The Llama 4 Herd: Architecture, Training, Evaluation, and Deployment Notes

Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, Adam Stelle, Adeel Cheema, Adela Kabiljo, Adina Katz, Adithya Gangidi, Aditya Tayade, Adolfo Victoria, Adrian Samatan Alastuey, Adrien Conrath, Afroz Mohiuddin, Ahmed Sharif, Ahnaf Siddiqui, Ahuva Goldstand, Aijung Li, Aidan Boyd, Aidin Kazemi Daliri, Aisha Iqbal, Ajay Menon , Ajit Mathews, Akhil Mathur, and others* †

Abstract

This document consolidates publicly reported technical details about Meta's Llama 4 model family. It summarizes (i) released variants (Scout and Maverick) and the broader "herd" context including the previewed Behemoth teacher model, (ii) architectural characteristics beyond a high-level MoE description—covering routed/shared-expert structure, early-fusion multimodality, and long-context design elements reported for Scout (iRoPE and length generalization strategies), (iii) training disclosures spanning pre-training, mid-training for long-context extension, and post-training methodology (lightweight SFT, online RL, and lightweight DPO) as described in release materials, (iv) developer-reported benchmark results for both base and instruction-tuned checkpoints, and (v) practical deployment constraints observed across major serving environments, including provider-specific context limits and quantization packaging. The manuscript also summarizes licensing obligations relevant to redistribution and derivative naming, and reviews publicly described safeguards and evaluation practices. The goal is to provide a compact technical reference for researchers and practitioners who need precise, source-backed facts about Llama 4.

Keywords: large language models, mixture-of-experts, multimodal models, long context, evaluation, deployment, licensing

1 Overview and scope

Llama 4 refers to a set of foundation and instruction-tuned models announced by Meta in April 2025, including open-weight releases (Scout and Maverick) and a previewed teacher model (Behemoth)[1]. The first released variants are Llama 4 Scout (17B active parameters, 16 experts; 109B total parameters) and Llama 4 Maverick (17B active parameters, 128 experts; 400B total parameters), both described as natively multimodal (text and image inputs; text and code outputs) and multilingual (12 supported languages) [2, 3, 4]. Meta also previewed a substantially larger teacher model, *Llama 4 Behemoth*, which was reported as still in training at the time of the Scout/Maverick release [5, 6]. Meta's release announcement provides the top-level framing for the Llama 4 "herd," including the positioning of Scout and Maverick and the preview of Behemoth[1].

*The complete author list appears in Appendix A.

†This manuscript is an independent synthesis of publicly available materials and is not an official Meta publication, with credit given to the Meta Llama contributors. All product names and trademarks are the property of their respective owners.

The release announcement also makes several comparative performance claims (e.g., Scout "best in its class" with 10M context, Maverick emphasizing performance-per-cost and an "experimental chat" LMArena ELO reported as 1417). These statements are marketing-facing claims and should be interpreted separately from the model-card benchmark tables reproduced in Section 5 [1].

This document is an independent survey of public materials. Reported benchmark numbers are attributed to the model cards unless stated otherwise; they should be treated as *developer-reported* results with the usual caveats around evaluation harnesses, prompting, and postprocessing.

Some public reporting describes Llama 4 as a broader "multimodal system" spanning multiple media types. In this manuscript, "multimodal" refers specifically to the capabilities documented for the released open-weight Scout and Maverick checkpoints (text-and-image inputs with text/code outputs) as described in the official model cards[2, 3, 5].

2 Model family and specifications

Table 1 summarizes the high-signal specifications that recur across official distribution channels.

Item	Llama 4 Scout	Llama 4 Maverick
Released variants	Base (pretrained) and instruction-tuned checkpoints[2, 3]	Base (pretrained) and instruction-tuned checkpoints; FP8 quantized weights are distributed for the instruction-tuned Maverick artifact[3, 7]
Architecture	MoE, early-fusion backbone for native multimodality[2, 8, 4]	MoE, early-fusion backbone for native multimodality[2, 8, 4]
Activated / total params	17B active; 109B total; 16 experts[2, 8]	17B active; 400B total; 128 experts[3, 4, 1]
Modalities	Text+image input; text+code output[2]	Text+image input; text+code output[2]
Supported languages	12 languages listed in the model card (Arabic, English, French, German, Hindi, Indonesian, Italian, Portuguese, Spanish, Tagalog, Thai, Vietnamese)[3]	Same 12-language list[3]
Pretraining coverage	Pretraining spans a broader set of languages (200 total languages reported)[3]	Same statement[3]
Knowledge cutoff	August 2024 (reported)[2]	August 2024 (reported)[2]
Token count	$\sim$ 40T tokens (reported)[2]	$\sim$ 22T tokens (reported)[2]
Context length (model)	10M tokens (reported)[2]	1M tokens (reported)[3]

Table 1: Model specifications and metadata as reported in official model cards and partner documentation.

2.1 Distributed artifacts and identifiers

Table 2 lists the canonical, citable identifiers used by major distribution channels. When reporting results, authors should name the specific artifact (and, ideally, the repository revision).

Artifact	Identifier (distribution channel)
Scout (base)	`meta-llama/Llama-4-Scout-17B-16E` (Hugging Face)[2]
Scout (instruct)	`meta-llama/Llama-4-Scout-17B-16E-Instruct` (Hugging Face)[9]
Maverick (instruct, bf16)	`meta-llama/Llama-4-Maverick-17B-128E-Instruct` (Hugging Face)[3]
Maverick (instruct, FP8)	`meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8` (Hugging Face)[7]
Llama Guard 4	`meta-llama/Llama-Guard-4-12B` (Hugging Face)[10]

Table 2: Commonly referenced Llama 4 artifacts and their canonical identifiers.

2.2 Architecture details beyond "MoE + early fusion"

Public release materials describe several architectural choices that are useful for interpreting the Llama 4 design space, particularly for long-context and multimodal workloads[1, 2, 3].

MoE routing and layer structure. Meta describes Llama 4 as its first Llama generation using a mixture-of-experts backbone. For Maverick, Meta reports alternating dense and MoE layers to improve inference efficiency; in the MoE layers, each token is routed to a shared expert and one routed expert among 128, so only a subset of parameters are activated per token even though all expert weights are resident in memory[1].

iRoPE for length generalization (Scout). Meta describes a long-context architecture choice for Scout that interleaves attention layers without positional embeddings with RoPE-based layers and applies inference-time attention temperature scaling to improve length generalization. This design is presented as a key component of Scout's long-context behavior[1].

Vision encoder and early fusion. The models are described as natively multimodal via early fusion: text and vision tokens are integrated in a unified backbone, enabling joint training over large-scale text, image, and video data[1]. Meta also reports an updated vision encoder based on MetaCLIP, trained separately while conditioning on a frozen Llama model to better adapt the encoder to the language backbone[1].

