NOW AVAILABLE The 512GB Mac Studio is on sale now, shipping February to March. See every configuration.

THE AI TOOLCHAIN

NO. 012

Head

THE WEEK'S TOP PICKS

PUBLISHED NOVEMBER 6, 2026 · FRIDAYS

// SHARE THIS WEEK ▽ Top 5 ▽ Top Pick

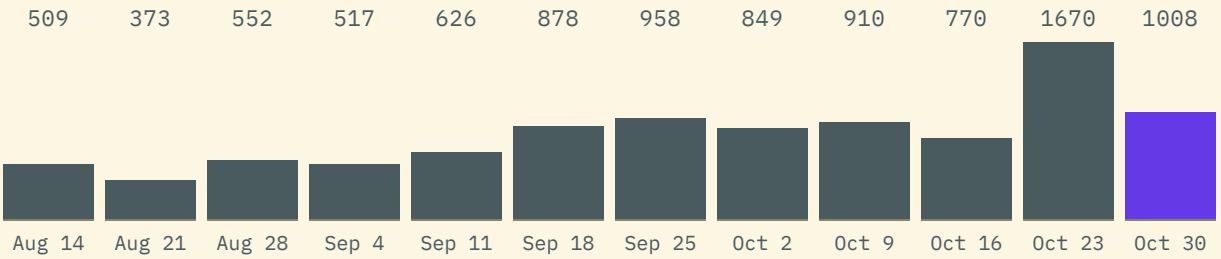
The best tooling updates that shipped this week.

The week in numbers · Oct 30–Nov 5

1,008	351	5	688
changes we read →	tools shipped →	product launches →	breaking changes →

Changes per week, last 12 weeks · Oct 30–Nov 5

Changes we read in that week's dailies, including docs and API changes. A gap means no daily was published.



▸ // HOW THESE PICKS ARE MADE SHOW

// BENCHMARKS MOVED FOR TOOLS WE TRACK HIDE

Mistral AI FrontierMath (tiers 1-3) Mistral Large 4 high

vim ? help

46.3% −12.3 pts 2026-11-02 · Epoch run ↗

Mistral AI SimpleQA Verified Mistral Large 4 high

22.4% +2.4 pts 2026-11-02 · Epoch run ↗

Mistral AI GPQA Diamond Mistral Large 4 high

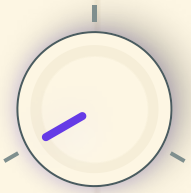
90.2% +2.1 pts 2026-11-02 · Epoch run ↗

Scores from Epoch AI Benchmarking Hub (CC BY 4.0), evaluated 2026-10-30 to 2026-11-06.

VIEW

ISSUE VIEW

scroll



SCROLL | EXPAND IN PLACE | PREVIEW PANE

INFORMATION DENSITY

everything



BRIEFED | EVERYTHING

Every change in this week's picks, in full.

all 52 changes

Do you prefer **scroll**? 👍 👎

show prereleases 1

This week's 5

01 **SGLang** Feature · Inference Serving & Optimization

02 **vLLM** Feature · Inference Serving & Optimization

03 **Liquid AI** Launch · Open-Weight Models

04 **Arcade** Feature · AI Agent Frameworks

05 **Perceptron** Launch · Frontier Models

The picks in full

01 SGLang v0.5.23

Inference Serving & Optimization

SGLang is a high-performance serving framework for large language models and multimodal models.

// WHY IT MADE THE LIST — DEPTH 5/5 — IMPACT 5/5 — NOVELTY 4/5

Paged Experts now runs mixture-of-experts models larger than GPU memory on a single consumer GPU, up to 30x faster than CPU offload — bringing oversized MoE models within reach of one card; the release also adds a LoRA engine for FP8/NVFP4 quantized MoE and Google TPU7x serving.

> SGLang's v0.5.23 release adds Google TPU7x support, a Paged Experts feature for running oversized MoE models on a single consumer GPU, a new LoRA engine for quantized MoE models, and low-latency megakernel/monokernel plugins for NVIDIA and AMD hardware, alongside broad speculative-decoding, caching and throughput improvements.

└──▷ WHAT SHIPPED · 10 FEATURES most completely described first ? what's the number?

