

The draft model is dissolving *into the target*.

This window opens where issue 02's compile stopped — midday Sunday 23 August — and runs to the end of Sunday 30, so nothing fell between issues. llama.cpp merged DFlash2, a speculative drafter that ships as a purpose-trained head rather than a smaller sibling, and a third party published one for a two-day-old open model. The KV-cache-eviction literature spent the same week auditing its own benchmarks, and the accounting failed. And a new §3 covers the labs, which managed three transcription announcements in five days.

COMPILED 31 AUG 2026

5 SECTIONS

~60 PRIMARY CITATIONS

WINDOW 7½ DAYS

AT A GLANCE

01

llama.cpp merged DFlash2 speculative decoding, and the draft is not a smaller model — it is a head trained for the target. Reported $1.77\times$ to $1.85\times$ decode throughput on Qwen3.8-27B Q4_K_M on an M5 Pro, accepting ~ 5 tokens per block of 8, depending on how the drafter itself is quantized #27342. One rough edge: GGUFs converted before 27 August must be reconverted or vision breaks.

02

Enumerating the optimal KV-eviction subset shows selection was never the problem — the benchmarks were. Over 168,192 attention rows from five models, keeping the largest weights already closes all but a median 2–5% of the gap to full attention; an 87.6-point retrieval margin between published methods traces to storage accounting, because per-head selections stored as masks free no memory 2608.25230. A second paper finds attention magnitude uncorrelated with a token's causal contribution (Spearman $\rho = -0.004$) 2608.27128.

03

Z.AI shipped GLM-5.3-Flash under MIT: 320B total, 18B active, natively multimodal. Reported to approach Claude Opus 4.8 on coding and agentic benchmarks at one-tenth the price of GLM-5.2 [model card](#). The 753B GLM-5.3 shipped beside it under a custom licence. The licence, not the score, is the event: MIT at this capability level resets what "open weights" is expected to mean.

04

Transcription had a week. OpenAI shut down the Assistants API and marked [whisper-1](#) and the [gpt-4o-transcribe](#) family for retirement on 26 August [changelog](#) — the same day Google put Gemini 3.5 Transcribe into public preview at a reported 2.6% word error rate [post](#), two days before Hugging Face's ASR leaderboard added its first Global South language [post](#).

CONTENTS

§1 [llama.cpp](#) — DFlash2, Metal tunings, lazy weights

§2 The papers — the KV-eviction audit

§3 From the labs — OpenAI, Google, Anthropic, Hugging Face

§4 Weights that shipped

§5 What I would actually do

New this issue: §3 covers what the large labs shipped — technical news only, every item cited to the lab's own post, and the section gets deleted in a week when they ship nothing. Two terms were added to the glossary and one was extended.

§1 · [LLAMA.CPP](#)

DFlash2, Metal tunings, and lazy weights

124 pull requests merged between midday 23 and 30 August, ending at build [b10705](#). Grouped by what changes for someone actually running the thing.

The draft head lands

Issue 02 covered DSpark drafters arriving for recurrent models. This week the other speculative family landed: DFlash2 support merged on 27 August #27342, reaching master via #27816. Where a classic draft model is a smaller LLM guessing on the target's behalf, a DFlash2 drafter is a head trained against a specific target: it predicts a whole block of tokens in one pass, adds grouped dynamic depthwise convolution — part of the kernel is predicted from the input — and scores candidates against their predecessors through a pair of codebooks. There is no flag to pass; if the drafter checkpoint is DFlash2, it is used.

The PR's own evaluation, on Qwen3.8-27B Q4_K_M on an Apple M5 Pro (64 GB), over the first eight GSM8K problems at temperature 1.0 with a block size of 8:

DRAFTER QUANT	DECODE	SPEEDUP	ACCEPTANCE
none (autoregressive)	10.42 t/s	1.00×	—
BF16	19.31 t/s	1.85×	4.92 / 8
Q8_0	18.43 t/s	1.77×	5.08 / 8
Q4_K_M	18.89 t/s	1.81×	5.03 / 8

Author-reported on one machine and eight problems, so the magnitudes are claims — but note what varies and what does not: quantizing the drafter all the way to Q4_K_M costs almost nothing, because acceptance holds around five tokens per block either way. The follow-ups arrived within days: missing NVFP4 scales passed to attention #28000, and synthetic acceptance-rate options for benchmarking speculation without a real drafter #27711. DSpark kept spreading too, to Nanbeige4.2-3B #27730 and Nemotron3.5 #27804.

