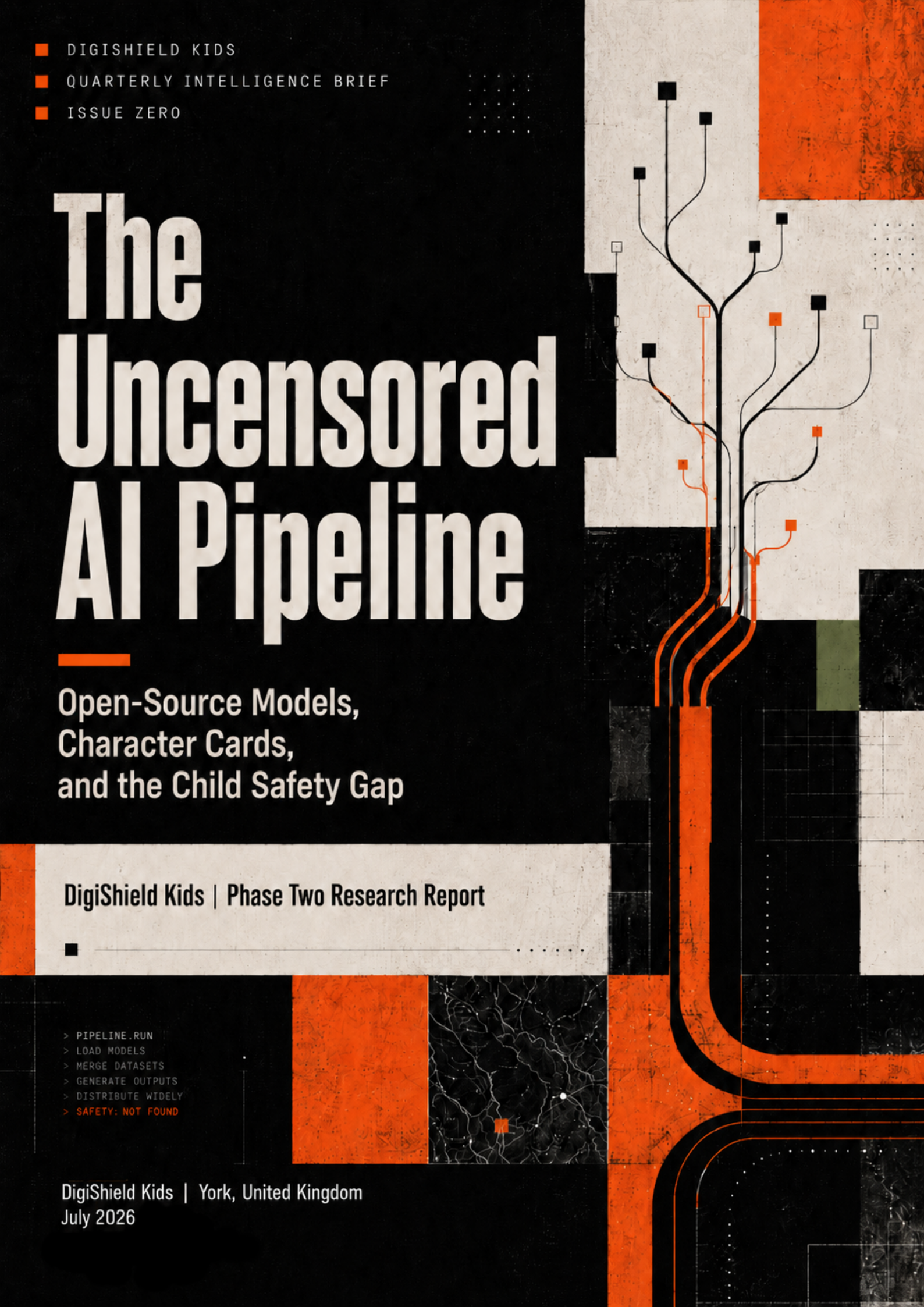
DIGISHIELD KIDS
QUARTERLY INTELLIGENCE BRIEF
ISSUE ZERO

The Uncensored AI Pipeline

Open-Source Models, Character Cards, and the Child Safety Gap

DigiShield Kids | Phase Two Research Report



DigiShield Kids | York, United Kingdom
July 2026

DIGISHIELD KIDS
QUARTERLY INTELLIGENCE BRIEF
ISSUE ZERO

The Uncensored AI Pipeline:

