THE AI TOOLCHAIN

NO. 015

Head

THE WEEK'S TOP PICKS

PUBLISHED NOVEMBER 27, 2026 · FRIDAYS

// SHARE THIS WEEK ▽ Top 5 ▽ Top Pick

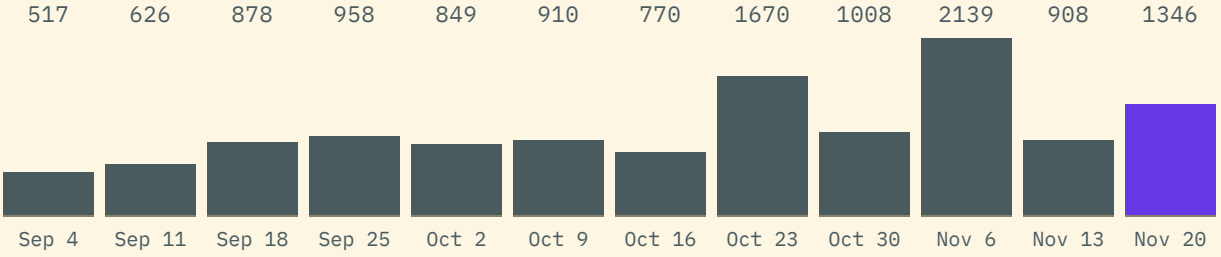
The best tooling updates that shipped this week.

The week in numbers · Nov 20–26

1,346	346	3	604
changes we read →	tools shipped →	product launches →	breaking changes →

Changes per week, last 12 weeks · Nov 20–26

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

Gemini API APEX-Agents Gemini 4 Neon high

vim ? help

82.7% +4.7 pts released 2026-11-18 · leaderboard ↗

OpenAI APEX-Agents GPT-6.2 Sol `max`

66.6% -1.7 pts released 2026-11-17 · leaderboard ↗

Gemini API APEX-Agents Gemini 4 Neon `xhigh`

82.9% +4.9 pts released 2026-11-18 · leaderboard ↗

OpenAI APEX-Agents GPT-6.2 Astra `max`

72.2% +3.9 pts released 2026-11-17 · leaderboard ↗

Gemini API CursorBench Gemini 4 Neon `high`

63.7% +17.3 pts released 2026-11-18 · leaderboard ↗

Gemini API CursorBench Gemini 4 Neon `low`

45.5% -0.9 pts released 2026-11-18 · leaderboard ↗

Gemini API CursorBench Gemini 4 Neon `xhigh`

65.4% +19.0 pts released 2026-11-18 · leaderboard ↗

Gemini API CursorBench Gemini 4 Neon `medium`

57.7% +11.3 pts released 2026-11-18 · leaderboard ↗

Gemini API ARC-AGI-2 Gemini 4 Hydrogen `xhigh`

92.5% +42.9 pts released 2026-11-12 · leaderboard ↗

Gemini API ARC-AGI-2 Gemini 4 Hydrogen `medium`

89.2% +39.6 pts released 2026-11-12 · leaderboard ↗

Gemini API ARC-AGI-2 Gemini 4 Neon `medium`

93.8% +44.2 pts released 2026-11-18 · leaderboard ↗

Gemini API ARC-AGI-2 Gemini 4 Neon `low`

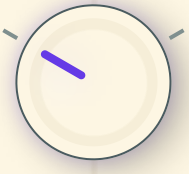
93.8% +44.2 pts released 2026-11-18 · leaderboard ↗

Scores from Epoch AI Benchmarking Hub (CC BY 4.0), evaluated 2026-11-20 to 2026-11-27.

VIEW

ISSUE VIEW

scroll



SCROLL

EXPAND IN PLACE

INFORMATION DENSITY

everything



BRIEFED | EVERYTHING

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

all 33 changes

Do you prefer **scroll**? 👍 👎

This week's 5

01	**Perplexity API**	Feature · Web Search & Scraping for Agents
02	**Claude App**	Feature · AI Chat Apps
03	**ElevenLabs**	Feature · Speech & Voice AI
04	**Together AI**	Launch · Inference Providers & GPU Clouds
05	**browser-use**	Feature · Browser & Computer Use

The picks in full

01 Perplexity API

updated · Nov 25, 2026

Web Search & Scraping for Agents

Perplexity API enables developers to integrate AI-powered search and reasoning capabilities into applications with real-time information retrieval.

// WHY IT MADE THE LIST

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

Perplexity ships three first-of-kind APIs in one week: the Context API fetches and scores web pages by URL with yes/no criteria; the Monitor API creates scheduled Search and Page monitors with webhooks; and `snapshot_as_of` lets agents query any page as it existed up to 365 days ago across Search, Context, and Agent APIs — none of these have prior art in tracked tools.

> Perplexity API added a `perplexity_context` MCP tool for structured page retrieval and a `snapshot_as_of` parameter for historical time-travel across its Search, Context, and Agent APIs, alongside a new Agent API model and a breaking change to the `xhigh` effort level's underlying model.

