DIGISHIELD KIDS

OPEN-SOURCE AI INTELLIGENCE NOTE | NO. 2

The Floor Has Dropped

Hardware thresholds for capable open-weights inference, December 2025 to June 2026

DigiShield Kids | Short Report

DigiShield Labs Ltd | York, United Kingdom

June 2026

Open-weights models run outside any service layer once downloaded. Abliterated variants circulate through the same distribution channels as base models, often within days of major releases. Hardware and deployment friction have therefore become central access constraints. Over the last six months, the practical floor for locally useful open-weights inference has moved sharply downward: from enthusiast GPU setups, to ordinary laptops for multimodal models, to phone-class deployment for compact agentic models.

FINDINGS

Three thresholds, six months

The table below documents three discrete drops in the minimum hardware required to run a locally useful open-weights model. "Locally useful" means fit for instruction-following, local assistant, agentic, coding, or multimodal workflows; not necessarily frontier-equivalent. Publicly reported baselines are cited; DigiShield Kids assessments on abliterated derivative availability are our own and identified as such.

THRESHOLD	HARDWARE	MODEL / EVIDENCE	ASSESSMENT
Laptop-class	16 GB VRAM or unified memory. Consumer laptop class; no datacentre GPU required.	Gemma 4 12B (Google, 3 June 2026). Runs text, image, and audio on a standard 16GB laptop under Apache 2.0. LM Studio and Ollama shipped same-day support [1, 2].	Abliterated derivatives assessed as available within days of release, consistent with established lag pattern across prior frontier releases.
Phone-class	~0.5 GB RAM. Runs fully offline on a mainstream Android or iOS device	MiniCPM5-1B (OpenBMB, May 2026). Q4-quantised GGUF circa 500 MB. Developer shipped a local companion application on the same model [3, 4].	At this size class, unrestricted quantised variants are the predictable next packaging step. The inference engine is already in the consumer product.
Frontier-on-budget	Single consumer GPU + second-hand Intel Optane DIMMs. ~380 GB total, slow token generation	1T-parameter open-weights model run on a single-GPU workstation using commodity Optane memory, reported May 2026 [5, 6]. Exotic build; not mass-access.	Demonstrates no institutional compute requirement for frontier-class private inference. Relevant to threat actors; not relevant to casual access.