Open-Source Models, Character Cards, and the Child Safety Gap

DigiShield Kids | Phase Two Research Report
Greymark Research Ltd | York, United Kingdom
Published 7 July 2026
Authored by Dr Andrew Langley
DOI: 10.5281/zenodo.21240874

STATUS
This is the published version of this report, issued after close of the responsible disclosure window on 3 July 2026. Ahead of publication, DigiShield Kids notified the relevant regulatory bodies, platform operators, and frontier model developers of the findings and gave each a response period. The disclosure log at Annex B records each recipient, the date of notification, and the response received. Core ecosystem counts and pipeline-derived quantitative findings are derived from public APIs and publicly accessible metadata. Supplementary figures, including traffic estimates, hardware pricing, public-source dates, and legal references, are drawn from the public sources cited in the relevant sections. No private communications, child user data, or non-public personal data were collected. Where public account identifiers or repository names were processed, they were handled under data-minimisation principles for child-safety research and regulatory disclosure.

Executive Summary

This report presents the findings of DigiShield Kids Phase Two research into the open-source AI model ecosystem and its implications for child safety. It is the founding document of the DigiShield Quarterly Intelligence Brief, a series that will track ongoing developments in this space. Phase One (No Guardrails, March 2026) documented the failure of consumer-facing AI chatbot platforms to protect children from harmful content. Phase Two examines the infrastructure behind those failures: the supply chain of uncensored AI models, the character card deployment ecosystem, and the technical mechanisms that together constitute a child safety risk with no current regulatory answer.

Three findings define this report.

Finding 1: Scale that existing research significantly underestimates

DigiShield's HuggingFace intelligence pipeline, run on 18 May 2026, identified 22,117 tier-qualified model records associated with the uncensored and abliteration ecosystem, including 8,732 high-confidence records, carrying 587 million cumulative API-reported downloads. A September 2025 academic study identified 8,608 such repositories. A 2.6x increase in eight months is the growth rate of an industry.[R1]

Finding 2: There was never a window

Regulatory frameworks that assume AI developers can implement safety measures before harmful versions reach the public are working from a false premise. Across a panel of 57 frontier model releases from 2023 to 2025, the median time between a model's release and the appearance of an abliterated variant on HuggingFace was 13 days. In 60% of cases an abliterated variant appeared within 14 days; in 28% of cases within 7 days. The same-day case is Gemma 2 2B (lag = 0). The window that regulation assumes exists has never reliably existed.[R2]

Finding 3: Agentic automation has removed the last practical friction

As of early 2026, the process of removing safety alignment from a frontier AI model can be initiated with eight one-word prompts to an autonomous AI agent. The agent handles abliteration, benchmarking and upload to public repositories without further human input. What required specialist technical skill in 2023 can now be initiated by a low-skill actor using agentic tooling.

These three findings, taken together, describe a supply chain operating at scale, accelerating in speed, and becoming progressively easier to access. The consumer platforms documented in Phase One sit at the end of that chain, not at its origin. Regulatory focus on consumer platform compliance alone addresses symptoms rather than causes.

The report closes with a regulatory analysis covering what current UK law reaches and, more specifically, what it does not. The most tractable near-term enforcement lever is payment rail intervention at the commercial inference tier, not content moderation at the consumer platform tier. The report's recommendations reflect this distinction.

The model access pathways and platforms documented in this report are child-accessible in the configurations tested. DigiShield confirmed that at least one access route in each relevant pathway could be reached from a standard UK consumer device without verified age assurance. Some commercial inference pathways involve payment or token access; the

child-accessibility finding concerns the absence of meaningful age verification, not the absence of monetisation across every configuration.