MetaP hyperparameter transfer. Meta reports a training approach (MetaP) intended to stabilize hyperparameter selection (e.g., per-layer learning rates and initialization scales) and improve transferability across changes in batch size, width/depth, and total training tokens[1].

2.3 Behemoth as teacher and codistillation into released models

Meta previews *Llama 4 Behemoth* as a teacher model for the released Scout and Maverick variants. In the release announcement, Behemoth is described as a multimodal MoE model with 288B active parameters, 16 experts, and nearly 2T total parameters, and as still in training at the time of the Scout/Maverick release.[1]

Meta further states that Maverick was *codistilled* from Behemoth, describing a distillation objective that combines hard targets with soft targets via a dynamically weighted loss during training. The announcement also notes an efficiency motivation: codistillation during pretraining amortizes the cost of teacher forward passes across much of the student training mixture, with additional teacher-generated targets produced for newer data incorporated after that point.[1]

2.4 Native multimodality and "early fusion"

Both Scout and Maverick are described as *natively* multimodal using an *early-fusion* backbone[2, 8, 4]. In early fusion, visual inputs are incorporated into the same transformer stream (rather than via a late-stage adapter-only pathway). From a systems perspective, this typically simplifies the inference API (single model call) but places stricter requirements on preprocessing and sequence construction.

Public materials report different multi-image test regimes: the model cards describe evaluation with a smaller number of images per prompt (e.g., up to five in some reported tests), while Meta's release announcement reports post-training tests with good results up to eight images[3, 1]. In practice, supported image counts and total image token budgets may vary by serving provider and client library.

2.5 MoE compute vs. memory trade-offs

The MoE design implies that only a fraction of the total parameters are activated per token, which can reduce *compute* relative to dense models of comparable total parameter count; however, *memory footprint* for weights remains driven by total parameters because all expert weights must typically reside in GPU memory (or be streamed efficiently). Cloudflare's deployment note explicitly highlights the challenge: serving requires loading full model weights ("over 200 GB" for Scout-class weights) and maintaining substantial KV cache for long contexts[8]. A practical implication is that many hosted offerings cap usable context length below the model's architectural maximum (Section 8).

3 Training data, freshness, and reported training footprint

3.1 Data sources and cutoff

Meta reports that Scout and Maverick were pretrained on multimodal data drawn from a mixture of publicly available data, licensed data, and information from Meta's products and services, including publicly shared posts from Instagram and Facebook and interactions with Meta AI[2]. The pretraining data cutoff is reported as August 2024[2].

Meta's release announcement adds several high-level pretraining details. It emphasizes that early fusion enables joint pretraining over large volumes of text, images, and video data. In the portion of the announcement describing the *Behemoth* pre-training run, Meta reports an overall data mixture of more than 30T tokens spanning diverse text, image, and video datasets[1]. This *Behemoth*-context figure should not be conflated with the per-model token-count disclosures in the Llama 4 model cards (reported as $\sim$ 40T for Scout and $\sim$ 22T for Maverick)[2].

The announcement also states that Llama 4 pretraining covered 200 languages, including more than 100 languages with at least 1B tokens each, and that the multilingual token volume is substantially larger than in Llama 3[1].

Token-count disclosures differ by source and scope: the model cards report per-model token counts for Scout and Maverick, while the release announcement reports an overall training-mixture scale; readers should cite the specific source corresponding to the intended claim[2, 1].

The announcement further describes a continued training phase ("mid-training") that incorporates new recipes and specialized datasets, including long-context extension, as part of enabling

Scout's reported 10M-token input length[1]. In addition, Meta reports training both Scout and Maverick on a wide variety of images and video-frame stills to improve general visual understanding, including temporal activity cues[1]. In this context, "video" refers to training on sampled frames/stills; the released open-weight Scout and Maverick checkpoints are documented as accepting images (not raw video streams) as inputs[2, 3].

Multi-image regime. Meta reports that the models were pretrained with up to 48 images per prompt and that post-training produced good results up to eight images[1]. This differs from the more conservative multi-image regimes described in the model cards; practitioners should treat multi-image limits as implementation- and endpoint-dependent[3].

3.2 Reported compute and emissions

The Scout model card reports cumulative pretraining compute of 7.38M H100-80GB GPU-hours across Scout and Maverick, with a breakdown of 5.0M GPU-hours (Scout) and 2.38M GPU-hours (Maverick), plus a location-based emissions estimate of 1,999 tons CO_2eq and a market-based estimate of 0 tons CO_2eq (due to renewable energy matching as described by Meta)[2]. These values are self-reported and intended as transparency disclosures.

Back-of-the-envelope throughput (informative). Using the reported token counts and GPU-hours ($\sim$ 40T / 5.0M and $\sim$ 22T / 2.38M), one obtains rough averages on the order of 8–9M tokens per GPU-hour. This is only a coarse sanity check because token counts are approximate, GPU utilization varies across phases, and multimodal training pipelines include non-token compute.

4 Post-training methodology

Meta's announcement provides a relatively detailed narrative of post-training for Maverick and Scout[1].

4.1 Maverick: SFT, online RL, and DPO with a multimodal curriculum

Meta describes a post-training pipeline organized as lightweight supervised fine-tuning (SFT), followed by online reinforcement learning (RL), and then a lightweight direct preference optimization (DPO) stage to address corner cases[1]. A central theme is maintaining capability across modalities while improving reasoning and conversational quality. Meta describes a curriculum strategy for mixing modalities without sacrificing single-modality performance[1].

Meta also reports that overly aggressive SFT/DPO can restrict exploration during online RL; to counter this, it describes filtering out more than half of "easy" examples (using Llama models as judges) and focusing SFT on a harder subset, followed by a multimodal online RL stage biased toward harder prompts. The announcement further describes a continuous online RL approach that alternates model updates with ongoing prompt filtering to retain medium-to-hard prompts, and a final lightweight DPO stage to improve response quality in corner cases[1].

4.2 Scout: context training and length generalization

Meta reports that Scout is both pretrained and post-trained at 256K context length, and presents this as enabling improved length generalization[1]. The release also describes evaluations focused on extreme-length regimes (e.g., retrieval-style "needle" tests and negative log-likelihood over very long code sequences), and presents Scout's long-context behavior as supported by the iRoPE design and training strategy[1].

5 Evaluation results

Meta reports benchmark results for both pretrained (base) and instruction-tuned variants, with evaluations performed on bf16 models[2]. For multimodal tasks, prior Llama baselines are often noted as not supporting multimodality (and thus are not directly comparable).

5.1 Full benchmark tables (developer-reported)

Tables 3 and 4 reproduce the full benchmark sets as presented in the model card[2].