ROUGH EDGE

The DFlash2 merge changed conversion: **GGUFs converted before 27 August 2026 must be reconverted, or vision silently breaks** on multimodal models #27342. If you update past b10700 and images stop working, this is why — reconvert, or re-download a GGUF produced after the merge.

Apple Silicon: a tuning campaign, chip by chip

Nine PRs this week did one unglamorous thing: tune the flash-attention vector kernels per Apple GPU generation — M1 Max #27932, M2 #27940, M3 Pro #27963, M3 Max, M5 and M5 Pro #27863, M3 Ultra #27999, M4 #27875, M4 Pro #27915, and the remaining quantized-KV cases for M2 #28017. If you run llama.cpp on a

Mac, this is free decode throughput for updating; the per-chip tables replace one-size-fits-all launch parameters. A Metal memory leak from missing autoreleasepools was fixed in the same sweep #27758.

Weights that load themselves late

A new `TENSOR_READ_LAZY` path reads tensors from disk only when first needed rather than up front #27794, was hardened #27837, exposed in `llama-bench` #27881, and ended the week renamed to `--lazy-mode` with a `-lzm` shorthand #27969 — four PRs from feature to ergonomics in eight days. In the same spirit, `llama-quantize` now caps its working memory instead of pulling whole big tensors into RAM #27795. Both matter on exactly the machines this publication cares about: the ones with less memory than the model.

New model plumbing, three days after release

Qwen3.8-Flash-Next shipped on Sunday 24 August (§4) and its architecture landed in `llama.cpp` on Wednesday #27742, with a graph-split reduction the day after #27880. Vulkan gained the Lightning Indexer ops for DeepSeek V4 #27453, and the OpenVINO backend moved to 2026.3.1, bringing `whisper.cpp` support and Qwen3.5 on NPUs #27843 — worth noting next to what §3 says about hosted Whisper. The project also quietly bumped its version to 0.3.0 #27696.

Server

A real correctness fix: tool calls were being dropped from the trailing assistant message, which corrupts any agent loop that resumes from server-side history #27626. And `LLAMA_SERVER_SLOTS_N_DIFF` allows slots of different sizes #27600, useful when one server hosts both a chat workload and a long-context one.

§2 · PAPERS

The KV-eviction audit

The dominant cluster this week is KV-cache eviction — not new methods, but the literature checking its own foundations. Four papers, four instruments, one conclusion: the published rankings measure something other than selection quality.

Selection was never the problem

The one to read is **Trust the Mass**. Instead of proposing an eviction rule, it enumerates the exact best subset of keys to keep, on 168,192 attention rows from five models — and finds that the trivial rule, keep the largest attention weights, is already near-optimal: the true best subset closes only a median 2–5% of the remaining gap to full attention. If selection can only move things that little, the large published margins between eviction methods must come from somewhere else, and the paper finds where: *storage*. The strongest query-agnostic methods hold the full cache in memory, because their per-head selections are stored as masks; only ragged per-head storage actually frees bytes. Enforcing a nominal memory budget on one fixed selection costs 14 to 62 benchmark points, and an 87.6-point retrieval margin between methods dissolves under honest accounting 2608.25230.

TwinKV reaches a compatible verdict through a different probe. Doing leave-one-out ablation on individual cached tokens, it finds attention magnitude — the signal most eviction policies rank by — essentially uncorrelated with a token's causal contribution to the answer: Spearman $\rho = -0.004$. Its constructive half is a training-free repair pass: given any policy's retained set, detect evicted tokens whose key has no surviving near-duplicate (orphans) and swap them back in for retained tokens that do 2608.27128.