01 Google TPU7x support via SGLang-TorchTPU NEW

90

SGLang now runs on Google TPU7x through the `SGLang-TorchTPU` PyTorch-native plugin, distributed via the `lmsysorg/sglang:v0.1.0-tpu` Docker image, with tensor and expert parallelism across hosts.

Pull the TPU7x Docker image to serve an LLM on Google Ironwood TPU hardware with SGLang-TorchTPU.

```
$ docker pull lmsysorg/sglang:v0.1.0-tpu
```

— Names plugin, exact Docker image, and runnable pull command v0.5.23

02 Low-latency megakernel/monokernel decode plugins NEW

85

New opt-in NVIDIA megakernel (`sgl-nvidia-mega`) and AMD monokernel (`sgl-amd-mono`) plugins enable single-user decode up to 44% faster on GB300 and 75% faster on MI350X for DeepSeek-V4.1-Flash.

— Named flags, hardware targets and exact speedup numbers given v0.5.23

03 Prefill and decode throughput improvements IMPROVED

68

`DeepSeek-V4.1` on GB300 now delivers up to 1.5x more prefill throughput under load and about 28% lower first-token latency on 32K to 64K prompts, while MoE models such as `DeepSeek-V4.1-Flash`, `GLM-5.3-Flash`, and `gpt-oss` reach the first token 30% to 58% sooner on 8K prompts on H200.

— Precise benchmarks and models named, but no reproduction steps v0.5.23

04 **Paged Experts for oversized MoE models** NEW 60

The opt-in Paged Experts feature lets MoE models larger than GPU memory run on a single consumer GPU, up to 30x faster than CPU offload.

— Mechanism and speedup given, but no exact flag or command v0.5.23

05 **Speculative decoding engine enhancements** NEW 60

Adds support for seeding rejection-sampling draft proposals for deterministic inference in speculative decoding, out-of-tree platform hooks so external platforms can provide their own speculative worker and EAGLE draft graph runners, and CUDA attention DP for DFlash and DFlash2 speculative decoding.

— Names three distinct speculative-decoding additions without commands v0.5.23

THINNER COVERAGE BELOW

06 **LoRA engine for quantized MoE models** NEW 52

A new opt-in LoRA engine serves LoRA adapters on FP8 and NVFP4 quantized MoE models.

— States capability and quantization formats but no command v0.5.23

07 **HiCache L3 storage for MLA prefix caches** NEW 50

DeepSeek-style MLA model prefix caches can now be kept in HiCache file-based L3 storage when running decode context parallelism.

— Names mechanism and storage tier but no config example v0.5.23

08 **Removed speculative-skip-dp-mlp-sync flag** BREAKING 45

The `--speculative-skip-dp-mlp-sync` flag has been removed.

— Exact flag named, clear migration action for users relying on it v0.5.23

09 **Ten new supported models** NEW 40

SGLang now supports Muse Spark Open, Beam-1.0, MiniMax-M3.1-Flash, MiniCPM-V 4.7, Cosmos3-Nano-Sim, FastH3 Omni, FLUX.3 Image Klein, HY-Nomas, and MOSS-Image.

— Lists named models but no usage instructions v0.5.23

10 **MSA-v1 sparse attention backend and NVFP4 KV pool** NEW

35

Adds the MSA-v1 sparse attention backend and an NVFP4 KV pool.

— Bare feature names with no further mechanism or usage detail v0.5.23

└──▷ **BREAKING ON UPGRADE**

! The `--speculative-skip-dp-mlp-sync` flag has been removed.

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

02 vLLM

announced · Nov 4, 2026

Inference Serving & Optimization

A high-throughput and memory-efficient inference and serving engine for LLMs

// WHY IT MADE THE LIST

DEPTH 5/5 IMPACT 5/5 NOVELTY 4/5

Day-0 serving for Kimi K3.1 Pro — including its Dedicated Sparse Attention layers and token-sharded prefill reaching up to 1.69x faster time-to-first-token at 340K tokens — plus same-day Mistral Large 4 support, making two trillion-scale models serveable across NVIDIA, AMD, TPU, and PPU hardware.