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

01 perplexity_context tool for page retrieval

Not checked yet NEW

95

A new `perplexity_context` tool on the MCP server, backed by `POST /v1/context`, retrieves page content as Markdown by default with optional HTML, links, query-relevant highlights, yes/no criteria scores, and page metadata; it supports freshness controls (`if_modified_since`, `max_age_seconds`, `cache_only`, `snapshot_as_of`) and `timeout_ms`. It is reachable on the hosted MCP server at `https://api.perplexity.ai/mcp` via OAuth or API key, or locally via `@perplexity-ai/mcp-server` version 1.5.0 in stdio mode.

Pull Markdown content and freshness-filtered highlights for a known URL without hitting the live page — useful for auditing what a site said at a fixed point in time.

```
$ curl -X POST https://api.perplexity.ai/v1/context \
  -H 'Authorization: Bearer <API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{
    "urls": ["https://example.com/report"],
    "snapshot_as_of": "2026-11-01T00:00:00Z",
    "cache_only": true,
    "highlights": true
  }'
```

— Endpoint, flags, and connection methods all named with a runnable example.

snapshot-20261126

02 **snapshot_as_of time-travel across Search, Context, Agent APIs**

Not checked yet NEW 90

Passing `snapshot_as_of` (RFC 3339 with time zone, e.g. `2026-11-01T00:00:00Z`) to `POST /search`, `POST /v1/context`, or the Agent API `web_search` and `fetch_url` tools via `filters` fetches a historical snapshot of any page up to 365 days in the past; pages with no stored copy at that time return `status.type` of `no_content`. The parameter is available in the Perplexity Python library 0.50.0 as a `datetime` object and the TypeScript SDK 0.45.0 as a `Date` object.

Run a historical web search scoped to pages as they existed 90 days ago, to compare search results over time.

```
$ curl -X POST https://api.perplexity.ai/search \
  -H 'Authorization: Bearer <API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{
    "query": "critical infrastructure vulnerabilities",
    "snapshot_as_of": "2026-08-26T00:00:00Z"
  }'
```

— Endpoints, SDK versions, and exact parameter format all specified. snapshot-20261126

03 **xhigh effort level switches to gpt-6.2-sol** Not checked yet

NEW INTEGRATION BREAKING 80

The `xhigh` effort level now routes to `xhigh@2026-11-18.0`, backed by `openai/gpt-6.2-sol` with no fallback model, replacing the previous Claude Opus 5.5 model and its `openai/gpt-6.1-sol` fallback. Callers relying on the old behaviour must pin `xhigh@2026-11-11.0` to preserve the previous model configuration.

— Exact version pins and models named, with a migration path. snapshot-20261120

04 **google/gemini-4-neon model support** Not checked yet NEW INTEGRATION

NEW 60

The Agent API now supports `google/gemini-4-neon`, priced at $3.00 per million input tokens, $0.15 per million cached input tokens, and $15.00 per million output tokens (reasoning tokens included).

— Pricing and model name given, but no usage example. snapshot-20261124

NEW INTEGRATION Orchestrator API GA — Anthropic, OpenAI, Google, xAI, open-weight models: anthropic/claude-opus-5, anthropic/claude-opus-5-5, anthropic/claude-fable-5, anthropic/claude-fable-5-1, anthropic/claude-fable-5-2

NEW SUPPORTED PLATFORM Decider for Mac — local Decisions API on Apple silicon: pplx-decider-flash-v1.2

└──▷ **BREAKING ON UPGRADE**

! The `xhigh` effort level now uses `openai/gpt-6.2-sol` with no fallback model; callers that relied on the previous Claude Opus 5.5 / `openai/gpt-6.1-sol` fallback chain must pin `xhigh@2026-11-11.0` to preserve that behaviour.

updated · Nov 25, 2026

Perplexity API adds the Context API, Monitor API, Batch API, Audit API, Computer MCP incognito mode and usage_get tool, MCP Tasks support, Orchestrator API GA with multi-format requests and auto-routing, and Agent API prompt-cache controls.

New Context API: `POST /v1/context` returns Markdown, HTML, links, relevant passages, and metadata for provided URLs; supports up to five yes/no scoring criteria; client.context.create() in Python 0.48.0 and TypeScript SDK 0.43.0; $1.00 per 1,000 fetched URLs. Not checked yet

New Monitor API: `POST /v1/monitors` creates Search or Page monitors on a schedule with optional webhooks; `client.monitors` in Python 0.47.0 and TypeScript SDK 0.42.0; $0.001 per succeeded run. Not checked yet

New Batch API: submit up to 50,000 Agent API requests as a JSONL file; `client.files` and `client.batches` in the Perplexity Python library and TypeScript SDK; batch requests exempt from Agent API request rate limits. Not checked yet

New Audit API: `GET https://api.perplexity.ai/v1/admin/audit-events` for Enterprise organizations; cursor-based pagination; covers up to 30 days per request. Not checked yet

NEW INTEGRATION Orchestrator API now in general availability: one API key for Anthropic, OpenAI, Google, xAI, and open-weight models; supports OpenAI Chat

Completions, OpenAI Responses, and Anthropic Messages formats; `perplexity/auto` for automatic model routing. `Not checked yet`

`+3 more`

`NEW INTEGRATION` Gemini 4 Neon on Orchestrator API: google/gemini-4-neon, google/gemini-4-hydrogen

`NEW INTEGRATION` GPT-6.2 Sol on Orchestrator API: openai/gpt-6.2-sol, openai/gpt-6.2-astra

`NEW SUPPORTED PLATFORM` Decider for Mac — local Decisions API on Apple silicon: pplx-decider-flash-v1.2

└─▷ **BREAKING ON UPGRADE**

! `answer_question`, `confirm_action_approve`, and `confirm_action_deny` now require `followup_uuid` and fail without it.