Core ecosystem counts and pipeline-derived quantitative findings are derived from public APIs and publicly accessible metadata. Supplementary figures, including traffic estimates, hardware pricing, public-source dates, and legal references, are drawn from the public sources cited in the relevant sections. Detection methodology is deterministic and fully reproducible. Findings that remain under active legal review are not included in this publication.

This report is an evidential account of what DigiShield's monitoring found. DigiShield presents its findings as a witness to the landscape described; it does not issue regulatory determinations or assert legal conclusions. Where statutory provisions are discussed, the analysis identifies questions that, in DigiShield's reading, appear raised by the evidence. Those questions are for Ofcom, DSIT, and, where relevant, the courts to resolve.

Terms used in this report

Abliteration. The process of removing safety alignment from an open-weight AI model by targeting and suppressing the model's internal refusal direction vector. Abliteration removes content restrictions across all subject areas simultaneously and leaves no observable signature in the model's public metadata.

Uncensored model. An open-weight language model from which safety alignment has been removed by abliteration or equivalent means. In this report, uncensored models were identified through their explicit self-documentation in public metadata on HuggingFace.

Tier-qualified corpus. The 22,117 model records meeting DigiShield's detection criteria for analysis. Of 22,836 total pipeline entries, 719 collected through author attribution carry no uncensored detection signal in their own metadata and are excluded from all analysis modules in this report.

Open-weight model. An AI model whose underlying parameter weights are publicly released and downloadable. Open-weight models can be run, modified, and redistributed without the permission of the original developer. This report uses 'open-source' where it reflects common sector usage or source terminology, but the technical risk described is more precisely an open-weight risk: public release of downloadable model weights that can be modified, quantised, redistributed, and run locally.

Character card. A structured text file containing a persona definition, scenario instructions, example dialogue, and a system prompt that shapes how an AI model behaves during conversation. Character cards define the instruction layer; they are not the AI model itself.

Kill chain. The end-to-end sequence of stages from a frontier model's release to child access. An intelligence-sector term for a full pathway, used here to make the ecosystem legible as a regulatory map.

Local inference. Running an AI model on a local device (a personal computer or mobile phone) without sending data to a remote server. The model weights, prompt, and outputs remain on the user's device; no network traffic is generated.

Model record. A public HuggingFace model repository or derivative entry captured by the pipeline. It does not necessarily denote a unique base model or family; the corpus includes quantisations, forks and derivatives. Where this report says "models" in summary counts, it refers to model records unless a distinct base model is specified.

Abliteration lag. The time elapsed between the public release of a base model and the first abliterated version of that model appearing on HuggingFace. DigiShield tracks abliteration lag as an indicator of how quickly safety controls on newly released models are circumvented.

CSAM. Child Sexual Abuse Material: images or video depicting the sexual abuse or exploitation of a child, the generation, possession, and distribution of which is a criminal offence in the United Kingdom.

Quantisation. A compression technique that reduces an AI model's file size and hardware requirements by representing its parameters at lower numerical precision. Quantised versions of uncensored models are identified by author-attribution methodology in DigiShield's corpus.

Section 1: The Problem Framing

What Phase Two adds

Phase One of DigiShield's research programme tested 59 consumer AI chatbot platforms against a standardised set of child safety criteria. The headline findings: 85 per cent bypass rate across platforms, 91.5 per cent rated Poor or Critical, 517 shared boilerplate passages across nominally independent safety policies. Phase One established the consumer-facing failure.

Phase Two asks, where do the capabilities enabling those failures originate? The consumer platforms documented in Phase One generally sit downstream of model providers, inference services, and character-card ecosystems. They draw on an open-source ecosystem in which safety alignment is routinely removed and models are made available for download at no cost, with limited practical downstream enforcement of licence or terms-of-use restrictions once weights are copied, modified, or redistributed. Understanding child safety risk in AI chatbots requires understanding that ecosystem.

