

# El draft model se está disolviendo *dentro del modelo objetivo.*

Esta ventana empieza donde se detuvo la compilación del número 02 — el mediodía del domingo 23 de agosto — y llega hasta el final del domingo 30, de modo que nada cae entre números. llama.cpp integró DFlash2, un drafter especulativo que se publica como cabeza entrenada a medida y no como un hermano pequeño del modelo, y un tercero publicó uno para un modelo abierto con dos días de vida. La literatura de KV-cache eviction dedicó la misma semana a auditar sus propios benchmarks, y la contabilidad no cuadró. Y una nueva §3 cubre los laboratorios, que encadenaron tres anuncios de transcripción en cinco días.

---

COMPILADO 31 AGO 2026    5 SECCIONES    ~60 CITAS PRIMARIAS    VENTANA 7½ DÍAS

---

DE UN VISTAZO

01

llama.cpp integró el speculative decoding de DFlash2, y el draft no es un modelo más pequeño: es una cabeza entrenada para el objetivo. Se reporta de  $1,77\times$  a  $1,85\times$  de throughput de decode en Qwen3.8-27B Q4\_K\_M en un M5 Pro, aceptando  $\sim 5$  tokens por bloque de 8, según cómo se cuantice el propio drafter #27342. Una arista: los GGUF convertidos antes del 27 de agosto deben reconvertirse o la visión se rompe.

02

**Enumerar el subconjunto óptimo de KV eviction demuestra que la selección nunca fue el problema — los benchmarks sí.** Sobre 168.192 filas de atención de cinco modelos, quedarse con los mayores pesos ya cierra todo salvo una mediana del 2–5% de la distancia a la atención completa; un margen de 87,6 puntos en retrieval entre métodos publicados se explica por la contabilidad del almacenamiento, porque las selecciones por cabeza guardadas como máscaras no liberan memoria <sup>2608.25230</sup>. Un segundo artículo encuentra que la magnitud de la atención no correlaciona con la contribución causal de un token (Spearman  $\rho = -0,004$ ) <sup>2608.27128</sup>.

03

**Z.AI publicó GLM-5.3-Flash bajo MIT: 320B totales, 18B activos, multimodal nativo.** Se reporta que se acerca a Claude Opus 4.8 en benchmarks de código y agénticos a una décima parte del precio de GLM-5.2 <sup>model card</sup>. A su lado salió el GLM-5.3 de 753B con licencia propia. El acontecimiento es la licencia, no la puntuación: MIT a este nivel de capacidad redefine lo que se espera que signifique "pesos abiertos".

04

**La transcripción tuvo su semana.** OpenAI apagó la Assistants API y marcó `whisper-1` y la familia `gpt-4o-transcribe` para retirada el 26 de agosto <sup>changelog</sup> — el mismo día en que Google puso Gemini 3.5 Transcribe en preview pública con un word error rate reportado del 2,6% <sup>post</sup>, dos días antes de que el leaderboard de ASR de Hugging Face añadiera su primera lengua del Sur Global <sup>post</sup>.

---

## CONTENIDO

§1 llama.cpp — DFlash2, tunings de Metal, pesos perezosos

§2 Los papers — la auditoría del KV eviction

§3 Desde los laboratorios — OpenAI, Google, Anthropic, Hugging Face

§4 Pesos que se publicaron

§5 Qué haría yo

Novedad de este número: la §3 cubre lo que publicaron los grandes laboratorios — solo noticias técnicas, cada una citada al post del propio laboratorio, y la sección se elimina la semana en que no publiquen nada. Se añadieron dos términos al glosario y se amplió uno.

# DFlash2, tunings de Metal y pesos perezosos

124 pull requests fusionados entre el mediodía del 23 y el 30 de agosto, terminando en la build `b10705`. Agrupados por lo que cambian para quien de verdad ejecuta esto.

## Aterriza la cabeza de draft

El número 02 cubrió la llegada de drafters DSpark para modelos recurrentes. Esta semana aterrizó la otra familia especulativa: el soporte de DFlash2 se fusionó el 27 de agosto [#27342](#), llegando a master vía [#27816](#). Donde un draft model clásico es un LLM más pequeño adivinando por el objetivo, un drafter DFlash2 es una cabeza entrenada contra un objetivo concreto: predice un bloque entero de tokens en una pasada, añade una convolución depthwise dinámica agrupada — parte del kernel se predice desde la entrada — y puntúa candidatos contra sus predecesores mediante un par de codebooks. No hay flag que pasar; si el checkpoint del drafter es DFlash2, se usa.