> vLLM v0.33.0rc3 adds a new `/v1/decisions` endpoint, token embeddings with late chunking, several new speculative-decoding techniques (draft `lm_head` quantization, FLy, XPress, MonoKernel), and an AWS EFA-optimized Docker image, while removing Medusa/MLP speculative decoding, the legacy Ray compiled-graph executor, and CUDA 12.9 build artifacts.

└──▷ **WHAT SHIPPED · 0 FEATURES** most completely described first ? what's the number?

THINNER COVERAGE BELOW

NEW SUPPORTED PLATFORM Kimi K3.1 Pro model support

NEW INTEGRATION Mistral Large 4 model support (FP8 and NVFP4 checkpoints)

└──▷ **BREAKING ON UPGRADE**

- ! Medusa and the MLP speculator are removed; any configuration referencing them will no longer work.
- ! The legacy Ray compiled-graph executor is removed; deployments using it must migrate to the current executor.
- ! `--mamba-cache-mode` is deprecated.
- ! Requests sent while generation is paused now receive HTTP 503 instead of being queued.
- ! The NIXL connector now requires NIXL 1.5 or newer; older NIXL installs will fail.
- ! LM Format Enforcer is deprecated.
- ! `--ir-op-priority` is deprecated.
- ! CUDA 12.9 release wheels and `-cu129` / `-cu129-ubuntu2404` images are no longer published; use the CUDA 13.0 artifacts instead.

> announced · Nov 4, 2026
>
> vLLM adds Kimi K3.1 Pro support with DSA sparse attention, token-sharded prefill, DCP with exact top-k merge, mixed KV quantization, and zero-mismatch RL.

NEW SUPPORTED PLATFORM Serve Kimi K3.1 Pro with `vllm serve moonshotai/Kimi-K3.1-Pro` using `--tool-call-parser kimi_k3`, `--reasoning-parser kimi_k3`, and `--load-format fastsafetensors` on NVIDIA GPUs, AMD GPUs, Google TPUs, and Alibaba T-Head Zhenwu PPUs.

Enable speculative decoding for Kimi K3.1 Pro via `--speculative-config` with `'method':'dspark'` and `'draft_sample_method':'probabilistic'` using the `Inferact/Kimi-K3.1-Pro-DSpark-preview` draft model.

vLLM now supports Kimi K3.1 Pro's Dedicated Sparse Attention (DSA) layers, packing MLA latent, indexer keys, and KDA state into a single contiguous block for use with FlashInfer and FlashMLA sparse kernels.

Token-sharded prefill attention now redistributes query tokens across GPUs via all-to-all instead of splitting by head, reaching up to 1.69× faster time to first token at 340K tokens on 8× GB300.

Decode context parallelism (DCP) with DSA now uses an exact top-k merge that sends ~10× fewer bytes, making the attention module 1.3–1.5× faster at DCP8 versus a generic merge.

+4 more

announced · Oct 29, 2026

vLLM adds Mistral Large 4 serving with DSpark speculative decoding, decode/prefill context parallelism, cross-layer KV sharing, and multimodal image input support.

NEW SUPPORTED PLATFORM NEW INTEGRATION Serve Mistral Large 4 via `vllm serve` with FP8 or NVFP4 checkpoints, using `--tool-call-parser mistral`, `--enable-auto-tool-choice`, and `--reasoning-parser mistral`.

Enable DSpark speculative decoding with `--speculative-config '{"method": "dspark", "num_speculative_tokens": 3}'` for approximately 2x output throughput on TP8.

Shard the latent KV cache for long-context workloads using `--decode-context-parallel-size 8`, raising KV capacity up to 7.9x on TP8.

Allow multiple images per prompt with `--limit-mm-per-prompt '{"image": N}'`.

Cross-layer KV sharing added via `kv_sharing_target_layer_name` on MLAAttention and `AttentionLayerBase`, available model-agnostically.

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

03 Liquid AI

announced · Oct 27, 2026

Open-Weight Models

Liquid AI builds efficient Liquid Foundation Models (LFMs) for on-device, edge, and cloud AI with low latency, privacy, and hardware-aware deployment.