Table 3: Pretrained (base) benchmark results reproduced from the official Llama 4 model card. All evaluations are reported on bf16 models.[2]

Category	**Benchmark**	**# Shots**	**Metric**	**L3.1 70B**	**L3.1 405B**	**L4 Scout**	**L4 Maverick**
Reasoning & Knowledge	MMLU	5	macro_avg/acc_char	79.3	85.2	79.6	85.5
Reasoning & Knowledge	MMLU-Pro	5	macro_avg/em	53.8	61.6	58.2	62.9
Reasoning & Knowledge	MATH	4	em_maj1@1	41.6	53.5	50.3	61.2
Code	MBPP	3	pass@1	66.4	74.4	67.8	77.6
Multilingual	TyDiQA	1	average/f1	29.9	34.3	31.5	31.7
Image	ChartQA	0	relaxed_accuracy	No multimodal support		83.4	85.3
Image	DocVQA	0	ANLS	No multimodal support		89.4	91.6

5.2 Interpretation notes

Several patterns stand out in the reported tables:

- **Maverick leads Scout consistently** on the reported reasoning and coding benchmarks (e.g., GPQA Diamond and LiveCodeBench), consistent with a larger expert pool (128 experts) at equal active parameter count[2].
- **Instruction-tuning closes gaps and improves multimodal tasks** substantially relative to base results on ChartQA/DocVQA (as expected for vision-language alignment)[2].
- **Long-context evaluation is presented via MTOB** with chrF scores in two translation directions, with the model card explicitly contrasting against a 128K context baseline[2].

6 Quantization and inference packaging

Meta reports that Scout can fit within a single H100 GPU with on-the-fly int4 quantization. For Maverick, Meta distributes BF16 weights as well as an FP8 quantized *instruction-tuned* variant; the FP8 weights are described as fitting on a single H100 DGX host while maintaining quality[2, 7].

Table 4: Instruction-tuned benchmark results reproduced from the official Llama 4 model card. All evaluations are reported on bf16 models.[2] MMMU Pro values are reported as the average of Standard and Vision tasks.[2]

Category	**Benchmark**	**# Shots**	**Metric**	**L3.3 70B**	**L3.1 405B**	**L4 Scout**	**L4 Maverick**
Image Reasoning	MMMU	0	accuracy	No multimodal support		69.4	73.4
Image Reasoning	MMMU Pro	0	accuracy	No multimodal support		52.2	59.6
Math & Vision	MathVista	0	accuracy	No multimodal support		70.7	73.7
Image Understanding	ChartQA	0	relaxed_accuracy	No multimodal support		88.8	90.0
Image Understanding	DocVQA (test)	0	ANLS	No multimodal support		94.4	94.4
Coding	LiveCodeBench (10/01/2024–02/01/2025)	0	pass@1	33.3	27.7	32.8	43.4
Reasoning & Knowledge	MMLU Pro	0	macro_avg/acc	68.9	73.4	74.3	80.5
Reasoning & Knowledge	GPQA Diamond	0	accuracy	50.5	49.0	57.2	69.8
Multilingual	MGSM	0	average/em	91.1	91.6	90.6	92.3
Long context	MTOB (half book) eng→kgv / kgv→eng	–	chrF	Context window is 128K		42.2/36.6	54.0/46.4
Long context	MTOB (full book) eng→kgv / kgv→eng	–	chrF	Context window is 128K		39.7/36.3	50.8/46.7

The Hugging Face usage example for Maverick references `transformers` $\geq$ 4.51.0 and uses `attn_implementation="flex_attention"` in `Llama4ForConditionalGeneration`[3]. This suggests that client-library details may matter for throughput and long-context performance, particularly in attention kernels and KV cache handling.

7 Prompt formats and chat templates

Meta publishes prompt formats and model-specific guidance for Llama 4. Reproducible evaluation and stable deployment behavior typically depend on using the correct template (system/user/assistant formatting) and generation settings as documented[11].

8 Deployment limits and provider-specific context windows

A recurring operational theme is the gap between *architectural* context length and *served* context length.

- **Cloudflare Workers AI (launch note):** Scout is highlighted as supporting up to 10M tokens architecturally, but Workers AI initially supports a context window of 131,000 tokens and planned increases[8].

- **Amazon Bedrock (serverless):** AWS reports Bedrock support for a 1M token context window for Maverick and a 3.5M token context window for Scout at the time of the announcement, with stated plans to expand Scout further[12].

- **Amazon SageMaker JumpStart (blog overview):** the JumpStart overview text describes Maverick with a 128K context window and Scout with 10M, illustrating that availability can reflect platform constraints and packaging choices rather than the maximum described in model cards[13].

Operational implication. In long-context deployments, memory devoted to KV cache grows roughly linearly with context length and batch size; this can become the dominating memory consumer at 1M+ tokens. Consequently, providers may expose a reduced context limit to preserve latency, throughput, and multi-tenant stability. Practitioners should treat "context length" as a *per-endpoint contract* rather than a fixed model constant.

9 Model governance, redistribution, and attribution

Llama 4 is distributed under a custom community license agreement rather than an OSI-approved open source license[14, 15]. Key obligations (summarized) include:

- **Redistribution requirements:** distributing the model (or derivatives) requires providing a copy of the agreement and prominently displaying "Built with Llama" in relevant product-facing materials[14].

- **Derivative model naming:** if Llama Materials or their outputs are used to create/train/fine-tune an AI model that is distributed, "Llama" must be included at the beginning of the new model name[14].

- **Attribution file:** redistributed copies must retain a specified attribution notice within a `NOTICE` text file[14].

- **Large-scale commercial threshold:** if the licensee (or affiliates) exceeds 700M monthly active users on the version release date, an additional license from Meta is required before exercising rights under the agreement[14, 6].

- **EU restriction (multimodal models):** public license/policy materials have included a restriction under which the rights granted for Llama 4 multimodal models are not granted to individuals domiciled in, or entities with a principal place of business in, the European Union (with an exception stated for end users of products/services incorporating such models)[14, 15].

The Open Source Initiative argues that Meta's Llama licenses do not satisfy the Open Source Definition and points to restrictions that, in OSI's view, limit fundamental freedoms associated with open source[15]. Users should evaluate the license text directly for their intended redistribution and deployment scenarios.

10 Safety and system-level protections

The model card materials emphasize that LLMs should be deployed as part of a broader system with guardrails; Meta references a set of system protections (e.g., Llama Guard, Prompt Guard, Code Shield) in its distribution materials[3]. Meta also released Llama Guard 4 (12B), a multimodal safety classifier trained to predict safety labels aligned to the MLCommons hazard taxonomy (with an additional category for code-interpreter abuse in certain contexts)[10].

Meta's announcement describes stress testing via adversarial dynamic probing and introduces "Generative Offensive Agent Testing" (GOAT), presented as a method to simulate multi-turn interactions by medium-skill adversaries to broaden coverage beyond traditional red-teaming[1].