La evaluación del propio PR, en Qwen3.8-27B Q4\_K\_M en un Apple M5 Pro (64 GB), sobre los ocho primeros problemas de GSM8K a temperatura 1,0 y bloque de 8:

| CUANT. DEL DRAFTER       | DECODE    | ACELERACIÓN | ACEPTACIÓN |
|--------------------------|-----------|-------------|------------|
| ninguno (autorregresivo) | 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   |

Reportado por el autor en una sola máquina y ocho problemas, así que las magnitudes son afirmaciones — pero fíjate en qué varía y qué no: cuantizar el drafter hasta Q4\_K\_M apenas cuesta nada, porque la aceptación se mantiene en torno a cinco tokens por bloque en todos los casos. Las continuaciones llegaron en días: escalas NVFP4 que faltaban pasadas a la atención [#28000](#), y opciones sintéticas de tasa de aceptación para medir especulación sin un drafter real [#27711](#). DSpark también siguió extendiéndose, a Nanbeige4.2-3B [#27730](#) y Nemotron3.5 [#27804](#).

La fusión de DFlash2 cambió la conversión: **los GGUF convertidos antes del 27 de agosto de 2026 deben reconvertirse, o la visión se rompe en silencio** en modelos multimodales #27342. Si actualizas más allá de `b10700` y las imágenes dejan de funcionar, es por esto — reconvierde, o descarga un GGUF producido después de la fusión.

## Apple Silicon: una campaña de tuning, chip a chip

Nueve PRs hicieron esta semana una sola cosa poco glamurosa: ajustar los kernels vectoriales de flash attention por generación de GPU de Apple — M1 Max #27932, M2 #27940, M3 Pro #27963, M3 Max, M5 y M5 Pro #27863, M3 Ultra #27999, M4 #27875, M4 Pro #27915, y los casos restantes de KV cuantizada para M2 #28017. Si ejecutas llama.cpp en un Mac, esto es throughput de decode gratis por actualizar; las tablas por chip sustituyen a parámetros de lanzamiento de talla única. En la misma pasada se corrigió una fuga de memoria de Metal por autoreleasepools ausentes #27758.

## Pesos que se cargan tarde y solos

Una nueva vía `TENSOR_READ_LAZY` lee los tensores de disco solo cuando se necesitan por primera vez, en lugar de por adelantado #27794; se endureció #27837, se expuso en `llama-bench` #27881, y acabó la semana renombrada a `--lazy-mode` con el atajo `-lzm` #27969 — cuatro PRs de la función a la ergonomía en ocho días. En el mismo espíritu, `llama-quantize` ahora limita su memoria de trabajo en lugar de subir tensores grandes enteros a RAM #27795. Ambas cosas importan exactamente en las máquinas que le interesan a esta publicación: las que tienen menos memoria que el modelo.

## Fontanería para modelos nuevos, tres días después del lanzamiento

Qwen3.8-Flash-Next salió el domingo 24 de agosto (§4) y su arquitectura aterrizó en llama.cpp el miércoles #27742, con una reducción de graph splits al día siguiente #27880. Vulkan ganó las operaciones del Lightning Indexer para DeepSeek V4 #27453, y el backend de OpenVINO pasó a 2026.3.1, con soporte de whisper.cpp y Qwen3.5 en NPUs #27843 — conviene tenerlo al lado de lo que la §3 dice del Whisper alojado. El proyecto también subió discretamente su versión a 0.3.0 #27696.

## Servidor

Una corrección de verdad: las tool calls se estaban perdiendo del último mensaje del asistente, lo que corrompe cualquier bucle de agente que reanude desde el historial del servidor #27626. Y

`LLAMA_SERVER_SLOTS_N_DIFF` permite slots de distintos tamaños #27600, útil cuando un mismo servidor aloja una carga de chat y otra de contexto largo.