! `attachment_urls` is no longer supported on `confirm_action_approve` and `confirm_action_deny`.

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

02 Claude App

November 24, 2026

AI Chat Apps

Claude App is Anthropic's Claude assistant on web, desktop and mobile, where chat and Cowork now run as one unified Claude experience.

// WHY IT MADE THE LIST

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

New to coverage: Claude App now lets users install AI-generated artifacts as native Mac or Windows desktop apps with their own window and icon, ships computer use for Team and Enterprise plans, and adds Claude Meetings BETA where Claude joins Google Meet calls as a participant to take notes and build artifacts — three distinct capability expansions in one week from Anthropic's consumer surface.

> Claude App's biggest additions this window are installable artifact apps for Mac/Windows and a beta that lets Claude join Google Meet calls as a participant, alongside computer-use beta, file transfer with connected apps, and monthly recaps reaching general availability.

└──▷ WHAT SHIPPED · 7 FEATURES most completely described first ? what's the number?

01 Monthly recap (Reflect) reaches general availability

Not checked yet IMPROVED 75

Monthly recap in Settings > Reflect is now out of beta on Free, Pro, Max, and Ultra plans on web, Claude Desktop, iOS, and Android; recaps can now draw on Claude Code sessions and Claude Cowork tasks, and stored recaps can be deleted from the page.

Review or delete your stored monthly recaps after the feature moves out of beta.

📍 Go to Settings › Reflect to view your monthly recap history. Use the delete option on the page to remove any stored recaps.

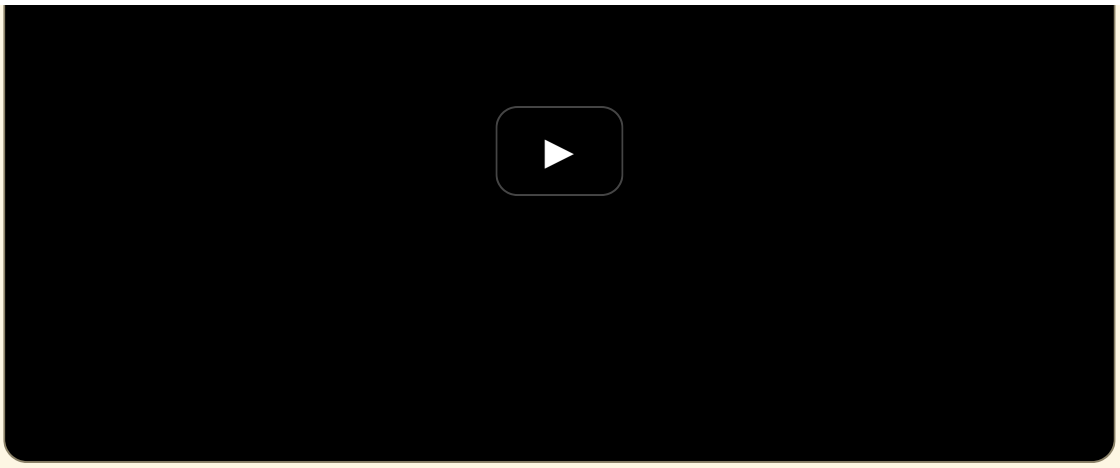
— Exact settings path, plans, platforms and new data sources named November 24, 2026

02 Release channel setting in Claude Desktop

Not checked yet NEW 70

A new release channel setting lets users choose whether Claude Desktop updates to a stable (less frequent, more tested) or latest (daily) release track, via Settings > Release Channel.





Opt into daily Claude Desktop updates to get fixes and new features as soon as they ship

📍 In Claude Desktop, go to Settings › Release Channel and select `Latest` to receive updates approximately once a day.

— Exact setting location and both track options named November 17, 2026

---

03 **Claude Meetings joins Google Meet calls** Not checked yet NEW

70

Claude Meetings BETA, available on Team and Enterprise plans, lets Claude join a Google Meet call as a visible participant to take notes, answer questions, and build artifacts such as docs, whiteboards, and slides during the call.

Add Claude to a live Google Meet call on a Team or Enterprise plan to get real-time notes and artifact generation without leaving the meeting.

📍 In Claude or Claude Desktop, open Claude Meetings BETA and add Claude as a participant to your active Google Meet call.

— Scope and artifact types named, but add-to-call steps are brief November 12, 2026

THINNER COVERAGE BELOW

04 **Install artifacts as native desktop apps** Not checked yet NEW

55

An artifact Claude makes for you can be installed as a native app on Mac or Windows BETA; it gets its own window and icon and can interact with the operating system like any other app.

Turn a Claude-generated artifact into a standalone desktop app you can launch from your dock or taskbar.

📍 In the artifact view, open the artifact menu and select `Install as app`. The app will be installed on your Mac or Windows computer with its own window and icon.