The same announcement also reports changes intended to reduce refusal rates and improve response balance on debated political and social topics, including (i) a lower overall refusal rate relative to the prior generation, (ii) improved parity in refusal behavior across viewpoints, and (iii) a reduced rate of strong political lean in responses on a contentious prompt set (reported as comparable to Grok in that evaluation)[1]. These are developer-reported behavioral metrics; downstream deployments may differ depending on system prompts, safety layers, and fine-tuning choices.

10.1 Benchmarking and release-variant caveats

When citing third-party leaderboards, it is important to verify whether the evaluated system corresponds exactly to the publicly released checkpoints. For example, reporting around LMArena indicated that an "experimental chat" Maverick variant used for leaderboard submission was not identical to the public release, prompting criticism and policy changes on the benchmark side[16]. Accordingly, this manuscript treats the official bf16 model-card evaluations as the primary quantitative reference for the released artifacts[2].

11 Conclusion

Llama 4 represents a substantive shift in the Llama ecosystem toward sparse, multimodal, long-context models: Scout and Maverick combine a mixture-of-experts backbone with early-fusion vision-language processing, while release materials further describe long-context design choices for Scout (iRoPE and inference-time attention scaling) and a training program that includes pre-training, mid-training for context extension, and a post-training pipeline centered on lightweight SFT, online RL, and lightweight DPO. Meta also positions Behemoth as a large teacher model and reports codistillation into Maverick, underscoring the role of teacher–student training in achieving the released models' capability–cost profile. Developer-reported benchmark tables indicate that Maverick outperforms Scout on a range of reasoning and coding tasks, and that instruction tuning yields large gains on multimodal evaluations; however, real-world performance depends materially on prompt formats, inference kernels, and serving constraints. In deployment, the limiting factor is typically not the architectural maximum context length but the effective context window and memory budget provided by specific platforms, including KV-cache scaling and weight residency under MoE. Finally, Llama 4's custom community license and accompanying attribution and derivative naming requirements are operationally significant and should be treated as core engineering and compliance considerations alongside safety tooling and system-level mitigations.

12 Release chronology (public reporting)

The Scout and Maverick release date is reported as April 5, 2025, in partner model documentation[4]. Reuters and other outlets reported the public release and noted a preview of the Behemoth teacher model[5, 6].

A Contributors

Aaron Adcock, Aayushi Srivastava, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pande, Abhinav Pandey, Abhinav Sharma, Abhishek Kadian, Abhishek Kumawat, Adam Kelsey, Adam Stelle, Adeel Cheema, Adela Kabiljo, Adina Katz, Adithya Gangidi, Aditya Tayade, Adolfo Victoria, Adrian Samatan Alastuey, Adrien Conrath, Afroz Mohiuddin, Ahmed Sharif, Ahnaf Siddiqui, Ahuva Goldstand, Aijung Li, Aidan Boyd, Aidin Kazemi Daliri, Aisha Iqbal, Ajay Menon , Ajit Mathews, Akhil Mathur, Akshat Agarwal, Alan Schelten, Alana Shine, Alejandro Castillejo Muñoz, Aleksei Guliaev, Alex Radovic, Alex Song, Alex Vaughan, Alexander Simeonov, Alexandre Rezende, Alexandre Rezende, Alexei Baevski, Alexey Roubaud, Allen Ma, Alvin Lee, Alyssa Pereira, Aman Ahmed, Aman Shankar, Amanda Kallet, Amar Budhiraja, Ameya Khandekar, Amine Benhalloum, Amir Gershman, Amit Nagpal, Amit Zohar, Amr Sharaf, Anant Desai, Anastasia Razdaibiedina, Anca Agape, Andranik Kurghinyan, Andre Perunicic, Andrea Madotto, Andrei Darabanov, Andrés Alvarado, Andrew Brown, Andrew Cohen, Andrew Fang, Andrew Freeman, Andrew Gallagher, Andrew Gu, Andrew Prasetyo Jo, Andrew Ryan, Andrew Steffen, Andrew Wei, Andrey Rusakov, Andrii Golovei, Andy Shang, Angela Fan, Angela Fan, Angela Flewellen, Animesh Pathak, Anirudh Goyal, Ankit Ramchandani, Ankur Pai, Ankur Singh, Ankush Garg, Anlu Xing, Anna Cai, Anna Grosul, Anna Prochowska, Anna Sun, Annie Dong, Annie Franco, Anqi Hu, Anshul Chawla, Anthony Hartshorn, Antonia Sheng, Antony Thomas, Anuj Goyal, Anusha De, Anvit Bodiwala, Anvit Bodiwala, Aobo Yang, Aparajita Saraf, Apurva Samudra, Aran Mun, Arash Rahnama, Archi Mitra, Archie Sravankumar, Archit Gupta, Aria Haghighi, Ariel Stolerman, Arkabandhu Chowdhury, Arnab Choudhury, Artem Korenev, Arthur Guo, Arthur Hinsvark, Arun Mallya, Arvind Neelakantan, Arya Talebzadeh, Ashish Shah, Ashmitha Jeevaraj Shetty, Ashwin Bharambe, Asif Islam, Aston Zhang, Austen Gregerson, Avi Lewis, Aya Ibrahim, Ayaz Minhas, Ayelet Dahan, Ayelet Regev Dabah, Bangsheng Tang, Bar Ulman, Bardiya Sadeghi, Bartosz Jedrzejewski, Barys Skarabahaty, Beibei Zhu, Beibin Li, Ben Bharier, Benjamin Leonhardi, Benjamin Muller, Bennett Plessala, Bernie Huang, Beth Loyd, Bhargavi Paranjape, Bhavik Sheth, Bill Bonner, Bill Holland, Bill Wang, Bingzhe Liu, Binh Tang, Bo Liu, Bo Wu, Boduo Li, Bokai Yu, Bor-Chun Chen, Boris Araya, Boris Vidolov, Botao Chen, Boya Peng, Boyu Ni, Bradley Davis, Bram Wasti, Brandon Adams, Brandon Taylor, Brandon Wu, Brant Swidler, Brian Chiang, Brian Clerkin, Brian Fuller, Brooks Cutter, Bruno Novais, Bryan Gmyrek, Bysshe Easton, Cait Campos, Canaan Case, Carl Chengyan Fu, Carly Burton, Caro Diaz, Catherine Cole, Ce Liu, Cedric Fougerat, Cen Peng, Cen Peng, Cen Zhao, Changhan Wang, Changkyu Kim, Chantal Shaib, Chao Zhou, Charlotte Caucheteux, Chau Nguyen, Chawin Sitawarin, Chaya Nayak, Chelsea Asher, Chen Fan, Chen Zhu, Cheng Cheng, Cheng Zhang, Chenguang Zhu, Chengxiong Ruan, Chengzhu Yu, Chenheli Hua, Chenxi Whitehouse, Cheryl Holloway, Ching-Hsiang Chu, Ching-Yao Chuang, Chinmay Karande, Chirag Nagpal, Chloé Bakalar, Chloe Bi , Chris Cai, Chris Marra, Chris McConnell, Chris Thi, Chris Tindal, Chris Waterson, Christian Deverall, Christian Fuegen, Christian Keller, Christine Cheng, Christine Jou, Christine Smith, Christine Wang, Christoph Feichtenhofer, Christophe Touret, Christopher Luc, Christy Sauper, Chuanhao Zhuge, Chun-Yi Sung, Chunqiang Tang, Chunyang Wu, Clara Siegel, Cody Heale, Cody Wilbourn, Colin White, Congying Xia, Corinne Wong, Cornel Rat, Cristian Canton Ferrer, Cyrille Habis, Cyrus Nikolaidis, D Lohachov, Da Ju, Dalton Flanagan, Damien Allonsius, Damon Civin, Dan Johnson, Daniel Bolya, Daniel Francisco, Daniel Fried, Daniel Hawthorne, Daniel Haziza, Daniel Ho, Daniel Kreymer, Daniel Li, Daniel Machlab, Daniel McKinnon, Daniel Obenshain, Daniel Rodriguez, Daniel Song, Daniel Tse, Danielle Pintz, Danny Livshits, Daryl James Rodrigo, Dat Huynh, Daulet Askarov, David Brandfonbrener, David Esiobu, David Kant, David Levin, David Renardy, David Soofian, David Stevens, David Xu, David Zhang, Deep Shah, Delia David, Demi Douglas, Denis Boyda, Desh Raj, Devamanyu Hazarika, Dheeraj Mekala, Dhruv Choudhary, Dhruv