---

§2 · PAPERS

## La auditoría del KV eviction

El clúster dominante de la semana es el KV-cache eviction — no métodos nuevos, sino la literatura comprobando sus propios cimientos. Cuatro artículos, cuatro instrumentos, una conclusión: los rankings publicados miden otra cosa distinta de la calidad de la selección.

### La selección nunca fue el problema

El que hay que leer es **Trust the Mass**. En lugar de proponer una regla de eviction, enumera el mejor subconjunto exacto de claves a conservar, sobre 168.192 filas de atención de cinco modelos — y encuentra que la regla trivial, quedarse con los mayores pesos de atención, ya es casi óptima: el mejor subconjunto verdadero solo cierra una mediana del 2–5% de la distancia restante a la atención completa. Si la selección solo puede mover tan poco, los grandes márgenes publicados entre métodos tienen que venir de otro sitio, y el artículo encuentra de dónde: del *almacenamiento*. Los métodos query-agnostic más fuertes mantienen la caché completa en memoria, porque sus selecciones por cabeza se guardan como máscaras; solo el almacenamiento irregular por cabeza libera bytes de verdad. Imponer un presupuesto nominal de memoria sobre una selección fija cuesta de 14 a 62 puntos de benchmark, y un margen de 87,6 puntos en retrieval entre métodos se disuelve bajo una contabilidad honesta 2608.25230.

**TwinKV** llega a un veredicto compatible por otra sonda. Haciendo ablación leave-one-out sobre tokens individuales de la caché, encuentra que la magnitud de la atención — la señal por la que ordenan casi todas las políticas de eviction — no correlaciona con la contribución causal del token a la respuesta: Spearman  $\rho = -0,004$ . Su mitad constructiva es una pasada de reparación sin entrenamiento: dado el conjunto retenido por cualquier política, detectar los tokens expulsados cuya clave no tiene un casi-duplicado superviviente (huérfanos) y recuperarlos a cambio de tokens retenidos que sí lo tienen 2608.27128.

Los otros dos aportan teoría y sustrato. Una formalización probabilística demuestra que el eviction óptimo es computacionalmente duro, y luego muestra que, planteado probabilísticamente, se reduce a estimar una esperanza, que el muestreo puede aproximar — el primer tratamiento formal de un problema que los métodos desplegados llevaban resolviendo por heurística 2608.28293. Y un estudio controlado a escala GPT-2 encuentra que la propia función de atención importa para el eviction *aprendido*: la atención sigmoide hace

un modelo denso algo peor pero un sustrato de eviction mucho mejor — los modelos con puertas sigmoides borran entradas con un cambio de perplejidad despreciable y superan a las implementaciones de H<sub>2</sub>O y KeyDiff de los autores bajo un protocolo igualado de caché viva, donde las puertas softmax no lo hacen 2608.23296.

#### EL PANEL DE ADVERTENCIA — LOS BENCHMARKS DE EVICTION MIDEN CONTABILIDAD DE ALMACENAMIENTO

Junta los dos primeros resultados y la conclusión práctica es contundente: **el margen publicado de un método de KV eviction dice poco hasta saber qué libera físicamente**. Las máscaras no liberan memoria; el almacenamiento irregular por cabeza sí, e imponer un presupuesto real cuesta 14–62 puntos en los mismos benchmarks de donde salieron los márgenes 2608.25230. Si ejecutas eviction estilo H<sub>2</sub>O en producción, la pregunta que te deja esta semana no es "qué método puntuó más alto" sino "¿mi implementación guarda las selecciones como máscaras?" — porque si lo hace, estás pagando calidad por una memoria que nunca recuperaste.