// WHY IT MADE THE LIST

DEPTH 5/5 IMPACT 4/5 NOVELTY 5/5

LFM3 is a new open-weight multimodal model family (text, vision, audio, omni, decision) on a hybrid tri-operator architecture with a 256K context, fine-tunable on a single GPU and runnable on phones, Raspberry Pi, and laptops — and LFM3-VL-2B is the first LFM vision model to accept video input.

> Liquid AI launched LFM3, a new generation of Liquid Foundation Models with a hybrid tri-operator architecture and a 256K-token context window, alongside video-capable vision models, agentic tool-use support, and expanded on-device deployment targets.

└──▷ **WHAT SHIPPED · 4 FEATURES** most completely described first ? what's the number?

01 LFM3 hybrid tri-operator architecture NEW

70

LFM3 is built on a hybrid tri-operator architecture combining gated short convolutions, sliding-window attention, and global attention, with a 256K-token context window, pre-trained on 22T tokens. It is available via Hugging Face, Playground, and the Liquid AI API.

— Names architecture, context window, and training scale, but no code/API example

launch-20261027-4bd98cf0

THINNER COVERAGE BELOW

02 On-device fine-tuning and deployment targets NEW

55

LFM3 models can be fine-tuned locally on a single GPU and run on-device across phones (Samsung Galaxy S26 Ultra, OnePlus 15), Raspberry Pi 5, Apple M4 Max, AMD Ryzen AI laptops, and the Qualcomm Hexagon NPU.

— Names concrete hardware targets but no setup instructions launch-20261027-4bd98cf0

03 Agentic harness support for LFM3-1.7B NEW

50

LFM3-1.7B now supports agentic harnesses including Hermes Agent, Pi, and OpenClaw, with post-training across tool calling, web browsing, and enterprise workflows.

— Names specific harnesses and post-training scope, no usage steps

launch-20261027-4bd98cf0

04 Video input support in LFM3-VL-2B NEW

45

LFM3-VL-2B now accepts video as input, the first LFM vision model to do so, enabling questions over recorded video from phones, lectures, or camera feeds.

— Describes capability and use cases but no concrete API or limit

launch-20261027-4bd98cf0

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

04 Arcade

announced · Oct 29, 2026

AI Agent Frameworks

Arcade is a security platform that enables safe, controlled third-party API access through sandbox environments and permission management.

// WHY IT MADE THE LIST

DEPTH 4/5 IMPACT 4/5 NOVELTY 5/5

MCP Apps support renders editable Gmail, Slack, and Outlook Mail composers and native file pickers directly inside agent chat, with authorization enforced at both interface construction and click time — a first real implementation of the MCP interactive-UI extension in an actions runtime.

> Arcade's biggest window update is a broad set of deployment-hardening controls — per-deployment egress restrictions, worker secret isolation, cloud workload identity, and sizing knobs — alongside in-place project API key rotation, mid-execution tool prompts, automatic remote MCP catalog refresh, and credential scrubbing from logs and execution history. The window also carries a heavy load of breaking changes — toolkit major-version bumps with tool removals, an OAuth scope change, `values.yaml` schema validation, and automatic DB migrations — that require reviewing the Upgrade notes before upgrading.

└──▷ **WHAT SHIPPED · 23 FEATURES** most completely described first ? what's the number?

01 Deployment configuration controls: egress, secrets, identity, sizing

NEW 95

Several new deployment options: send invitation emails from deployments using `coordinator.email`; restrict deployed servers to a specific in-cluster destination without broader private-network access using `deployments.network.egress.allow`; withhold organization and platform secrets from tool servers registered or deployed through the API using `engine.workerSecretSharing`; attach cloud workload identity or a mounted key to the engine pod using `engine.serviceAccount.annotations`, pod labels/annotations, and extra volumes; set default resource requests and limits for deployed workers using `deployments.sizing.*`, with `deployments.resourceQuota.enabled` now the single quota switch; and reserve environment variable names starting with `ARCADE_RUNNER_` for the deployment runner so they cannot be set through a deployed server's configuration.