The other two supply theory and substrate. A probabilistic formalisation proves optimal eviction computationally hard, then shows that framed probabilistically it reduces to expectation estimation, which sampling can approximate — the first formal treatment of a problem the deployed methods have been solving by heuristic 2608.28293. And a controlled GPT-2-scale study finds the attention function itself matters for *learned* eviction: sigmoid attention makes a slightly worse dense model but a much better eviction substrate — sigmoid-gated models delete entries with negligible perplexity change and beat the authors' H₂O and KeyDiff implementations under a matched live-cache protocol, where softmax gates do not 2608.23296.

THE CAVEAT PANEL — EVICTION BENCHMARKS MEASURE STORAGE ACCOUNTING

Put the first two results together and the practical claim is blunt: **a KV-eviction method's published margin tells you little until you know what it physically frees**. Masks do not free memory; ragged per-head storage does, and imposing a real budget costs 14–62 points on the same benchmarks the margins came from 2608.25230. If you run H2O-style eviction in production, the question this week hands you is not "which method ranks highest" but "does my implementation store selections as masks" — because if it does, you are paying quality for memory you never got back.

New jobs for the speculative head

The second cluster explains why §1's DFlash2 merge is not an isolated event: the speculative apparatus is being reused for things besides speed. **AgentSpec** profiles why speculation degrades at large batch sizes in agent serving — high rejection rates and under-used dynamic token budgets — and constrains drafting to

the structured spans where acceptance is high 2608.24004. **AsymSpec** breaks the assumption that drafter and verifier share a context: the small drafter reads the full input while the large verifier sees a compressed view, with a contrastive fusion of logits and a divergence-aware acceptance gate holding verification stable 2608.26004 — inverting the usual roles, because reading everything is cheap for a small model and expensive for a large one. And **Speculative Probing** repurposes the speculative module as a classifier: append a trained soft prompt and the draft head becomes a context-aware monitor for safety filtering at speculative-decoding cost 2608.28099. Three papers in one week treating the draft head as standing infrastructure to build on — which is the issue's thesis, stated by the literature itself.

On-policy distillation, one week later

Issue 02 closed by saying its four-paper convergence on OPD's per-token locality was a hypothesis to revisit. The field revisited it immediately. A critical review of on-policy *self*-distillation names the dominant failure mode — collapse, the progressive narrowing of the reasoning paths a model can produce — and traces three fix families back to that one symptom; its framing of the teacher is the cleanest yet: *no stronger than the student, only better informed* 2608.25936. **VISTA** makes the teacher adapt: outcome-verified rollouts pull the privileged teacher toward the student's distribution rather than treating it as a fixed target 2608.28306. And a token-level analysis of the reverse-KL gradient shows why OPD updates concentrate where they do: the per-token gradient norm factorises into the teacher–student gap times a factor that grows as the sampled token gets less likely under the student, so low-probability tokens dominate training 2608.25643. The cluster is now two weeks old. Still not a trend — but the diagnosis has survived a second week of scrutiny, which is more than most do.

Three that resist the tide

Sliding-window beats linear attention is the week's best contrarian title, and delivers what it says: across multiple LLMs and downstream tasks, plain sliding-window attention with sinks matches or beats post-trained linear-attention retrofits 2608.28444. The comparison is about retrofits, not natively hybrid designs — but if you were considering linearising a pretrained model, the simpler baseline wants an audition first. **H-Scale** works the newest corner of quantization: NVFP4's group-of-16 scales on Blackwell, selected with a Hessian-guided proxy from calibration activations instead of plain reconstruction error 2608.28113. And a security paper formalises the **validation–deployment gap**: because quantization is a many-to-one mapping, a model can pass every full-precision check and activate a backdoor only after being quantized — demonstrated across quantizers 2608.27512. Issue 02's caveat panel said compressed models are not behaviourally neutral by accident; this one says they can be non-neutral *on purpose*. Certify the artifact you deploy, not the one you evaluated.

From the labs

New standing section: what OpenAI, Google, Anthropic, Hugging Face and Meta shipped this week. Technical news only, cited to the lab's own posts, and deleted in a week when nothing ships. Open-weight releases live in §4, not here.