## Nuevos oficios para la cabeza especulativa

El segundo clúster explica por qué la fusión de DFlash2 de la §1 no es un hecho aislado: el aparato especulativo se está reutilizando para cosas distintas de la velocidad. **AgentSpec** perfila por qué la especulación se degrada con batches grandes al servir agentes — tasas de rechazo altas y presupuestos dinámicos de tokens infrautilizados — y restringe el drafting a los tramos estructurados donde la aceptación es alta 2608.24004. **AsymSpec** rompe el supuesto de que drafter y verificador comparten contexto: el drafter pequeño lee la entrada completa mientras el verificador grande ve una vista comprimida, con una fusión contrastiva de logits y una puerta de aceptación consciente de la divergencia que mantiene estable la verificación 2608.26004 — invirtiendo los papeles habituales, porque leerlo todo es barato para un modelo pequeño y caro para uno grande. Y **Speculative Probing** reutiliza el módulo especulativo como clasificador: añade un soft prompt entrenado y la cabeza de draft se convierte en un monitor con contexto para filtrado de seguridad al coste del speculative decoding 2608.28099. Tres artículos en una semana tratando la cabeza de draft como infraestructura estable sobre la que construir — que es la tesis de este número, enunciada por la propia literatura.

## Destilación on-policy, una semana después

El número 02 cerraba diciendo que su convergencia de cuatro artículos sobre la localidad por token del OPD era una hipótesis que revisar. El campo la revisó de inmediato. Una revisión crítica de la *auto*-destilación on-policy nombra el modo de fallo dominante — el colapso, el estrechamiento progresivo de los caminos de razonamiento que un modelo puede producir — y remonta tres familias de arreglos a ese único síntoma; su caracterización del teacher es la más limpia hasta ahora: *no más fuerte que el student, solo mejor informado* 2608.25936. **VISTA** hace que el teacher se adapte: rollouts verificados por resultado atraen al teacher privilegiado hacia la distribución del student en lugar de tratarlo como objetivo fijo 2608.28306. Y un análisis

a nivel de token del gradiente de la KL inversa muestra por qué las actualizaciones del OPD se concentran donde se concentran: la norma del gradiente por token se factoriza en la brecha teacher–student por un factor que crece cuanto menos probable es el token muestreado bajo el student, así que los tokens de baja probabilidad dominan el entrenamiento 2608.25643. El clúster tiene ya dos semanas. Sigue sin ser una tendencia — pero el diagnóstico ha sobrevivido a una segunda semana de escrutinio, que es más de lo que consigue la mayoría.

## Tres que van a contracorriente

**Sliding-window beats linear attention** es el mejor título contracorriente de la semana, y cumple lo que dice: en varios LLM y tareas, la sliding-window attention con sinks iguala o supera a los retrofits post-entrenados de linear attention 2608.28444. La comparación va de retrofits, no de diseños nativamente híbridos — pero si estabas pensando en linealizar un modelo preentrenado, el baseline más simple merece una audición primero. **H-Scale** trabaja la esquina más nueva de la cuantización: las escalas por grupo de 16 de NVFP4 en Blackwell, elegidas con un proxy de segundo orden a partir de activaciones de calibración en lugar del error de reconstrucción a secas 2608.28113. Y un artículo de seguridad formaliza la **brecha validación–despliegue**: como la cuantización es una aplicación de muchos a uno, un modelo puede pasar todas las comprobaciones a precisión completa y activar una puerta trasera solo tras ser cuantizado — demostrado entre distintos cuantizadores 2608.27512. El panel de la cautela del número 02 decía que los modelos comprimidos no son conductualmente neutros por accidente; este dice que pueden no serlo *a propósito*. Certifica el artefacto que despliegas, no el que evaluaste.

---

§3 · LABORATORIOS

## Desde los laboratorios

Nueva sección permanente: lo que OpenAI, Google, Anthropic, Hugging Face y Meta publicaron esta semana. Solo noticias técnicas, citadas a los posts de los propios laboratorios, y la sección se elimina la semana en que no haya nada. Los lanzamientos de pesos abiertos viven en la §4, no aquí.

## Transcripción, tres veces en cinco días

El 26 de agosto OpenAI apagó la Assistants API — ruta de migración: las APIs Responses y Conversations — y, en la misma entrada del changelog, marcó `whisper-1`, `gpt-4o-transcribe`, `gpt-4o-mini-transcribe` y `gpt-4o-transcribe-diarize` para retirada el 26 de febrero de 2027, apuntando a `gpt-transcribe` y `gpt-`

`live-transcribe` [changelog](#). La API alojada que lleva el nombre de Whisper desaparece; los pesos abiertos de Whisper, y `whisper.cpp` — que el backend de OpenVINO acaba de incorporar (§1) — no se ven afectados, y se convierten en la opción de continuidad para quien apreciara la interfaz antigua.

