as capable models begin to saturate established benchmarks. The development of more ecologically valid, adversarially robust, and domain-specific evaluation frameworks is an important priority for the field as a whole.

V. CONCLUSION

This review has examined five landmark machine learning papers published between 2023 and 2024, spanning large language models, universal image segmentation, efficient neural architectures, open foundation models, and autoregressive visual generation. Together, these works document a field undergoing rapid and consequential transformation across multiple dimensions simultaneously.

Llama 2 established that responsible open release of RLHF-aligned large language models is both technically feasible and immensely impactful for the research community. SAM redefined image segmentation as a promptable, universal capability and set a new standard for dataset scale and diversity. Mistral 7B demonstrated that architectural efficiency—specifically through Grouped-Query Attention and Sliding Window Attention—can enable compact models to outperform systems nearly twice their size, challenging assumptions about the primacy of scale. Gemma demonstrated that Gemini-class training quality and responsible deployment infrastructure could coexist in a publicly released model family. VAR reconceived autoregressive image generation as a multi-scale prediction problem, achieved state-of-the-art generation quality at a fraction of the inference cost of diffusion models, and established visual generation scaling laws for the first time.

Collectively, these papers reflect a machine learning research culture that is increasingly attentive to the practical concerns of deployment—efficiency, safety, accessibility, and governance—alongside the scientific pursuit of capability. This dual orientation, if sustained, may prove to be the defining characteristic of the current era of machine learning research.

Future work should focus on several priority areas: multi-modal unification of language and vision foundation models; robustness improvements to RLHF-based alignment that withstand adversarial prompting; domain-adaptive pretraining pipelines for underrepresented image domains; systematic study of scaling laws for visual and multi-modal generation beyond ImageNet benchmarks; and the development of evaluation frameworks that better capture real-world utility. The contributions reviewed in this paper provide both the technical building blocks and the conceptual framing for this next generation of machine learning research.

REFERENCES

[1] Touvron, H., Martin, L., Stone, K., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288. Meta AI.

[2] Kirillov, A., Mintun, E., Ravi, N., et al. (2023). Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. arXiv:2304.02643.

[3] Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. arXiv:2310.06825. Mistral AI.

[4] Gemma Team, Google DeepMind. (2024). Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.

[5] Tian, K., Jiang, Y., Yuan, Z., Peng, B., & Wang, L. (2024). Visual AutoRegressive Modeling: Scalable Image Generation via Next-Scale Prediction. NeurIPS 2024 Best Paper. arXiv:2404.02905.

[6] OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774.

[7] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback (InstructGPT). NeurIPS 2022.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

[9] Anil, R., Chowdhery, A., et al. (2023). PaLM 2 Technical Report. arXiv:2305.10403. Google.

[10] He, K., Chen, X., Xie, S., et al. (2022). Masked Autoencoders Are Scalable Vision Learners. CVPR 2022.

© 2025 Uday Vaidya, DAV Institute of Engineering & Technology, Jalandhar