— UI path given but mechanism of OS interaction unspecified November 24, 2026

05 **File transfer with connected apps** Not checked yet NEW

55

Send files to connected apps and save files from them inside Claude, in beta and rolling out gradually on Pro, Max, Ultra, Team, and Enterprise plans (Enterprise requires an owner to turn it on).

Save a file returned by a connected app directly to your computer during a Claude conversation.

> 📍 With a supported app connected, ask Claude to retrieve or generate a file through that app. When the app offers a file, Claude will present a Save file option inline in the conversation.

— Plans and rollout named but save flow described only loosely November 24, 2026

06 **Computer use beta for Team and Enterprise** Not checked yet NEW

55

Computer use BETA is now available on Team and Enterprise plans, letting Claude see a user's screen and act in desktop apps they allow, across Claude Desktop, Claude on the web, and Claude Mobile where Computer access is on.

— Named toggle and plans but no step-by-step setup given November 20, 2026

07 **Refreshed mobile look for iOS 26** Not checked yet IMPROVED

35

A refreshed mobile look is rolling out for Claude for iOS, requiring iOS 26 or later; no opt-in needed, just keep the app updated.

— Version requirement named but visual change undescribed November 20, 2026

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

03 ElevenLabs

November 23, 2026 (2026-11-23)

Speech & Voice AI

ElevenLabs is a voice AI platform with text-to-speech, speech-to-text, voice agents, dubbing and music generation, available in the app and through an API.

// WHY IT MADE THE LIST

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

Agents-as-code via agents.push()/agents.pull() in the Python and JS SDKs lets teams version-control and CI-deploy ElevenLabs voice agents for the first time, and the 100k-character TTS limit (up from 10k) makes single-call long-form audio generation practical; Scribe v3 Live Medical adds HIPAA-eligible zero-retention real-time medical transcription to live agent calls.

> ElevenLabs shipped agents-as-code support and a new Swift SDK for building ElevenAgents programmatically, raised Text to Speech requests to 100,000 characters, added video ad variant generation and multi-market localization to Ads Engine, and introduced Scribe v3 Live Medical with a HIPAA-eligible Zero Retention Mode for healthcare voice agents.

└──▷ WHAT SHIPPED · 14 FEATURES most completely described first ? what's the number?

01 Agents-as-code in Python and JS SDKs

Not checked yet NEW

95

Define ElevenAgents agents in code with the `Agent` builder in the Python and JavaScript SDKs (v2.80.0+), then push or update them — including every webhook and client tool — with agents.push(), pull the project layout with agents.pull(), and validate configurations in CI without an API key using validate().

Push a fully defined agent — settings, tools, and webhooks — to ElevenAgents from code, so your agent config lives in version control alongside your application.

python

```python
from elevenlabs.agents import Agent

agent = Agent(
    name='support-bot',
    llm='gpt-6.2-sol',
    tools=[...],
)
agents.push(agent)
```

Validate an agent definition in CI without an API key to catch configuration errors before deploying.

```python
validate(agent)
```

— Runnable push and validate code examples included November 23, 2026 (2026-11-23)

02 **100,000-character Text to Speech requests** Not checked yet

IMPROVED 88

Text to Speech requests to `eleven_v4` and `eleven_v4_turbo` now accept up to 100,000 characters per request, up from 10,000, enabling roughly 100 minutes of audio from a single call across every Text to Speech endpoint that accepts these models.

Send a long-form document to TTS in one request now that `eleven_v4` accepts up to 100,000 characters, eliminating manual chunking for content up to ~100 minutes of audio.

```
$ curl -X POST https://api.elevenlabs.io/v1/text-to-
speech/<voice_id> \
  -H 'xi-api-key: <api_key>' \
  -H 'Content-Type: application/json' \
  -d '{"text": "<up_to_100000_chars>", "model_id": "eleven_v4"}'
```

— Concrete limit change with a runnable curl example November 23, 2026 (2026-11-23)

03 **Swift SDK v4 conversation client** Not checked yet NEW 82

Swift SDK v4.0.0 introduces a single `ConversationClient` with `startVoiceConversation` for WebRTC voice calls and `startTextOnlyConversation` for WebSocket text-only chats, plus compile-time credential matching, `clientTools` registration, `chatHistory` streaming, and `onDisconnect` reporting.

— Names exact APIs and methods but no code sample given November 23, 2026 (2026-11-23)

04 **Union and array tool parameters** Not checked yet NEW 80

Declare union tool parameters with `AnyOfJsonSchemaProperty` using `any_of` shapes and a `discriminator` property, and declare array parameters with `QueryArrayJsonSchemaProperty` sent as repeated keys such as `?status=open&status=pending`.

— Names exact schema types and query syntax used November 23, 2026 (2026-11-23)

05 **Real-time medical transcription in ElevenAgents** Not checked yet

NEW 78

ElevenAgents now supports Scribe v3 Live Medical, a real-time medical transcription model optimised for drug names and clinical terms in live agent calls, with up to 10,000 custom terms (plan names, therapies, devices, locations) to improve accuracy for organisation-specific vocabulary, and mid-call language switching so a caller who gives a member ID in English and continues in Spanish is transcribed correctly throughout.