El mismo día, Google puso **Gemini 3.5 Transcribe** en preview pública: dos variantes de API ( `gemini-3.5-transcribe` y una gemela `-live` en streaming), reportada con un word error rate medio del 4,0% en streaming y del 2,6% en audio pregrabado en más de 85 idiomas, con identificación de hablantes y vocabulario personalizado [post](#), 26 ago. Dos días después, el Open ASR Leaderboard de Hugging Face añadió su primera lengua del Sur Global [post](#), 28 ago. Tres movimientos en cinco días tienen forma de convergencia; también son tres puntos de datos, y el chip de NOTA de la §5 los trata como tales.

## El resto de la semana

**Anthropic** abrió una research preview del **Model Hardware Standard**, una especificación para que agentes operen equipamiento físico de laboratorio y fabricación — microscopios, manipuladores de líquidos, brazos robóticos — a través de una interfaz estandarizada, construida para funcionar con el Model Context Protocol [post](#), 27 ago. MCP tardó en torno a un año en pasar de especificación a fontanería por defecto; un hermano para hardware merece vigilancia por la misma trayectoria, y es la razón de que exista esta sección: no aparecería en un barrido de arXiv.

**Google** también publicó **Gemini Omni 1.1 Flash** para desarrolladores: generación de vídeo de producción con extensión de escenas hasta 40 segundos, control de primer y último fotograma, borradores baratos a 360p reportados hasta un 60% más rápidos a un tercio del coste, y reescalado a 4K [post](#), 27 ago. Fuera del alcance de esta publicación evaluarlo, dentro registrarlo: la generación de vídeo está adoptando el flujo borrador-y-reescala en el que ya se asentó la generación de imágenes.

**Hugging Face** tuvo un blog cargado además del leaderboard: Sentence Transformers ganó entrenamiento y fine-tuning de modelos de embeddings multivector de interacción tardía [post](#), 26 ago; Multiverse Computing escribió sobre **Quantization-Aware Healing**, una receta que, según reportan, recupera — y en algunas tareas supera — la calidad de precisión completa en modelos comprimidos a 4 bits [post](#), 25 ago (el artículo subyacente es del 21 de agosto, fuera de ventana: 2608.20953); IBM publicó unas notas de construcción de la familia Granite 4.2 [post](#), 25 ago; y Gradio lanzó un constructor de workflows [post](#), 25 ago.

**Apodex** publicó el informe técnico de Apodex 1.1, el paper más votado de la semana en la lista diaria de Hugging Face (203 al compilar): "working capability" — progreso sostenido y verificable a través de ficheros, búsqueda y código — escalado en dos ejes, diversidad de entornos y coordinación multiagente 2608.23283. El Apodex-1.1-mini de pesos abiertos (36B, Apache 2.0) es anterior a la ventana — pesos del 17 de agosto, informe del 24 — así que figura aquí como contexto y no en la §4 [model card](#).

**Meta** no publicó nada en la ventana; sus últimos posts son la línea Muse de principios de agosto. Queda registrado para que la sección sea honesta sobre quién estuvo callado, que también es información.

# Pesos que se publicaron

Una semana ruidosa tras la tranquila del número 02: dos casas frontera chinas y un drafter. Solo pesos abiertos; las especificaciones son las declaradas en las model cards.