Transcription, three times in five days

On 26 August OpenAI shut down the Assistants API — migration path: the Responses and Conversations APIs — and, in the same changelog entry, marked `whisper-1`, `gpt-4o-transcribe`, `gpt-4o-mini-transcribe` and `gpt-4o-transcribe-diarize` for retirement on 26 February 2027, pointing users at `gpt-transcribe` and `gpt-live-transcribe` [changelog](#). The hosted API named after Whisper is going away; the open Whisper weights, and `whisper.cpp` — which the OpenVINO backend just picked up (§1) — are unaffected, and become the continuity option for anyone who liked the old interface.

The same day, Google put **Gemini 3.5 Transcribe** into public preview: two API variants (`gemini-3.5-transcribe` and a streaming `-live` twin), reported at an average word error rate of 4.0% streaming and 2.6% for pre-recorded audio across 85+ languages, with speaker identification and custom vocabulary [post](#), 26 Aug. Two days later Hugging Face's Open ASR Leaderboard added its first Global South language [post](#), 28 Aug. Three moves in five days is convergence-shaped; it is also three data points, and the NOTE chip in §5 treats it accordingly.

The rest of the week

Anthropic opened a research preview of the **Model Hardware Standard**, a specification for agents operating physical lab and manufacturing equipment — microscopes, liquid handlers, robotic arms — through one standardised interface, built to work with the Model Context Protocol [post](#), 27 Aug. MCP took about a year to go from spec to default plumbing; a hardware sibling is worth watching for the same trajectory, and it is the reason this section exists: it would not surface in an arXiv sweep.

Google also shipped **Gemini Omni 1.1 Flash** for developers: production video generation with scene extension to 40 seconds, first/last keyframe control, cheap 360p drafting reported up to 60% faster at a third of the cost, and 4K upscaling [post](#), 27 Aug. Out of this publication's remit to evaluate, in it to record: video generation is getting the draft-then-upscale workflow that image generation settled on.

Hugging Face had a busy blog besides the leaderboard: Sentence Transformers gained training and fine-tuning for multi-vector late-interaction embedding models [post](#), 26 Aug; Multiverse Computing wrote up **Quantization-Aware Healing**, a recipe reported to recover — and on some tasks exceed — full-precision quality in compressed 4-bit models [post](#), 25 Aug (the underlying paper is from 21 August, out of window: 2608.20953); IBM published a build-notes piece on the Granite 4.2 family [post](#), 25 Aug; and Gradio shipped a workflow builder [post](#), 25 Aug.

Apodex published the technical report for Apodex 1.1, the week's most upvoted paper on Hugging Face's daily list (203 at compile time): "working capability" — sustained, verifiable progress through files, search and code — scaled along two axes, environment diversity and multi-agent coordination 2608.23283. The open-weight Apodex-1.1-mini (36B, Apache 2.0) predates the window — weights 17 August, report 24 August — so it sits here as context rather than in §4 [model card](#).

Meta shipped nothing in the window; its last posts are the early-August Muse line. Recorded so the section stays honest about who was quiet, which is information too.

§4 · RELEASES

Weights that shipped

A loud week after issue 02's quiet one: two Chinese frontier houses and a drafter. Open weights only; specifications are as stated on the model cards.

DATE	MODEL	WHY IT MATTERS
24 Aug	Qwen3.8-Flash-Next	MoE, 125B total with 6B active (512 experts, 10 routed + 1 shared), plus a 51B n-gram embedding table and a 4B MTP head — 180B on disk. Text, image and video input; 262,144 native context extensible to 1M; qwen-community-1.0 licence. Reported 62.5 on SWE-bench Pro and 84.5 on AndroidWorld model card . llama.cpp support merged three days later (§1); community GGUFs followed.
25 Aug	GLM-5.3	753B MoE under a custom glm-5.3 licence, presented as the coding flagship — reported "most capable open-weights model for coding", a 50% improvement over GLM-5.2 on internal benchmarks model card . Not local-hardware territory; it matters as the ceiling the licensed-open ecosystem now has to answer.