Prevent a high-privilege deployed server from reading organization-wide secrets it does not need.

```yaml
engine:
  workerSecretSharing: false
```

Allow a deployed server to call an internal service (e.g. a private artifact store) without opening broad private-network access.

```yaml
deployments:
  network:
    egress:
      allow: "internal-artifact-
store.infra.svc.cluster.local:443"
```

Attach a Kubernetes service-account annotation so the engine pod uses cloud workload identity (e.g. AWS IRSA) instead of a mounted static key.

```yaml
engine:
  serviceAccount:
    annotations:
      eks.amazonaws.com/role-arn:
arn:aws:iam::123456789012:role/arcade-engine-role
```

— Names every config key with runnable YAML examples for each `snapshot-20261030`

02 **Ephemeral storage default reduced for deployed workers**

IMPROVED 80

Deployed servers now request `1Gi` of ephemeral storage by default instead of `6Gi`, configurable via `deployments.sizing.ephemeralStorageRequest`.

Reduce ephemeral storage allocation for deployed MCP servers in a resource-constrained cluster.

```yaml
deployments:
  sizing:
    ephemeralStorageRequest: 1Gi
```

— Exact before/after numbers, config key, and example `snapshot-20261101`

03 **Toolkit major-version bumps and tool removals** NEW INTEGRATION

75

BREAKING

Toolkits move to new breaking major versions and several rename tools. Google Calendar's `DeleteEvent` and `FindTimeSlotsWhenEveryoneIsFree`, Workday's `ListTimeOffHistory`, Zoom's `ListUpcomingMeetings`, and several Slack tools with no alias are removed.

— Named tools removed but no migration path given beyond review `snapshot-20261030`

---

04 **Remote MCP catalog auto-refresh via remoteCatalogTtl** NEW

70

Remote MCP catalogs now refresh automatically on a configurable TTL instead of requiring a manual refresh, set with `engine.remoteCatalogTtl`.

Keep remote MCP tool catalogs fresh automatically so operators never need to trigger a manual refresh.

yaml

```yaml
engine:
  remoteCatalogTtl: 300s
```

— Named config key with a runnable example `snapshot-20261101`

---

05 **Compatibility index replaces connection uniqueness index**

70

BREAKING

The one-connection-per-user-and-provider database index is replaced with a compatibility index, which is then dropped; rollback is not clean, so operators should review the Upgrade notes before upgrading.

— Mechanism and rollback risk named, points to Upgrade notes `snapshot-20261101`

---

06 **Tool execution history deletion CLI command** NEW 65

A new engine CLI command deletes an organization's stored tool execution history. It reports each project's recorded executions and logging policy before acting, runs as a dry run by default, and only marks finished executions for deletion once committed with an explicit row cap and reason.

— Describes mechanism and safeguards but no exact command syntax `snapshot-20261105`

---

07 **Short authorization links via enableShortLinks** NEW

65

Enable short authorization links in place of long OAuth URLs with `engine.enableShortLinks`.

Surface short authorization links to users instead of raw OAuth URLs, and label this deployment in usage analytics.

```yaml
engine:
  enableShortLinks: true
usageReporting:
  deploymentDisplayName: 'prod-us-east'
features:
  releaseFeed:
    enabled: true
```

— Exact config key shown with a runnable example `snapshot-20261101`

08 **Usage reporting: deployment display name and link-click event**

NEW 65

Set a display name for usage analytics with `usageReporting.deploymentDisplayName`, and usage reporting now adds an `Authorization link clicked` event.

Surface short authorization links to users instead of raw OAuth URLs, and label this deployment in usage analytics.

```yaml
engine:
  enableShortLinks: true
usageReporting:
  deploymentDisplayName: 'prod-us-east'
features:
  releaseFeed:
    enabled: true
```

— Named config key and event, example provided for the key `snapshot-20261101`

09 **Outdated self-hosted install warning via releaseFeed.enabled**

NEW 65

Admins can enable a warning when a self-hosted install is unsupported or outdated with `features.releaseFeed.enabled`.

Surface short authorization links to users instead of raw OAuth URLs, and label this deployment in usage analytics.

```yaml
engine:
  enableShortLinks: true
usageReporting:
  deploymentDisplayName: 'prod-us-east'
features:
  releaseFeed:
    enabled: true
```

— Named config key with a runnable example `snapshot-20261101`

10 **Automatic database migrations on helm upgrade** BREAKING

65

Two additive database migrations run automatically on `helm upgrade` unless `engine.database.migrations` is set to `false`.