| FECHA  | MODELO                | POR QUÉ IMPORTA                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--------|-----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 24 ago | Qwen3.8-Flash-Next    | MoE, 125B totales con 6B activos (512 expertos, 10 enrutados + 1 compartido), más una tabla de embeddings de n-gramas de 51B y una cabeza MTP de 4B — 180B en disco. Entrada de texto, imagen y vídeo; contexto nativo de 262.144 extensible a 1M; licencia qwen-community-1.0. Reporta 62,5 en SWE-bench Pro y 84,5 en AndroidWorld model card. El soporte en llama.cpp se fusionó tres días después (§1); los GGUF de la comunidad llegaron detrás.               |
| 25 ago | GLM-5.3               | MoE de 753B bajo licencia propia glm-5.3, presentado como el buque insignia de código — reportado como "el modelo de pesos abiertos más capaz para programar", una mejora del 50% sobre GLM-5.2 en benchmarks internos model card. No es territorio de hardware local; importa como el techo al que el ecosistema abierto con licencia tiene ahora que responder.                                                                                                   |
| 25 ago | GLM-5.3-Flash         | MIT. 320B totales, 18B activos, atención híbrida sparse + linear, primer GLM-5 nativamente multimodal. Reporta 84,3 en Terminal-Bench 2.1 y 63,4 en Deep-SWE, "acercándose a Claude Opus 4.8" en código y trabajo agéntico a una décima parte del precio de GLM-5.2 model card. 379.000 descargas en su primera semana. La historia es la licencia permisiva a este nivel; 18B activos ponen el servicio multi-GPU cuantizado al alcance de un laboratorio pequeño. |
| 27 ago | Hy4-preview           | El siguiente Hunyuan de Tencent, por fases: un checkpoint de 780B bajo Apache 2.0 cuya card es parca en especificaciones model card. Los lanzamientos por fases son donde más discrepan fechas y especificaciones reportadas (el número 01 lo aprendió con Qwen3.8-Max) — comprueba el repositorio antes de repetir ningún número.                                                                                                                                  |
| 27 ago | GLM-5.3-Flash-DFlash2 | No es un modelo — es un drafter DFlash2 de 1,2B para GLM-5.3-Flash, del tercero Inco AI, publicado dos días después del modelo objetivo model card. La card dice que los números de benchmark se están volviendo a medir; la licencia es CC-BY-NC-ND-4.0, no comercial — léela antes de que se acerque a producción. La noticia es que exista: los drafters se están convirtiendo en artefactos de ecosistema del día dos, hechos por gente que no hizo el modelo.  |

## Qué haría yo

Ordenado por madurez, no por lo interesante que sea cada cosa.

### ADOPTAR

**Si conviertes tus propios GGUF de modelos multimodales, actualiza y reconvierte.** Las conversiones anteriores al 27 de agosto rompen la visión en silencio en builds posteriores a la fusión de DFlash2 #27342. Hazlo en una sola operación — actualiza llama.cpp más allá de b10705 , reconvierte, y llévate de paso los tunings por chip de Metal (§1) si estás en un Mac.

### ADOPTAR

**Si existe un drafter DFlash2 para un modelo que ejecutas, pruébalo — cuesta una descarga.** Qwen3.8-27B tiene uno hoy y el 1,77–1,85× reportado con la aceptación estable frente a la cuantización del drafter #27342 es la configuración buena: cuantiza el drafter, conserva la aceleración. La regla del número 02 sigue vigente — mide la aceptación en *tu* carga de trabajo, no el throughput en la suya.

### VIGILAR

**Si algo de lo que ejecutas llama a whisper-1 o a la familia gpt-4o-transcribe , pon la migración en el calendario ya.** La retirada es el 26 de febrero de 2027 changelog. La decisión que conviene tomar pronto es alojado-a-alojado frente a alojado-a-local: whisper.cpp con los pesos abiertos de Whisper es la vía que no te pueden retirar, y esta misma semana se aceleró en NPUs de Intel (#27843).

### VIGILAR

**GLM-5.3-Flash, más por la licencia que por el checkpoint.** MIT a 320B-A18B model card es un precedente contra el que se medirán otros lanzamientos. Los pesos en sí quedan fuera del uso local en una sola máquina, así que lo práctico es vigilar cuantizaciones y destilaciones — y nótese que el drafter del día dos lleva licencia *no comercial*. En un stack donde modelo, drafter y cuantización pueden venir de tres partes, la auditoría de licencias es ya por artefacto.

**Todos los métodos de KV eviction, hasta que se re-midan con una contabilidad honesta del almacenamiento.** Eso es el hallazgo de la auditoría de esta semana, no una sospecha general: la selección casi óptima es casi gratis (2608.25230), la señal de ordenación no correlaciona con la contribución causal (2608.27128), y los presupuestos impuestos cuestan 14–62 puntos. Si el eviction ya está en tu ruta de servicio, comprueba si las selecciones se guardan como máscaras — puede que estés pagando calidad por una memoria que nunca se liberó.