DATE	MODEL	WHY IT MATTERS
25 Aug	GLM-5.3-Flash	MIT. 320B total, 18B active, hybrid sparse + linear attention, first natively multimodal GLM-5. Reported 84.3 on Terminal-Bench 2.1 and 63.4 on Deep-SWE, "approaching Claude Opus 4.8" on coding and agentic work at one-tenth GLM-5.2's price <code>model card</code> . 379K downloads in its first week. The permissive licence at this level is the story; 18B active puts quantized multi-GPU serving in reach of a small lab.
27 Aug	Hy4-preview	Tencent's next Hunyuan, staged: a 780B checkpoint under Apache 2.0 whose card is thin on specs <code>model card</code> . Staged releases are where reported dates and specs disagree most (issue 01 learned this on Qwen3.8-Max) — check the repo before repeating any number.
27 Aug	GLM-5.3-Flash-DFlash2	Not a model — a 1.2B DFlash2 drafter for GLM-5.3-Flash , from third-party Inco AI, published two days after the target model <code>model card</code> . The card says benchmark numbers are being re-measured; the licence is CC-BY-NC-ND-4.0, non-commercial — read it before it goes anywhere near production. The existence is the news: drafters are becoming day-two ecosystem artifacts, made by people who did not make the model.

§5 · TAKEAWAYS

What I would actually do

Sorted by readiness, not by how interesting the thing is.

ADOPT

If you convert your own GGUFs for multimodal models, update and reconvert. Conversions made before 27 August silently break vision on builds after the DFlash2 merge #27342. Do it as one operation — update llama.cpp past `b10705`, reconvert, and collect the Metal per-chip tunings (§1) in the same move if you are on a Mac.

ADOPT

If a DFlash2 drafter exists for a model you run, try it — it costs one download. Qwen3.8-27B has one today and the reported 1.77–1.85× with acceptance stable across drafter quantization #27342 is the good configuration: quantize the drafter, keep the speedup. Issue 02's rule still applies — measure acceptance on *your* workload, not throughput on theirs.

WATCH

If anything you run calls `whisper-1` or the `gpt-4o-transcribe` family, put the migration on the calendar now. Retirement is 26 February 2027 [changelog](#). The decision worth making early is hosted-to-hosted versus hosted-to-local: `whisper.cpp` plus the open Whisper weights is the path that cannot be deprecated out from under you, and it got faster on Intel NPUs this very week ([#27843](#)).

WATCH

GLM-5.3-Flash, for the licence more than the checkpoint. MIT at 320B-A18B [model card](#) is a precedent other releases will be measured against. The weights themselves are beyond single-box local use, so the practical watch is quantizations and distills — and note the day-two drafter for it carries a *non-commercial* licence. In a stack where model, drafter and quantization can come from three parties, the licence audit is now per-artifact.

LAB ONLY

Every KV-eviction method, until re-measured under honest storage accounting. That is this week's audit finding, not a general suspicion: near-optimal selection is nearly free ([2608.25230](#)), the ranking signal is uncorrelated with causal contribution ([2608.27128](#)), and imposed budgets cost 14–62 points. If eviction is already in your serving path, check whether selections are stored as masks — you may be paying quality for memory that was never freed.

LAB ONLY

Agent-serving speculation: AgentSpec and AsymSpec. Both attack real deployment problems — batch-size degradation [2608.24004](#) and the context-compression accuracy tax [2608.26004](#) — and both are one week old with author-reported numbers. The idea to keep: acceptance rate is workload-structural. The drafts that survive are the structured spans, which is the same lesson the DSpark table taught in issue 02.

NOTE

What this week did not establish. Three transcription moves in five days reads like an industry turn and is three data points; the KV-eviction audit is the literature auditing itself for one week, and the right response is re-measurement, not abandonment — the constructive halves (TwinKV's repair pass, the sigmoid substrate) are as young as the methods they correct. And the labs section makes its debut in the week a search engine confidently attributed a February model launch to this window; every date above is from the lab's own page, which is the only reason the section can be trusted at all.

Sources

LLAMA.CPP §1

[llama.cpp releases](#) — build tags through [b10705](#) (30 Aug).

Merged pull requests 23–30 Aug were enumerated from the GitHub API (124 merged), and every PR cited below was fetched individually to confirm it merged.