— Named config key to opt out, but migration contents unspecified `snapshot-20261030`

THINNER COVERAGE BELOW

11 **Credential scrubbing from responses, logs, and history**

IMPROVED

55

Tool responses, errors, and execution history no longer echo back credentials a tool call carried; deployed workload logs no longer contain the secrets those servers were configured with or handed per call.

— Clear before/after behavior but no config surface to act on `snapshot-20261030`

12 **Mid-execution tool prompts with reconnect resume** NEW

55

Tools can now prompt the caller for input mid-execution and resume once answered, even across a reconnect, using a new elicitation exchange column.

— Mechanism named but no API or flag to invoke it `snapshot-20261101`

13 **Kubernetes worker self-registration and auto-cataloging** NEW

55

Kubernetes MCP deployments now self-register as managed workers and are cataloged automatically once running, requiring a new signing-secret column in the engine.

— Describes mechanism but no command or config to trigger it `snapshot-20261101`

14 **Additional OIDC scopes required for Google/Microsoft sign-in**

55

BREAKING

Google/Microsoft sign-in now requests additional OpenID Connect scopes — new connections write a profile identifier, which requires updated sign-in scope grants.

— Names the change but not the specific scopes added snapshot-20261030

15 **Project API key rotation and lifecycle management** NEW

50

Rotate, relabel, change expiration, or revive project API keys in place.

— No exact API or UI path named for the operations snapshot-20261030

16 **Quota-permission requirements before upgrading** BREAKING

50

Two self-hosted installs require a quota-permission check before upgrading, and `deployments.enabled: true` now installs quota permissions and readiness assumptions that differ from previous behavior.

— Flags a prerequisite without detailing the check itself snapshot-20261101 snapshot-20261030

17 **Apps interactive tool UI enabled by default** IMPROVED

45

Apps — interactive tool UI on the gateway — is now on by default for servers deployed through Arcade Deploy, pre-built toolkit workers, and tools found through recommendation.

— States scope of default-on change without a config toggle named snapshot-20261030

18 **Older toolkit version fallback for unversioned execute calls**

IMPROVED

45

Unversioned execute-API calls can now opt into an older toolkit version when the newest one's schema rejects the call's inputs.

— Explains trigger condition but no API detail or example snapshot-20261030

19 **`format=oxp` parameter removed from tools-listing endpoint**

45

BREAKING

The `format=oxp` parameter is removed from the tools-listing endpoint.

— Named parameter removed but endpoint path not specified snapshot-20261030

20 **`values.yaml` schema validation on upgrade** BREAKING

45

`values.yaml` is now schema-validated; custom values must be reviewed and corrected before upgrading.

— Named config file but no validation rules detailed `snapshot-20261030`

21 **Arcade Deploy runs custom MCP-protocol servers end-to-end** NEW

40

Arcade Deploy can now run a custom MCP-protocol server end to end, failing early if a dependency is missing.

— Thin description without a command or config example `snapshot-20261101`

22 **`supportImages.busybox` deprecated** DEPRECATED

35

`supportImages.busybox` is deprecated and will be removed in a future minor release.

— Bare deprecation notice with no replacement named `snapshot-20261030`

23 **Outgoing MCP connections support stateless protocol** IMPROVED

25

Outgoing MCP connections now support the newer stateless protocol path.

— One-line mention with no mechanism or config detail `snapshot-20261101`

└──▷ BREAKING ON UPGRADE

! The one-connection-per-user-and-provider database index is replaced with a compatibility index; the old index is then dropped. Rollback is not clean — review Upgrade notes before upgrading.

! Two self-hosted installs require a quota-permission check before upgrading — review Upgrade notes.

! Toolkits move to new breaking major versions and several rename tools — any caller that references tools by name must review the Upgrade notes before upgrading.

! `deployments.enabled: true` now installs quota permissions and readiness assumptions that differ from previous behavior — review before upgrading.

! `values.yaml` is now schema-validated; custom values must be reviewed and corrected before upgrading.

! The `format=oxp` parameter is removed from the tools-listing endpoint.

! Google Calendar's `DeleteEvent` and `FindTimeSlotsWhenEveryoneIsFree`, Workday's `ListTimeOffHistory`, Zoom's `ListUpcomingMeetings`, and several Slack tools with no alias are removed.