— Names model and term limit; no API call shown launch-20261118-29ab9a68

06 **HIPAA-eligible Zero Retention Mode** Not checked yet NEW 65

Enable HIPAA-eligible deployments by signing a BAA and turning on Zero Retention Mode so audio and transcripts are not stored on ElevenLabs infrastructure.

— Names exact toggle but no UI path or API field launch-20261118-29ab9a68

07 **Four new LLMs in ElevenAgents** Not checked yet NEW 61

ElevenAgents now supports four new LLMs — `gpt-6.2-sol`, `gpt-6.2-astra`, `gemini-4-hydrogen`, and `gemini-4-neon` — selectable via the agent LLM picker in the UI or the `llm` field through the API.

— Names every model but no usage example shown November 23, 2026 (2026-11-23)

08 **Video ad hook generation in Ads Engine** Not checked yet NEW 60

Ads Engine generates new video ad hooks by writing alternative openings based on past ad performance, creating voiceovers with a digital voice replica of the original speaker, lip-syncing the speaker to the new audio, and updating captions to match.

— Full pipeline described but no UI path or API shown launch-20261120-5b9157bb

09 **AI-generated soundtracks for video ads** Not checked yet NEW 60

Ads Engine generates new soundtracks for video ads using the ElevenLabs Music API, composing original music by analyzing the tone, pace, and emotion of the visuals, adding sound effects on key moments, and clearing the result for broad commercial use.

— Mechanism and licensing explained but no call shown launch-20261120-5b9157bb

THINNER COVERAGE BELOW

10 **Video ad localization and caption translation in Ads Engine**

Not checked yet NEW 58

Ads Engine now localizes video ads using Dubbing v2 to translate and dub each ad, lip-sync the speaker to the new audio, and optionally route each dub to a native-speaking localization professional for accuracy review, and also translates captions for localized

ads by removing burned-in originals and re-rendering captions in the new language while preserving the original ad's timing and brand fonts.

– Names the engine used but no endpoint or flag given Describes behaviour but no concrete control surface

launch-20261120-5b9157bb

11 **Healthcare Receptionist agent template** Not checked yet NEW

47

ElevenAgents adds a Healthcare Receptionist template with system prompt, call flow, and integration setup pre-configured so teams can reach a working agent in minutes.

– Names template but no navigation path given launch-20261118-29ab9a68

12 **ElevenAgents Architect assistant** Not checked yet NEW

45

ElevenAgents adds the `ElevenAgents Architect` built-in AI assistant, letting teams describe agent behaviour in plain language and have changes edited and tested before going live.

– Names the assistant but no setup command shown launch-20261118-29ab9a68

13 **Guardrails and simulated call testing** Not checked yet NEW

45

ElevenAgents adds custom guardrails, letting teams write plain-language rules — for example, routing any symptom question to a clinician — in addition to built-in topic guardrails, and simulated call testing so teams can verify an agent follows its configured rules before launch.

– Groups two pre-launch safety features with one example launch-20261118-29ab9a68

14 **TikTok ad account connection in Ads Engine** Not checked yet NEW

35

Ads Engine now connects to TikTok ad accounts, in addition to Google, Meta, and LinkedIn, to pull existing creative and performance data.

– One-line integration addition with no setup detail launch-20261120-5b9157bb

GOOD PICK?  GOOD REASON? 

04 Together AI

announced · Nov 25, 2026

Inference Providers & GPU Clouds

Together AI provides an API platform for running and fine-tuning open-source large language models at scale.

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

Together AI ships Hurdle, an open-source npm library for running automated agent evals against APIs with per-run assertions, LLM-based rubric scoring, and mid-run guards that block dangerous agent actions — no tracked tool ships this combination — alongside Together Storage Accelerator (free distributed NVMe caching that cuts model download times up to 20x on GPU Clusters) and Qwen3 TTS with WebSocket streaming.

> Together AI shipped a new open-source agent-eval framework (Hurdle), added Qwen3 TTS speech synthesis with streaming and WebSocket APIs, launched a free NVMe caching layer for GPU Clusters, and brought the 770B-parameter Tencent Hy4 vision-and-tool-calling model to serverless inference.

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

01 Hurdle agent-eval npm library

Not checked yet — NEW — 95

Hurdle is a new open-source npm library for running automated agent evals against APIs, SDKs and docs. Bootstrap a project with `npx hurdle init` (or `npx hurdle init --agent codex` to target Codex instead of Claude Code); define reusable eval suites in a `.suite.mjs` file via `defineSuite`, configuring `api`, `tasks`, `agents`, `conditions`, `repeats`, `verdict` and `judge` fields; write per-run assertions in `checks.mjs` via `defineChecks` using run.expect.request() and run.judge(); and block dangerous agent actions mid-run in `guard.mjs` via `defineGuard` and guard.block.request(). Run a suite and open the results viewer with `npx hurdle run --view`, or reopen saved results with `npx hurdle view`. A sandboxed Together API example ships via `npx hurdle init api-evals --example together-api` plus `npx hurdle setup api-evals`, with the target API and credentials configured through `authorizationCredential` and `hosts` in the suite's `api` block, keeping keys out of source via `.env.local`. Evals can run locally with no sandboxing, or inside isolated Together Sandbox environments where Guardian proxies traffic, enforces guards, and swaps in real API keys without exposing them to the agent.