Mahajan, Di Jin, Didac Suris Coll-Vinent, Didem Foss, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, DiJia Su, Dilip Madathil, Dinesh Govindasamy, Dinesh Yeduguru, Dmitry Vengertsev, Dong He, Dong Li, Dong Wang, Dongzhuo Li, Duc Le, Dunant Hin, Dustin Holland , Duy Nguyen, Duy Nguyen, Ed Dowling, Eden Litt, Egor Lakomkin, Ehab AlBadawy, Ehsan K. Ardestani, Elad Eckstein, Elahe Dabir, Elaine Montgomery, Elina Lobanova, Elior Abramoviz, Eliot Hedeman, Elissa Li, Elizabeth Hilbert, Ellen Xiaoqing Tan, Elliot Yun, Elodie Stener, Emilian Stoimenov, Emilien Garreau, Emily Dinan, Emily Hahn, Emily Wood, Emma Li, Emmanuel Ademuwagun, Emrah Seker, Eric Alamillo, Eric Gan, Eric Han, Eric Huang, Eric Michael Smith, Eric-Tuan Le, Ernie Chang, Eryk Helenowski, Eslam Elnikety, Esteban Arcaute, Ethan Myers, Eugene Nho, Eugene Poliukhovych, Evan Dunbar, Evgeniy Litvinenko, Evrim Altıntaş, Eyal Hochman, Eyal Shtrauch, Fabian Mastenbroek, Faiza Zeb, Faizan Ahmad, Farhad Farahbakhshian, Fei Kou, Fei Sun, Feiyu Chen, Felix Chung, Feng Tian, Feng Xu, Filip Radenovic, Filippos Kokkinos, Francesco Barbieri, Francesco Caggioni, Francisco Esparza, Francisco Guzmán, Frank Kanayet, Frank Seide, Frank Zhang, Fred Lewis, Freda Huang, Fulton Wang, Gabriel Synnaeve, Gabriela Jacques-Silva, Gabriella Schwarz, Gaganjit Ghardhora, Gal Elfer, Garrett Dickson, Gaurav Chaurasia, Gautam Sewani, Geet Shingi, Gefei Zuo, Geonhwa Jeong, George Puthanpurackal, Georgia Swee, Gerard Moreno-Torres Bertran, Gil Keren, Gina Ling, Gjergji Stasa, Gobinda Saha, Gor Safran , Gordy French, Goutham Rajendran, Govind Thattai, Grace Cineas, Graeme Nail, Greg Fletcher, Grégoire Mialon, Griffin Adams, Grigory Sizov, Guan Pang, Hady Elsahar, Hai Dang Tran, Hailey Nguyen, Haiping Wu, Hakan Inan, Hamid Eghbalzadeh, Han Fang, Han Zou, Hannah Doyle, Hannah Korevaar, Hannah Wang, Hannah Werbel, Hanwen Zha, Hany Morsy, Hao Ma, Haoci Zhang, Haonan Sun, Haozhu Wang, Hardik Shah, Haroun Habeeb, Harrison Rudolph, Harsh Gupta, Harsh Poddar, Harshil Parikh, Hejia Zhang, Heming Wang, Hengduo Li, Himanshu Sharma, Hoang Phi Nguyen, Hongbo Zhang, Honghao Qiu, Hongjiang Lv, Hongli Xu, Hongyuan Zhan, Hossein Hamooni, Howard Huang, Hu Xu, Hugo Laurençon, Hugo Touvron, Hung Dinh, Hunter Goldman, Hussein Mehanna, Huy Nguyen, Hweimi Tsuo, Ian Graves, Ian Yu, Ibrahim Damlaj, Idan Cohen, Igor Tufanov, Ilan Goldenstein, Ilias Leontiadis, Iliyan Zarov, Imad Ahmed, Innocent Djiofack, Iosif Spulber, Irina-Elena Veliche, Isabella Ramos, Ishan Misra, Itai Gal, Ivan Evtimov, Ivan Evtimov, Ivan Obraztsov, Jack Wu, Jacqueline Romero Vertino, Jaemo Koo, Jaewon Lee, Jake Jung, Jake Weissman, James Beldock, James Crnkovich, James Grinage, James Hongyi Zeng, James Kohli, James Tian, Jamie Cahill, Jan Geffert, Jan Seidel, Jan Seidel, Janey Tracey, Jang Hyun Cho , Janice Wei, Jarrod Kahn, Jasmyn Howell, Jason Long Vu, Jason Park, Jason Yan, Jason Yip, Jay Li, Jay Mahadeokar, Jaya Bharath R Goluguri, Jayasi Mehar, Jean-Baptiste Gaya, Jeet Shah, Jeff Hanson, Jeff Marcus, Jeff Walsh, Jeff Yang, Jelmer van der Linde, Jemma Fan, Jennifer Chan, Jenny Zhen, Jenya Lee, Jeremy Fu, Jeremy Reizenstein, Jeremy Teboul, Jesse He, Jessica Zhong, Ji Hou, Ji Yang, Jia Ding, Jiabo Hu, Jiacheng Zhu, Jiadong Guo, Jialiang Wang, Jialin Ouyang, Jianfeng Chi, Jianyu Huang, Jianyun Zhao, Jiaowen Yang, Jiatong Zhou, Jiawei Zhao , Jiawen Liu, Jie Wang, Jie You, Jiecao Yu, Jillian Schwiep, Jilong Wu, Jing Huang, Jing Li, Jing Yu Koh, Jing Zhang, Jingxiang Chen, Jingyi Yang, Jingyue Shen, Jinho Hwang, Jinxi Guo, Jiwan Khatiwada, Joanna Bitton, Joe Li, Joe Quanaim, Joel Beales, Johan Schuijt, John Chang, John Quan, Johnnie Chan, Jon Shepard, Jona Harris, Jonah Rubin, Jonathan Janzen, Jonathan Kaldor, Jorge Lopez Silva, Jose Leitao, Joseph Greer, Joseph Moon, Joseph Rocca, Joseph Tighe, Josh Fromm, Joshua Deng, Joshua Fernandes, Joshua Saxe, Joyce Zheng, Juan Pino, Julien Prigent, Jun Chen , Junjiao Tian, Junjie Qi, Junjie Wang, Junteng Jia, Kade Baker, Kai Londenberg, Kai Wang, Kainan Peng, Kaiyan Peng, Kaiyue Yang, Kalyan Vasudev Alwala, Kam Hou Yu, Kanika Narang, Karan Chadha, Karan Sikka, Karen Zhang, Karina Schuberts, Karishma Mandyam, Karthik Abinav Sankararaman, Karthik Padthe, Karthik Prasad, Karthik Sivakumar, Kartikeya Upasani, Kate Plawiak, Kate Saenko, Kateřina Žmolíková, Kathryn Stadler, Kathy Matosich, Katie Doulgass, Kaveh Hassani, Kay Ji, Ke