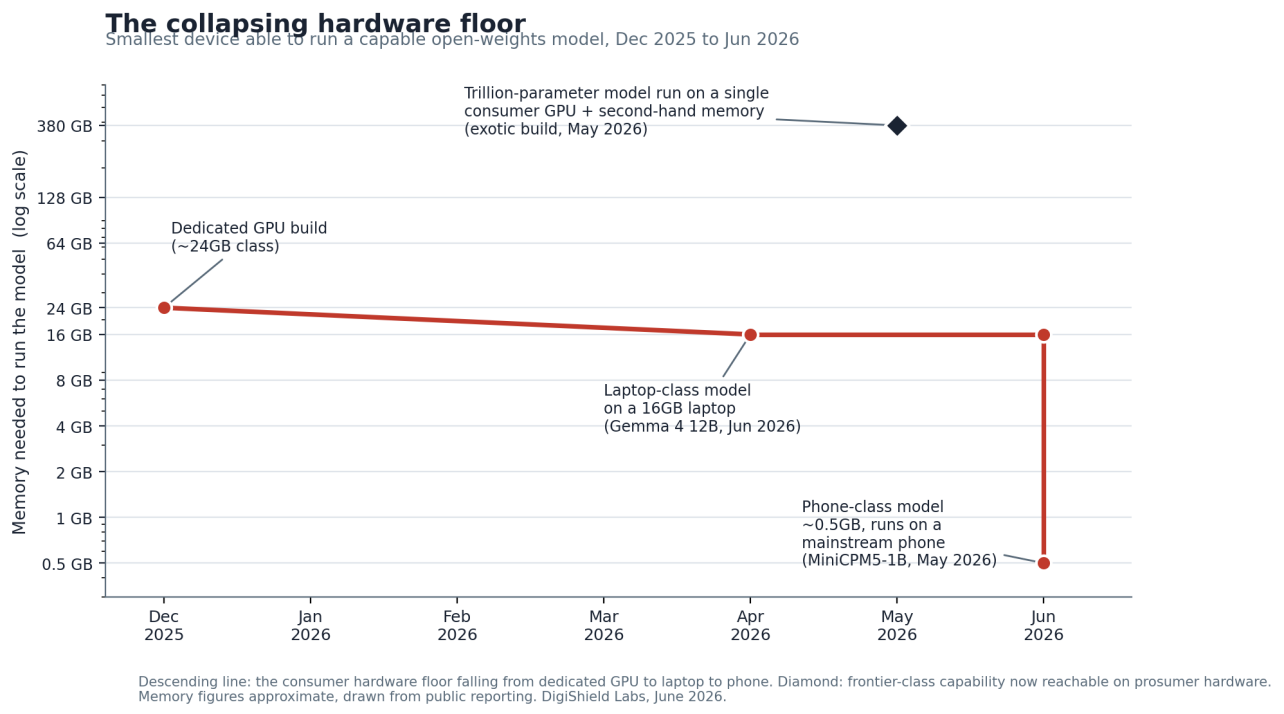

Fig. 1 — Minimum hardware floor for locally useful open-weights inference, Dec 2025 to Jun 2026. The descending line tracks the consumer floor; the diamond marks frontier-class capability on prosumer hardware. Markers are not capability-equivalent; they show successive drops in the minimum hardware needed for locally useful inference at different model tiers. DigiShield Kids, June 2026.

CONTEXT

Deployment packaging

The hardware compression is one half of the access problem. The other is toolchain friction. llama.cpp, Ollama, and LM Studio have reduced local model deployment to a download and a configuration file; GGUF has become the de facto consumer packaging format for quantised weights. In late May 2026, Felix Kjellberg (PewDiePie, 110M+ subscribers) open-sourced Odysseus, a local-AI workspace designed for personal hardware deployment with no telemetry, no account requirement, and no API dependency [7]. The project had attracted 68,800+ GitHub stars by 12 June 2026. That announcement did not reach a specialist audience. For most users, neither the hardware nor the toolchain now represents a meaningful barrier.

REGULATORY ANALYSIS

The service-layer problem

UK regulatory reach in the AI safety context is built around the service layer. The Online Safety Act reaches user-to-service relationships; the ICO Children's Code and age-assurance guidance reach organisations providing digital services to children. None of this architecture straightforwardly governs a GGUF file already running in llama.cpp, Ollama, LM Studio, or a local wrapper on a user's own device. There may be duties at the hosting, distribution, or platform layer, but the offline inference layer itself remains largely outside the current safety perimeter.

No statutory body currently holds clear jurisdiction at the model distribution layer. The tractable intervention points are upstream of inference: the model distribution layer (HuggingFace and equivalents), the abliteration toolchain, and the consumer packaging products now reaching mass audiences. None of these sits clearly within the current UK regulatory perimeter. Establishing which body should hold jurisdiction, and on what statutory basis, is the policy question this shift is now forcing.

SCOPE NOTE

What this report does and does not contain

Specific abliterated model identifiers, author pseudonyms, and repository locations are withheld. Base models named here are commercially released products covered in mainstream technology press. The cited sources substantiate hardware and capability claims only. DigiShield Kids intelligence on the abliterated derivative ecosystem is available through the DSI subscription tier.

REFERENCES

Sources

[1] "Google's new open source Gemma 4 12B analyzes audio, video and runs entirely locally on a typical 16GB enterprise laptop," VentureBeat, 3 June 2026.

[2] "Google Gemma 4 12B Brings Multimodal AI to 16GB Laptops, Free Under Apache 2.0," TechTimes, 4 June 2026. Gemma 150M downloads figure attributed to Google.

[3] "This Half-Gigabyte AI Model Runs Local Agents on Your Phone," Decrypt, May 2026.

[4] OpenBMB / MiniCPM project documentation and release notes, 2026.

[5] "768GB of cheap Intel Optane DIMM memory sticks used to run 1-trillion-parameter LLM on a system with a single GPU," Tom's Hardware, May 2026.

[6] "Used Optane memory helps run a 1 trillion parameter AI model on a single GPU workstation," Digital Citizen, May 2026.

[7] Felix Kjellberg (PewDiePie), "MY trillion $Dollar Project is finally OUT!", YouTube, 31 May 2026. youtube.com/watch?v=rAzT5lcezPs

Cited sources substantiate hardware and publicly reported model specifications only. DigiShield Kids assessments on abliterated derivative availability are identified as such throughout and are not sourced to these references. Where possible, primary sources (Google launch blog, Google developer guide, OpenBMB GitHub, Odysseus GitHub) should be substituted for secondary press citations before final publication.

DigiShield Kids is an independent child safety and AI research project. DigiShield Labs Ltd is the registered organisation, York, United Kingdom.

Detailed intelligence on the abliterated model ecosystem, character card deployment pipeline, and UK regulatory gap analysis is available through the DigiShield Intelligence subscription service. Enquiries: intelligence@digishieldlabs.com