Bootstrap a new Hurdle eval project and immediately run the starter task to verify your setup end-to-end.

```
$ mkdir my-evals && cd my-evals
npx hurdle init
npx hurdle run --view
```

Prevent an agent from accidentally creating a paid resource during an eval by blocking the relevant API endpoint with a guard.

```javascript
// guard.mjs
import { defineGuard } from 'hurdle';

export default defineGuard((guard) => {
  guard.block.request({
    id: 'no-reservation',
    method: 'POST',
    path: '/v1/compute/clusters',
  });
});
```

Run the same task across two models under two conditions (with and without an MCP skill) 15 times each, scoring with both code checks and an LLM judge.

javascript

```
// example.suite.mjs
import { defineSuite } from 'hurdle';

export default defineSuite({
  api: {
    authorizationCredential: 'TOGETHER_API_KEY',
    hosts: ['api.together.ai'],
  },
  tasks: ['./tasks/gpu-cluster'],
  agents: [
    { runner: 'claude-code', model: 'claude-opus-4-8' },
    { runner: 'claude-code', model: 'claude-sonnet-4-6' },
  ],
  conditions: {
    baseline: {},
    'with-skill': {
      skills: [{ name: 'together-gpu-clusters', sourceDirectory:
'skills/together-gpu-clusters' }],
    },
  },
  repeats: 15,
  verdict: 'hybrid',
  judge: { model: 'claude-opus-4-8' },
});
```

— Full command surface, file formats and sandbox mechanism named launch-20261125-13f1c697

02 **Qwen3 TTS speech synthesis API** Not checked yet NEW INTEGRATION

NEW 95

Together now hosts Qwen3 TTS (model `Qwen/Qwen3-TTS-12Hz-1.7B`) on Serverless Inference and dedicated endpoints at $20 per 1M characters, with nine preset voices each covering Chinese, English, Japanese, Korean, German, French, Russian, Portuguese, Spanish and Italian. Generate speech via client.audio.speech.create(), controlling tone/emotion through an `instructions` field passed via `extra_body`; stream audio over HTTP with `stream=True` and `response_format='raw'`, where each `conversation.item.audio_output.delta` server-sent event carries base64-encoded 16-bit mono PCM at 24 kHz; or connect to the WebSocket endpoint `wss://api.together.ai/v1/audio/speech/websocket` and send text via `input_text_buffer.append` / `input_text_buffer.commit` messages to receive audio as each sentence is synthesized. Output formats are `wav`, `mp3`, `raw` and `mulaw` at a `sample_rate` of 24000 (default) or 8000 (`opus`, `aac`, `flac` unsupported). All nine voices can be previewed against custom text up to 500 characters

at `voicefinder.together.ai`, which generates ready-made Python, TypeScript or cURL snippets.

Synthesize a line with a specific emotional delivery — here `excited and upbeat` — and save the result as an MP3.

```python
from together import Together
client = Together()
response = client.audio.speech.create(
    model="Qwen/Qwen3-TTS-12Hz-1.7B",
    input="Well, that changes everything.",
    voice="aiden",
    language="en",
    extra_body={"instructions": "excited and upbeat"},
    response_format="mp3",
)
response.write_to_file("excited.mp3")
```

Stream audio in real time over WebSocket when input text arrives token-by-token from a language model, committing once complete.

```python
import asyncio, base64, json, os, websockets
URL = (
    "wss://api.together.ai/v1/audio/speech/websocket"
    "?model=Qwen/Qwen3-TTS-12Hz-1.7B&voice=aiden&language=en"
    "&response_format=pcm&sample_rate=24000"
)
async def main():
    headers = {"Authorization": f"Bearer
{os.environ['TOGETHER_API_KEY']}"}
    async with websockets.connect(URL,
additional_headers=headers) as ws:
        await ws.send(json.dumps({"type":
"input_text_buffer.append", "text": "Your order has shipped. "}))
        await ws.send(json.dumps({"type":
"input_text_buffer.append", "text": "It should arrive on
Thursday."}))
        await ws.send(json.dumps({"type":
"input_text_buffer.commit"}))
        with open("order-update.pcm", "wb") as audio:
            async for message in ws:
                event = json.loads(message)
                if event["type"] ==
"conversation.item.audio_output.delta":
                    audio.write(base64.b64decode(event["delta"]))
                elif event["type"] ==
"conversation.item.audio_output.done":
                    break
asyncio.run(main())
```

Audition all nine Qwen3 TTS voices against your own text before committing to one for your application.

📍 1. Open `voicefinder.together.ai`. 2. Select `Qwen/Qwen3-TTS-12Hz-1.7B` in the model menu, then select `Find voices`. 3. Select `Preview` next to any voice to hear a sample. 4. Select `Try custom text`, enter up to 500 characters of your own text and select `Generate`. 5. Repeat for other voices to compare on the same text. 6. Select `Show code` to copy a ready-made Python, TypeScript or cURL request for the chosen voice.