! Google/Microsoft sign-in now requests additional OpenID Connect scopes — new connections write a profile identifier, which requires updated sign-in scope grants.

! Two additive database migrations run automatically on `helm upgrade` unless `engine.database.migrations` is set to `false`.

! `supportImages.busybox` is deprecated and will be removed in a future minor release.

announced · Oct 29, 2026

Arcade adds MCP Apps support, rendering interactive in-chat UIs with policy-gated controls and native file upload/download via the MCP protocol extension.

NEW INTEGRATION Arcade now supports MCP Apps (the Model Context Protocol interactive-UI extension), letting tools render editable forms, composers, and dashboards inside the chat instead of switching to a browser tab.

File uploads and downloads now work through MCP Apps: an in-chat interface opens the native file picker, accepts drag-and-drop files such as screenshots, and downloads large files straight to disk, bypassing the current MCP protocol gap for file parameters.

Arcade's authorization and governance runtime enforces access policy at the point of interface construction and again at click time, so UI elements such as Send buttons are suppressed when the calling user lacks the corresponding tool permission.

Agent routines that run unattended are not blocked waiting for human input — the interactive interface is available when a user wants to step in, but routines continue without it.

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

05 Perceptron

announced · Nov 4, 2026

Frontier Models

Perceptron is a layer of intelligence for the physical world. We are a research company building the future of Physical AGI.

// WHY IT MADE THE LIST

DEPTH 4/5 · IMPACT 5/5 · NOVELTY 5/5

Egocentric 1.0 is a self-service API that turns first-person human and robot video into structured robotics training data — timestamped subtask timelines, per-hand activity labels, 2D keypoints, and 3D hand mesh reconstruction — closing a gap teams previously filled by hand-labelling. No comparable tool ships this.

> Perceptron launched Egocentric 1.0, a self-service API that turns first-person human and robot video into structured robotics training data, backed by a new 3D hand reconstruction model.

└──▷ **WHAT SHIPPED · 2 FEATURES** most completely described first ? what's the number?

01 Egocentric 1.0 self-service video-to-data API NEW

77

Perceptron launched a self-service API that accepts first-person human and robot video and returns structured training data including timestamped subtask timelines, continuous per-hand activity labels, 2D keypoints, and 3D hand mesh reconstruction, priced at $9 per video-hour (default 4 fps) or approximately $15 at 8 fps. It supports video from household, robot arm, and manufacturing contexts, and requires no camera calibration for 3D hand reconstruction.

– Pricing and outputs named, but no endpoint or signup flow given.

launch-20261104-924d2e70

02 3D hand reconstruction model built on Mk1.5 NEW

68

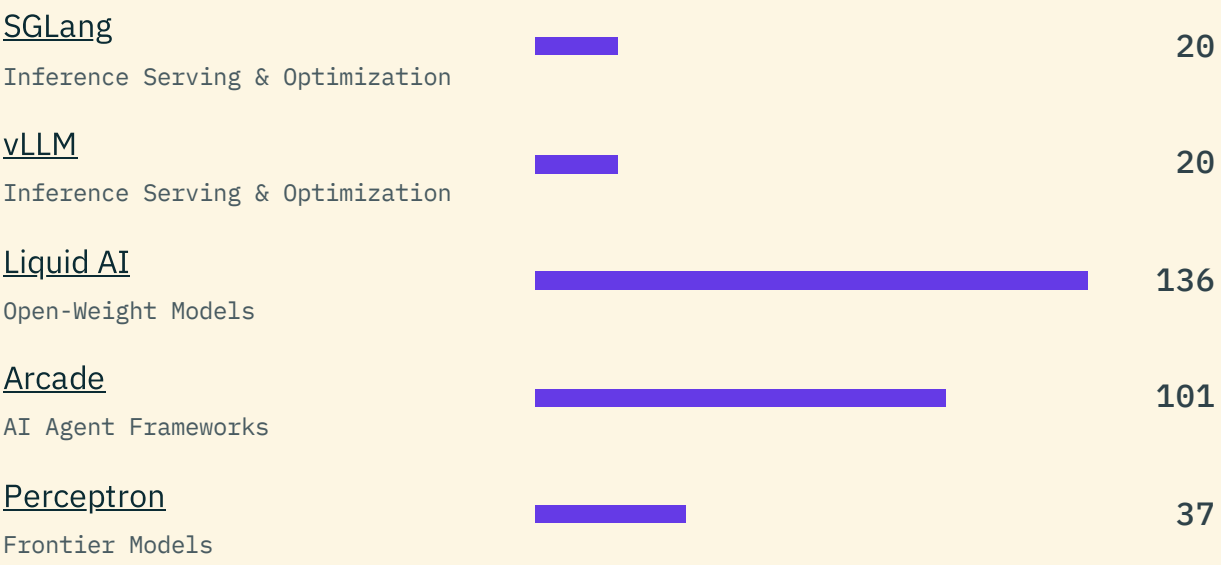
Perceptron added a new 3D hand reconstruction model built on Mk1.5 that estimates each hand's shape, finger articulation, and camera-relative position, achieving 12–25% lower hand-shape error than ACE-Ego-Hand across all four benchmark settings on HOT3D and ARCTIC.