Li, Kenneth Heafield, Kenny Yu, Keqian Li, Kevin Chih-Yao Ma, Kevin Hannan, Keyu Man, Kezhen Chen, Khalid El-Arini, Khrystyna Hutsulyak, Kieran Nash, Kiran Jagadeesh, Kody Bartelt, Konstantin Topaloglou-Mundy, Konstantinos Chatziioannou, Konstantinos Karanasos, Konstantinos Vougioukas, Kostas Tsiampouris, Kristen Hamill, Kristy Choi, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kun Huang, Kunal Bhalla, Kunal Chawla, Kunpeng Li, Kushal Lakhotia, Kyle Monk, Lakshya Garg, Lalit Chourey, Lars Hamre, Laura Gustafson, Lauren Deason, Laurence Rouesnel, Laurens van der Maaten, Lavender A, Lawrence Chen, Lawrence Jang, Leandro Silva, Leda Sari, Lee Hetherington, Lei Zhang, Leiyu Zhao, Lele Chen, Leo Chenghui Li, Leon Yang, Leon Zhan, Levi Corallo, Liang Tan, Licheng Yu, Lijuan Liu, Lilach Mor, Lincoln Lin, Linfeng Li, Lisa Titus, Liz Jenkins, Lovish Madaan, Lu Fang, Lu Yuan, Lucas Nava, Lucas Pasqualin, Lucas Switzer, Lucia Fang, Lucy Sun, Luka Tadic, Lukas Blecher, Lukas Landzaat, Luxin Zhang, Madhavi Rao, Madian Khabsa, Mahalia Miller, Mahendra Kariya, Mahesh Pasupuleti, Mahi Luthra, Manaal Faruqui, Manav Avlani, Manchen Wang, Mannat Singh, Manohar Paluri, Manoj Chakkaravarthy, Manoj Nair, Maquelle Tiffany, Marcin Pawlowski, Marcus Wu, Maria Lomeli, Mario Consuegra, Marion Boiteux, Marios Andreas Galanis, Marshall Chen, Martin Gleize, Maryam Fazel-Zarandi , Matan Hasson, Mathew Oldham, Mathieu Rita, Matt Dordal, Matt Setzler, Matt Staats, Matt Staats, Matt Wilde, Matthew Clark, Matthew Grange, Matthew Lennie, Matthew Schmohl, Max Raphael, Maxim Naumov, Maxim Samoylov, Maxime Lecanu, Maya Pavlova, Md Taha Bin Jawaid, Meghan Keneally, Melanie Kambadur, Meng Zhang, Mengchen Liu, Mengdi Lin, Mengjiao Wang, Mervyn Abraham, Miao Liu, Michael Au-Yeung, Michael Feldergraf, Michael Man, Michael Matheny, Michael Suo, Michael Tontchev, Michel Meyer, Michelle Ma, Mihir Patel, Mihir Sanjay Kale, Mik Vyatskov, Mikayla Alexander, Mike Andersland, Mike Clark, Mike Lewis, Mike Li, Mike Macey, Mike Macey, Mike Seltzer, Mikel Jimenez Fernandez, Mikhail Antonov, Mikhail Plekhanov, Milan Zhou, Min Si, Ming Qiao , Mingbo Ma, Mingjun Zhang, Mingyi Liang, Miquel Jubert Hermoso, Mirac Suzgun, Mirjam Skarica, Mitesh Kumar Singh, Mohammad Kabbani, Mohammad Rastegari, Mona Sarantakos, Monica Sim, Monika Gangapuram, Mor Moshe, Morrie Doulaty, Morvarid Metanat, Moya Chen, Mrinal Kumar, Munish Bansal, Murali Ramarao, Na Li, Nadav Azaria, Nahiyan Malik, Naman Goyal, Nancy Vargas Balderas, Nanshu Wang, Naoyuki Kanda, Natalia Gimelshein, Natalia Neverova, Nathan Aclander, Natt Sithiviraporn, Navneet Madhu Kumar, Ned Newton, Neeraj Bahl, Negar Ghorbani, Neil Patel, Neta-lee Golan, Nicholas Longenbaugh, Nick Egebo, Nikhil Johri, Nikhil Mehta, Nikhil Naik, Niko Moritz, Nikolay Bashlykov, Nikolay Bogoychev, Nikolay Pavlovich Laptev, Niladri Chatterji, Nile Jones, Nimish Shah, Ning Dong, Ning Li, Ning Li, Ning Zhang, Nishant Yadav, Noam Paz, Norman Cheng, Norman Cheng, Olaoluwa Adesanya, Oleg Repin, Oleksandr Maksymets, Omkar Salpekar, Omri Harosh, Onkar Pednekar, Onur Çelebi, Oran Gafni, Oren Edinger, Osama Hanna, Owais Khan Mohammed, Ozlem Kalinli, Paden Tomasello, Pankaj Singh, Paola Quevedo, Parag Jain, Paria Rashidinejad, Parker Tooley, Parth Parekh, Parth Thakkar, Parvin Taheri, Pasan Hapuarachchi, Pascal Kesseli, Patrick Alrassy, Paulo de Rezende Pinatti, Pavan Balaji, Pawan Sisodiya, Pedro Jose Ferreira Moreira, Pedro Rittner, Pedro Valenzuela, Peize Sun, Peizhao Zhang, Peng-Jen Chen, Pengchao Wang, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Carras, Peter Ney, Peter Weng, Petru Dumea, Phil Hayes, Philip Woods, Pierre Andrews, Pierre Ménard, Ping-Hao Wu, Pingchuan Liu, Piotr Dollar, Plamen Dzhelepov, Polina Zvyagina, Posten A, Prabhav Agrawal, Pradhapan Rajendran, Pradyot Prakash, Prajjwal Bhargava, Pramono, Pranay Shah, Pranshu Dave, Prash Jain, Pratik Dubal, Praveen Gollakota, Praveen Krishnan, Pritish Yuvraj, Projjal Ghosh, Punit Singh Koura, Puxin Xu, Qi Qi, Qi Zhou, Qian Guan, Qian Sun, Qiang Liu, Qing He, Qinqing Zheng, Qirui Yang, Qizhen Guo, Quanzeng You, Quentin Carbonneaux, Quentin Carbonneaux, Quentin Duval, Quintin Fettes, Rachad Alao, Rachel Batish, Rachel Guo, Rachel Rodriguez, Radhika Bhargava, Rafael Asuncion, Raghotham Murthy, Rahul Dutta, Rahul Jha, Rahul Kindi, Rahul Mitra, Raj Ganapathy, Raj Shah, Rajarshi Das, Rajat Shrivastava, Rajesh