— Model, endpoint, formats, pricing and runnable examples all named

launch-20261120-c2bc072a

---

03 **Together Storage Accelerator caching for GPU Clusters**

Not checked yet NEW 80

Together Storage Accelerator (TSA) is a free distributed NVMe caching layer running on every GPU node in a Together GPU Cluster, serving warm reads from local NVMe at up to 3.4 GB/s — roughly 20x faster than cross-cloud origin pulls — and fetching each object from origin only once regardless of how many nodes request it. Route S3-compatible storage (S3, R2, T4) through it by setting `AWS_ENDPOINT_URL=http://tsa.tsa.svc.cluster.local:18080` with no code changes to existing boto3/CLI workflows, or route Hugging Face Hub downloads by setting `HF_ENDPOINT=http://tsa.tsa.svc.cluster.local:18080`. It is enabled from a cluster's Settings page at no additional cost.

Route all S3-compatible reads (S3, R2, T4) through TSA's local cache so that the first node to fetch an object populates the NVMe cache for every other node in the cluster.

```
$ export AWS_ENDPOINT_URL=http://tsa.tsa.svc.cluster.local:18080
aws s3 cp s3://my-bucket/checkpoint.bin ./checkpoint.bin
```

— Exact env vars and endpoint given; UI path for enabling is not detailed

launch-20261119-a45f770f

---

04 **Tencent Hy4 model on serverless inference** Not checked yet NEW

80

Tencent Hy4, a 770B MoE model, is now available on Together Serverless Inference via the chat completions API using model string `tencent/Hy4`, with a 1M-token context window priced at $0.83 per 1M input tokens, $0.042 per 1M cached input tokens, and $2.50 per 1M output tokens. It is the first model in the Hy family to accept images, via an `image_url` block in message content, and supports structured output by setting `response_format` to `{"type": "json_object"}` or tool calling via the standard `tools` parameter.

Extract structured data from a chart image and return it as JSON using Hy4's vision and JSON output capabilities.

python

```python
import base64
from together import Together

client = Together()

with open("chart.png", "rb") as f:
    image = base64.b64encode(f.read()).decode()

response = client.chat.completions.create(
    model="tencent/Hy4",
    messages=[
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": "Read this chart. Return a JSON
object with a title and a list of bars, each with a label and a
value.",
                },
                {
                    "type": "image_url",
                    "image_url": {"url": f"
{image}"},
                },
            ],
        }
    ],
    response_format={"type": "json_object"},
)

print(response.choices[0].message.content)
```

— Model string, pricing and parameters named with a runnable example

launch-20261119-923aeea4

announced · Nov 21, 2026

Together AI adds Qwen4 Max with a 1M-token context window, streaming-only inference, built-in thinking controls, image input, tool calling, and automatic prompt caching.

NEW INTEGRATION Access Qwen4 Max via Together Serverless Inference using the model string `Qwen/Qwen4-Max`, which requires `stream=True` (or `"stream": true`

in the request body) — requests without streaming return a 400 error with code `streaming_required`. Not checked yet

Control thinking on Qwen4 Max by passing `reasoning={"enabled": False}` to disable it, or setting `reasoning_effort` to `"max"` for harder problems; thinking is on by default and streams in the `delta.reasoning` field, with thinking tokens counted as output tokens. Not checked yet

Qwen4 Max accepts text and image input, with images sent as `image_url` blocks in the message content. Not checked yet

Qwen4 Max supports tool calling through the standard `tools` parameter and can return more than one call per turn. Not checked yet

Prompt caching is automatic for Qwen4 Max on Together AI, with cached input tokens priced at `$0.075` per 1M tokens at list price (or `$0.0525` during the sale through November 22, 2026). Not checked yet

---

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

05 browser-use

announced · Nov 20, 2026

Browser & Computer Use

Make websites accessible for AI agents. Automate tasks online with ease.

// WHY IT MADE THE LIST

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

browser-use 1.0 replaces 24 discrete browser actions with a single persistent Python REPL backed by raw Chrome DevTools Protocol access, letting agents call any CDP command directly — a ground-up architectural shift that opens automation paths no built-in action previously covered.

> browser-use 1.0 overhauls its core architecture, replacing 24 discrete browser actions with a persistent Python REPL that has direct Chrome DevTools Protocol access, alongside a simplified Agent API and an MCP streaming fix for long-running calls.

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

01 Python REPL tool with raw CDP access

Not checked yet

BREAKING

91

browser-use 1.0 replaces the 24 discrete browser action tools from 0.x with a persistent Python REPL exposed as the `python` tool, with browser-harness helpers and cdp() pre-imported. Agents can call any Chrome DevTools Protocol command directly with cdp('Domain.method', ...) from inside the code tool, reaching capabilities no built-in action covered; agents relying on the old named actions must be rewritten to use this tool or custom `Tool` registrations.

Use raw CDP from inside the agent to unblock a hung page by stopping its load and blocking the stuck script before retrying.

```python
cdp('Network.enable')
cdp('Network.setBlockedURLs', urls=['*<stuck-script>*'])
goto_url('https://<exchange-stats-page>/')
```

— Names the tool, cdp() syntax and shows a working example launch-20261120-a4f9f6d6

02 Streaming output for long browser_exec MCP calls

Not checked yet

IMPROVED

70

Calling `browser_exec` again with empty code retrieves the remaining output after a long-running cell exceeds 50 seconds, so MCP clients like Cursor with a 60-second timeout limit get a result instead of failing.