**Especulación al servir agentes: AgentSpec y AsymSpec.** Ambos atacan problemas reales de despliegue — la degradación con el tamaño de batch 2608.24004 y el impuesto en precisión de comprimir el contexto 2608.26004 — y ambos tienen una semana de vida y números reportados por sus autores. La idea que quedarse: la tasa de aceptación es estructural de la carga de trabajo. Los drafts que sobreviven son los tramos estructurados, la misma lección que enseñó la tabla de DSpark en el número 02.

**Lo que esta semana no estableció.** Tres movimientos de transcripción en cinco días parecen un giro de industria y son tres puntos de datos; la auditoría del KV eviction es la literatura auditándose a sí misma durante una semana, y la respuesta correcta es re-medir, no abandonar — las mitades constructivas (la pasada de reparación de TwinKV, el sustrato sigmoide) son tan jóvenes como los métodos que corrigen. Y la sección de laboratorios debuta la misma semana en que un buscador atribuyó con toda confianza a esta ventana un lanzamiento de modelo de febrero; cada fecha de arriba viene de la página del propio laboratorio, que es la única razón por la que la sección merece confianza.

---

## Fuentes

LLAMA.CPP §1

[llama.cpp releases](#) — builds hasta **b10705** (30 ago).

Los pull requests fusionados del 23 al 30 de agosto se enumeraron desde la API de GitHub (124 fusionados), y cada PR citado abajo se consultó individualmente para confirmar su fusión.

[#27342](#) DFlash2: convolución local + selector de candidatos · [#27816](#) llegada a master · [#28000](#) escalas NVFP4 · [#27711](#) aceptación sintética para benchmarks

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

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

[#27758](#) fugas por autoreleasepool en Metal

Carga perezosa: [#27794](#) TENSOR\_READ\_LAZY · [#27837](#) endurecimiento · [#27881](#) flag en bench · [#27969](#) renombrado a --lazy-mode

[#27795](#) tope de memoria de trabajo en quantize

[#27742](#) arquitectura Qwen3.8-Flash-Next · [#27880](#) reducción de graph splits · [#27453](#) Lightning Indexer de DeepSeek V4 en Vulkan · [#27843](#) OpenVINO 2026.3.1, whisper.cpp, Qwen3.5 en NPU

[#27696](#) versión 0.3.0 · [#27626](#) server: tool calls del mensaje final · [#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

[Changelog de la API de OpenAI](#) — apagado de la Assistants API y retiradas de modelos de transcripción, 26 ago.

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

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

Blog de Hugging Face: [Open ASR Leaderboard](#) (28 ago) · [encoders multivector](#) (26 ago) · [Quantization-Aware Healing](#) (25 ago) · [notas de construcción de Granite 4.2](#) (25 ago) · [workflows de Gradio](#) (25 ago)

[2608.23283](#) informe técnico de Apodex 1.1 · [Apodex-1.1-mini](#) (pesos del 17 ago, fuera de ventana)

Fuera de ventana, citado como contexto: [2608.20953](#) artículo de Quantization-Aware Healing (21 ago).

#### PUBLICACIONES §4

[Qwen3.8-Flash-Next](#) · [GLM-5.3](#) · [GLM-5.3-Flash](#) · [Hy4-preview](#) · [GLM-5.3-Flash-DFlash2](#) — las especificaciones y cifras de benchmark son las declaradas en las model cards.

[Hugging Face daily papers](#) — usado para descubrimiento a lo largo de la ventana; cada paper que aportó se leyó en su abstract de arXiv antes de citarse.

---

Compilado el 31 de agosto de 2026 a partir del feed de releases y la API de GitHub de llama.cpp, la API de arXiv filtrada a esta ventana, los daily papers curados de Hugging Face y — novedad de este número — las salas de prensa y changelogs de los propios laboratorios. Las cifras citadas como aceleraciones, tasas de error o puntuaciones de benchmark son los números reportados por sus autores o fabricantes en sus propios benchmarks — trátalas como afirmaciones, no como mediciones, hasta que las reproduzcamos. La ventana se abre el mediodía del 23 de agosto, donde se detuvo la compilación del número 02; todo lo fechado fuera de ella está etiquetado como contexto. Los términos se definen en el [glosario](#).

---