Nishtala, Ramakant Shankar, Raman Shukhau, Ramon Calderer, Rangaprabhu Parthasarathy, Ranjan Subramanian, Raphael Bensadoun, Rares Bostan, Rashnil Chaturvedi, Ravi Agrawal, Ray Gao, Raymond Li, Rebecca Kogen, Ricardo Juan Palma Duran, Ricardo Silveira Cabral, Richard Lee, Richard Yuanzhe Pang, Riddhish Bhalodia, Riham Mansour, Rishabh Singh, Rishi Godugu, Ritun Patney, Rob Boyle, Robbie Goldfarb, Robert Caldwell, Robert Kuo, Roberta Raileanu, Robin Battey, Robin Sharma, Rochit Sapra, Rocky Wang, Rodolfo Granata, Rodrigo De Castro, Rodrigo Paim, Rohan Maheshwari, Rohan Varma, Rohit Girdhar, Rohit Patel, Roshan Sumbaly, Roy Sheaffer, Ruan Silva, Ruben Rodriguez Buchillon, Rui Hou, Ruiming Xie, Ruslan Mavlyutov, Ruslan Semenov, Rustam Dinov, Ruxiao Bao, Ryan Fox, Ryan Kilpatrick, Ryan Kwan, Ryan Lim, Ryan Smith, Saaketh Narayan, Sabrina Qiao, Sachin Mehta, Sachin Siby, Sagar Jain, Saghar Hosseini, Sagie Gur-Ari, Sahana Chennabasappa, Sahin Geyik, Sai Jayesh Bondu, Sai Mounika Chowdhary Nekkalapudi, Saif Hasan, Saisuke Okabayashi, Saketh Rambhatla, Salil Sawhney, Sam Dunster, Sam Zhao, Saman Keon, Samaneh Azadi, Sameet Sapra, Samuel Dooley, Samyak Datta, Sandeep Parab, Sang Michael Xie, Sanjay Singh, Sanyuan Chen, Sara Behn, Sara Khodeir, Sarah Shirazyan, Sargun Dhillon, Sarunya Pumma, Sasha Sidorov, Saskia Adaime, Saurabh Khanna, Sayem Wani, Scott Brenton, Sean Bell, Sean Kelly, Sean Koger, Sean Nunley, Sean Perry, Sebastian Caicedo, Sebastian Dahlgren, Sebastian Ruder, Seiji Yamamoto, Selam Mehretu, Selvan Sunitha Ravi, Sen Lyu, Senthil Chellapan, Serafeim Mellos, Sergey Edunov, Sergey Royt, Shaina Cohen, Shangfu Peng, Shannon Adams, Shaoliang Nie, Sharadh Ramaswamy, Sharan Narang, Shashank Pisupati, Shashi Gandham, Shaun Lim, Shaun Lindsay, Sheena Artrip, Shelly Sheynin, Shen Yan, Sheng Feng, Sheng Shen, Shengbao Zheng, Shenghao Lin, Shengjie Bi, Shengxin Cindy Zha, Shengye Wan, Shengyi Qian, Shengyong Cai, Shengzhi Shao, Shervin Shahidi, Shikai Li, Shimon Bernholtz, Shiqi Wang, Shishir G. Patil, Shiv Verma, Shiva Shankar P, Shiyang Chen, Sho Yaida, Shoubhik Debnath, Shreyas Siravara, Shruti Bhosale, Shuang Ma, Shun Zhang, Shuo Tang, Shuqiang Zhang, Shuyan Zhou, Sicong Che, Sidd Srinivisan, Siddharth Bhattacharya, Siddharth Patki, Sijia Chen, Sili Chen, Simon Vandenhende, Simone Merello, Sinong Wang, Sivan Barzily, Sixian Yi, Siyu Lin, SK Bong, Sky Yin, Sneha Agarwal, Sneha Agarwal, Soerian Lieve, Soji Sajuyigbe, Song Jiang, Songlin Li, Sonia Kim, Sopan Khosla, Soumi Maiti, Spencer Whitman, Sravya Popuri, Sreen Tallam, Srinivas Vaidyanathan, Srinivas Vaidyanathan, Sten Sootla, Stephane Collot, Stephanie Ding, Stephen Chen, Steven Cai, Suchin Gururangan, Sudarshan Govindaprasad, Sue Young, Suganthi Dewakar, Sujan Kumar Gonugondla, Sujeet Bhandari, Suman Gumudavelli, Suman Gumudavelli, Sumit Gupta, Summer Deng, Sungmin Cho, Suresh Ganapathy, Surjyendu Dhal, Susan Fedynak, Susana Contrera, Suyoun Kim, Sylvestre Rebuffi, Takshak Chahande, Tamar Herman, Tan Li, Tao Xu, Tara Fowler, Tarek Sheasha, Tarun Anand, Tarun Kalluri, Tarun Singh, Tatiana Shavrina, Ted Li, Teja Rao, Tejas Patil, Teng Li, Thach Bui, Thai Quach, Thamer Alharbash, Thanh Vinh Vo, Thawan Kooburat, Thilo Koehler, Thomas Georgiou, Thomas Scialom, Tian Ye, Tianhe Li, Tianjun Zhang, Tianyu Li, Tijmen Blankevoort, Timon Willi, Timothy Chou, Timothy Leung, TJ Lee, Todor Mihaylov, Tom Heatwole, Tong Xiao, Tony Cao, Tony Lee, Trang Le, Tristan Rice, Tsz Kei Serena Chan, Tuan Tran, Tudor Tiplea, Tyler Baumgartner, Uday Savagaonkar, Ujjwal Karn, Ulises Martinez Araiza, Umar Farooq, Uriel Cohen, Usman Sharif , Utkarsh Murarka, Van Phung, Varun Joginpalli, Varun Saravagi, Vasu Sharma, Vasudha Viswamurthy, Vedanuj Goswami, Vedika Seth, Venkat Ramesh, Venkat Ramesh, Vibhor Gupta, Victoria Montanez, Vidhya Natarajan, Vidya Sarma, Vignesh Ramanathan, Viktor Kerkez, Vinay Rao, Vincent Gonguet, Vincent Mauge, Virginie Do, Vish Vogeti, Vishrav Chaudhary, Viswesh Sankaran, Vítor Albiero, Vivek Miglani, Vivek Pai, Vlad Cojanu, Vlad Shubin, Vlad Tiberiu Mihailescu, Vladan Petrovic, Vladimir Ivanov, Vladislav Vorotilov, Vrushali Bhutada, Wai I Ng, Wei Cheng, Wei Sun, Wei Tu, Wei Wei, Wei Zhou, Wei-Ning Hsu, Weiwei Chu, Weizhe Yuan, Wenchen Wang, Wenjun Zhao, Wenwen Jiang, Wenyin Fu, Wenzhe Jiang, Whitney Meers, Will Constable, Will Wang, William R. Wong, Xavier Martinet, Xi Victoria Lin, Xi Yan, Xi Yin, Xian Li, Xianfeng Rui, Xianjun Yang, Xiaocheng Tang, Xiaodong Wang, Xiaofang Wang, Xiaolan Wang, Xiaoliang Dai, Xiaoliang Peng, Xiaopeng Li, Xiaozhu Meng, Xibei Zhang, Xide Xia, Xin Jin, xinbo Gao, Xinfeng Xie, Xingyi Zhou, Xu Ma, Xuan Ju, Xuanyi Zhao, Xubo Liu, Xuchao Jia, Xuedong Zhang, Xuefei Cao, Xuewei Wang, Xuewei Wu, Xunnan Xu, Xutai Ma, Xuyang Wang, Yan Cui, Yang Chen, Yang Li, Yang Shu, Yang Xia, Yanjun Chen, Yanjun Zhou, Yash Mehta, Yash Patel, Yash Tekena, Yashesh Gaur, Yasmine Babaei , Yaxuan Zhou, Ye Hu, Ye Qi, Yejin Lee, Yeming Wen, Yen-Cheng Liu, Yexin Bruce Wu, Yi Pan, Yi Yang, Yi-Hui Lin, Yifan Wang, Yifan Wu, Yifan Yang, Yifei Huang, Yiftah Ben Aharon, Yilin Yang, Yiling You, Ying Xu, Ying Zhang, Yingquan Yuan, Yingru Liu, Yingyi Ma, Yining Yang, Yiting Lu, Yonatan Komornik, Yongjie Lin, Yoni Goyhman, Yossi Moran Mamo, Youngjin Nam, Yu Wang, Yu Lu, Yu Zhao, Yu-Ho Hsieh, Yu-Jung Lo, Yuandong Tian, Yuanhan Zhang, Yuanhao Xiong , Yuanshun Yao, Yuchen Hao, Yuchen Zhang, Yuchuan Li, Yue Cao, Yue Yu, Yue Zhao, Yuhan Guo, Yuhao Wang, Yuheng Huang, Yujie Lu, Yujun Shi, Yulun Wang, Yun He, Yun Wang, Yundi Qian, Yunfan Wang, Yunhao Tang, Yuning Mao, Yunlu Li, Yuqi Dai, Yuriy Hulovatyy, Yushi Hu, Yuxuan Sun, Zach Rait, Zach Wentz, Zacharie Delpierre Coudert, Zachary Collins, Zahra Hankir, Zecheng He, Zeeshan Ahmed, Zeeshan Ahmed, Zef RosnBrick, Zhan Shu, Zhanna Rohalska, Zhaoduo Wen, Zhe Liu, Zhe Liu, Zhen Qiao, Zhenggang Xu, Zhengwen Zhou, Zhengxing Chen, Zhenyu Tang, Zhichen Wu, Zhicheng Ouyang, Zhihong Lei, Zhipeng Hong, Zhiping Xiu, Zhiwei Zhao, Zhong Meng, Zhou Jin, Zhouhao Zeng, Zichang Liu, Zihang Meng, Zihuan Qiao, Zinnia Zheng, Zixi Qi, Ziyi Luo, Zoe Foulkes Birkhead, Zoey Sun, Zohar Achdut.