– Benchmarked mechanism detailed but no direct user action. launch-20261104-924d2e70

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

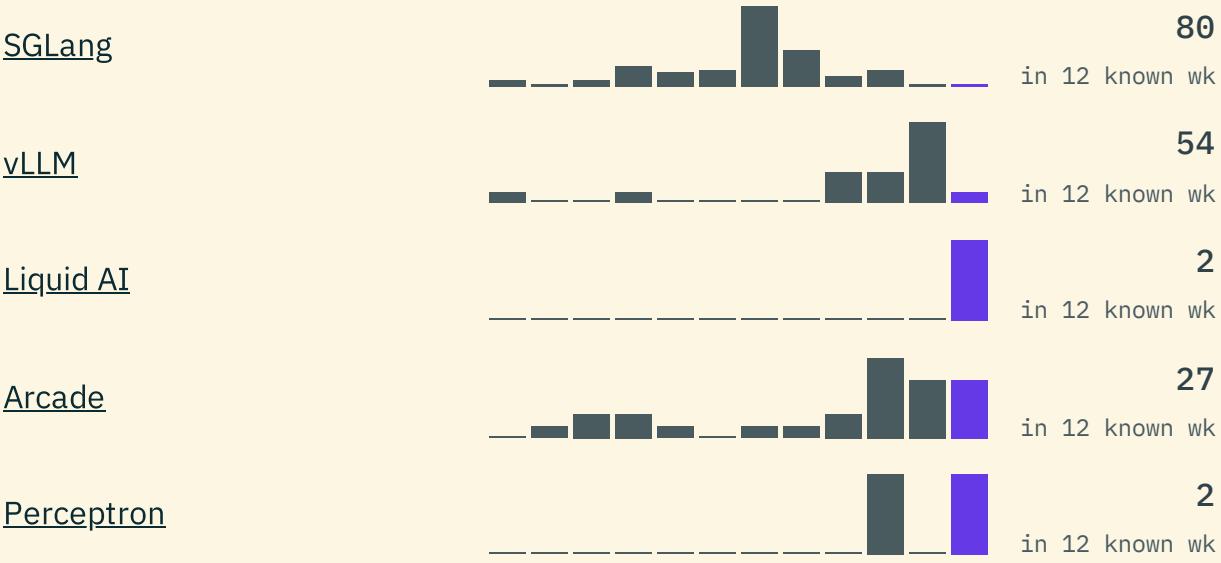
Where each pick stands · Oct 30–Nov 5

Tools with changes in each pick's category over the last 12 weeks, including the pick itself.



Each pick's release rhythm · Oct 30–Nov 5

Changes per week for the last 12 weeks. The bright bar is this week; a dash marks a week with no daily.



Which categories moved · Oct 30–Nov 5

The thick bar is this week. The thin bar is the average of the four weeks before, counting only weeks with a daily.





Who shipped the most · Oct 30–Nov 5

Changes per tool this week. Volume is not a quality score.



// END OF TRANSMISSION

The AI Toolchain · This Week's Highlights · November 6, 2026
The full firehose — every feature release, every day — is in the newsletter.

▷ Subscribe