[#27342](#) DFlash2: local convolution + candidate selector · [#27816](#) landing on master · [#28000](#) NVFP4 scales · [#27711](#) synthetic acceptance benchmarking

[#27730](#) DSpark for Nanbeige4.2-3B · [#27804](#) DSpark for Nemotron3.5

Metal fa-vec tunings: [#27932](#) M1 Max · [#27940](#) M2 · [#28017](#) M2 quantized KV · [#27963](#) M3 Pro · [#27863](#) M3 Max / M5 / M5 Pro · [#27999](#) M3 Ultra · [#27875](#) M4 · [#27915](#) M4 Pro

[#27758](#) Metal autoreleasepool leaks

Lazy loading: [#27794](#) TENSOR_READ_LAZY · [#27837](#) hardening · [#27881](#) bench flag · [#27969](#) rename to --lazy-mode

[#27795](#) quantize working-memory cap

[#27742](#) Qwen3.8-Flash-Next architecture · [#27880](#) graph-split reduction · [#27453](#)

DeepSeek V4 Lightning Indexer on Vulkan · [#27843](#) OpenVINO 2026.3.1, whisper.cpp, Qwen3.5 on NPU

[#27696](#) version 0.3.0 · [#27626](#) server: trailing tool calls fix · [#27600](#) server: LLAMA_SERVER_SLOTS_N_DIFF

PAPERS §2

[2608.25230](#) Trust the Mass: Forced Weights in KV-Cache Eviction

[2608.27128](#) TwinKV: A Composable Repair Pass for KV Cache Eviction

[2608.28293](#) A Probabilistic Interpretation of KV Cache Eviction

[2608.23296](#) Sigmoid Attention as a Better Substrate for Learned KV Cache Eviction

[2608.24004](#) AgentSpec: Speculative Decoding for Batch Inference of LLM Agents

[2608.26004](#) AsymSpec: Context-Asymmetric Speculative Decoding for Agentic LLMs

[2608.28099](#) Speculative Probing: LLM Monitoring at Speculative-Decoding Cost
[2608.25936](#) One Symptom, Three Levers: A Critical Review of On-Policy Self-Distillation
[2608.28306](#) VISTA: Verifier-Informed Student-to-Teacher Adaptation
[2608.25643](#) A Token-Level Analysis of Sampled-Token Reverse-KL On-Policy Distillation
[2608.28444](#) Sliding-window beats linear attention
[2608.28113](#) H-Scale: Hessian-Guided Scale Refinement for NVFP4
[2608.27512](#) Quantization-Triggered Backdoors in Language Models

LABS §3

[OpenAI API changelog](#) — Assistants API shutdown and transcription-model deprecations, 26 Aug.

[Gemini 3.5 Transcribe](#) (26 Aug) · [Gemini Omni 1.1 Flash](#) (27 Aug)

[Anthropic: Model Hardware Standard research preview](#) (27 Aug)

Hugging Face blog: [Open ASR Leaderboard](#) (28 Aug) · [multi-vector encoders](#) (26 Aug) · [Quantization-Aware Healing](#) (25 Aug) · [Granite 4.2 build notes](#) (25 Aug) · [Gradio workflows](#) (25 Aug)

[2608.23283](#) Apodex 1.1 technical report · [Apodex-1.1-mini](#) (weights 17 Aug, out of window)

Out of window, cited as context: [2608.20953](#) Quantization-Aware Healing paper (21 Aug).

RELEASES §4

[Qwen3.8-Flash-Next](#) · [GLM-5.3](#) · [GLM-5.3-Flash](#) · [Hy4-preview](#) · [GLM-5.3-Flash-DFlash2](#) — specifications and benchmark figures are as stated on the model cards.

[Hugging Face daily papers](#) — used for discovery across the window; every paper it surfaced was read at its arXiv abstract before being cited.

Compiled 31 August 2026 from the llama.cpp release feed and GitHub API, the arXiv API filtered to this window, Hugging Face's curated daily papers, and — new this issue — the labs' own newsrooms and changelogs. Figures quoted as speedups, error rates or benchmark scores

are the authors' or vendors' own reported numbers on their own benchmarks — treat them as claims, not measurements, until we reproduce them. The window opens midday 23 August, where issue 02's compile stopped; anything dated outside it is labelled inline as context. Terms are defined on the [glossary page](#).



LLM Radar is published weekly by [Montevive](#). [Atom feed](#) · ai@montevive.ai · [Privacy](#)