References

[1] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Meta AI Blog, 2025. Published 2025-04-05; accessed 2026-01-10.

[2] Meta. meta-llama/llama-4-scout-17b-16e: Model card and benchmarks. Hugging Face model card, 2025. Accessed 2026-01-10.

[3] Meta. meta-llama/llama-4-maverick-17b-128e-instruct: Model card and license excerpt. Hugging Face model card, 2025. Accessed 2026-01-10.

[4] NVIDIA. llama-4-maverick-17b-128e-instruct: Nvidia model card. NVIDIA Build / NIM model card, 2025. Accessed 2026-01-10.

[5] Reuters. Meta releases new ai model llama 4. Reuters, 2025. Published 2025-04-05; accessed 2026-01-10.

[6] Wes Davis. Meta releases two llama 4 ai models. The Verge, 2025. Published 2025-04-05; accessed 2026-01-10.

[7] Meta. meta-llama/llama-4-maverick-17b-128e-instruct-fp8: Model card. Hugging Face model card, 2025. Accessed 2026-01-10.

[8] Cloudflare. Meta's llama 4 is now available on workers ai. Cloudflare Blog, 2025. Published 2025-04-06; accessed 2026-01-10.

[9] Meta. meta-llama/llama-4-scout-17b-16e-instruct: Model card. Hugging Face model card, 2025. Accessed 2026-01-10.

[10] Meta. meta-llama/llama-guard-4-12b: Model card. Hugging Face model card, 2025. Accessed 2026-01-10.

[11] Meta. Llama 4: Model cards and prompt formats. Llama documentation, 2025. Accessed 2026-01-10.

[12] Danilo Poccia. Llama 4 models from meta now available in amazon bedrock serverless. AWS News Blog, 2025. Published 2025-04-28; accessed 2026-01-10.

[13] Marco Punio et al. Llama 4 family of models from meta are now available in sagemaker jumpstart. AWS Machine Learning Blog, 2025. Published 2025-04-07; accessed 2026-01-10.

[14] Meta. Llama 4 community license agreement. License text, 2025. Effective 2025-04-05; accessed 2026-01-10.

[15] Jordan Maris. Meta's llama license is still not open source. Open Source Initiative Blog, 2025. Published 2025-02-18; accessed 2026-01-10.

[16] The Verge. Meta got caught gaming ai benchmarks. The Verge, 2025. Published 2025-04; accessed 2026-01-10.