– Names the mechanism and timing but no runnable example `1.0.7`

03 **Three-line Agent().run_sync() entry point** Not checked yet NEW

65

Agent('...').run_sync(), imported via `from browser_use import Agent`, provides a three-line entry point for kicking off agent tasks, with the model and browser now defaulting to Browser Use Cloud.

Kick off an agent task with the new three-line API, using the default Browser Use Cloud browser and model.

python

```
from browser_use import Agent
result = Agent('Find the number of stars of the browser-use
repo').run_sync()
print(result.final_result())
```

– Gives a runnable snippet but little mechanism detail `launch-20261120-a4f9f6d6`

04 **Custom tool registration alongside python tool** Not checked yet

NEW

60

Additional custom tools can be registered alongside the built-in `python` tool using Tool(python, max_retries=TOOL_RETRIES, sequential=True).

– Names the exact call but no further explanation `launch-20261120-a4f9f6d6`

└–▷ **BREAKING ON UPGRADE**

! The 24 discrete browser action tools from 0.x are removed; agents relying on any of those named actions must be rewritten to use the `python` REPL tool with cdp() calls or custom `Tool` registrations.

GOOD PICK? 👍 👎 GOOD REASON? 👍 👎

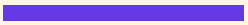
Where each pick stands · Nov 20–26

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

Perplexity API
Web Search & Scraping for Agents
4

Claude App
AI Chat Apps
4

ElevenLabs
Speech & Voice AI



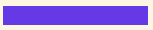
10

Together AI
Inference Providers & GPU Clouds



23

browser-use
Browser & Computer Use



6

Each pick's release rhythm · Nov 20–26

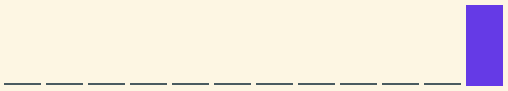
Changes per week for the last 12 weeks. The bright bar is this week; a dash marks a week with no daily.

Perplexity API



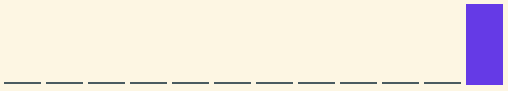
51
in 12 known wk

Claude App



4
in 12 known wk

ElevenLabs



4
in 12 known wk

Together AI



77
in 12 known wk

browser-use

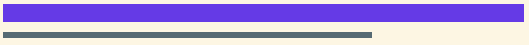


46
in 12 known wk

Which categories moved · Nov 20–26

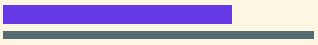
The thick bar is this week. The thin bar is the average of the four weeks before, counting only weeks with a daily.

AI Coding Agents



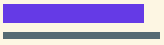
417
usual 295.2

AI Agent Frameworks



177
usual 241

MCP Servers & Tooling



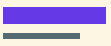
113
usual 125.2

ML Platforms & SDKs

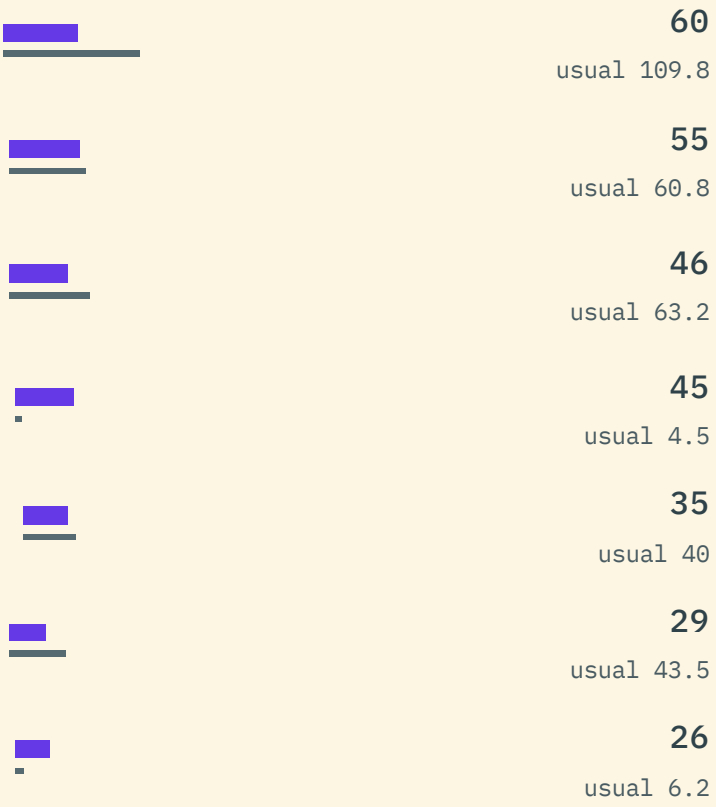


80
usual 85.5

AI / LLM Gateways



80
usual 60.5



Who shipped the most · Nov 20–26

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



[Upcoming conferences →]

// END OF TRANSMISSION

The AI Toolchain · This Week's Highlights · November 27, 2026

The full firehose — every feature release, every day — is in the newsletter.

▷ Subscribe