DigiShield is not aware of any public child-safety-focused mapping of this supply chain using deterministic, reproducible methodology. This report documents what that mapping found.

The kill chain

The following table represents the full supply chain from frontier model release to child access. The chain bifurcates at Stage 5 into two parallel inference pathways: local inference (Stage 5a), documented in this report, and commercial anonymous inference (Stage 5b). Each stage is documented with empirical evidence in the sections that follow, except Stage 5b.

01	**Frontier Release**	A major AI laboratory publishes a large language model under an open or permissive licence. Safety alignment is implemented at training time.
02	**Abliteration**	Within hours to days, the safety alignment is removed using published techniques. Median lag across a 57-model panel: 13 days; 60% within 14 days, 28% within 7 days. Same-day case: Gemma 2 2B (google/gemma-2-2b).
03	**Quantisation**	The abliterated model is compressed into GGUF format for consumer hardware. 229,716 GGUF files now indexed on HuggingFace. 6,633 models run on standard laptops. 2,823 on smartphones.
04	**Character Card Deployment**	Uncensored model capabilities are packaged into roleplay personas via character cards. Cards carry system prompts, personality definitions, and interaction rules. 252,521 cards on Chub.ai alone.
05a	**Local Inference Branch**	Cards are loaded into consumer frontend applications (SillyTavern, Ollama, KoboldCPP). The model runs locally on the user's own hardware with no network traffic and no platform oversight. This branch is the one documented in detail in this report.
05b	**Commercial Anonymous**	The same models are also served by a parallel commercial inference layer: decentralised compute networks, token-gated inference

06	**Inference Branch (Phase Three focus)**	services, and anonymous-tier operator services. These pathways can operate without consumer-facing content moderation and often without a single responsible operator at any regulatory threshold; payment can be made anonymously through token rails outside the card-rail enforcement perimeter. This branch is flagged here as a forward indicator and is the focus of DigiShield's Phase Three research.
	Child Access	A child accesses the system through a consumer platform with email-only registration, no age verification, or by running a downloaded model locally. No persistent intermediary enforcement surface exists at the point of local inference.

Fig.1. Documented kill chain from model release to child access through a chatbot platform. Each stage presents a different regulatory surface area and a different risk-analysis.

The regulatory implication of this structure is that enforcement at any single stage is insufficient if the stages before it remain unaddressed. A child safety requirement imposed on consumer platforms (Stage 5) does not affect the availability of uncensored models at Stage 2, the proliferation of GGUF files at Stage 3, or the character card deployment layer at Stage 4. Each stage is a distinct regulatory surface with distinct enforcement challenges.

What this report does not include

Two categories of finding from the Phase Two research programme are not included in this publication. The first concerns a specific intersection of character card content with signals relating to minors; that finding is subject to ongoing legal review and will be published through a separate process coordinated with the relevant authorities. The second concerns named commercial operators whose identification is contingent on a preliminary legal assessment currently underway. Where those operators are relevant to the regulatory analysis, they are described by their operational characteristics rather than by name.

The DigiShield Quarterly Intelligence Brief provides ongoing intelligence depth on actors, operational detail, and monthly pipeline updates that fall outside the scope of this founding document.

Section 2: The Supply Tier (HuggingFace Intelligence Pipeline)

Methodology

DigiShield built an automated intelligence pipeline querying the HuggingFace public API using deterministic keyword, tag, and author matching. The detection logic identifies model records whose metadata carries one or more of the following signals: use of the terms "uncensored", "abliterated", "decensored", "heretic", or "no-censorship" in model card text; known abliteration tags; or attribution to authors whose prior output has been manually verified as producing uncensored models. The methodology is fully documented and reproducible. No probabilistic or LLM-based classification is used; this is a deliberate design choice to ensure that findings can withstand methodological scrutiny in regulatory submissions.

The pipeline was last run on 18 May 2026. Corpus (tier ≥ 1): 22,117 model records.

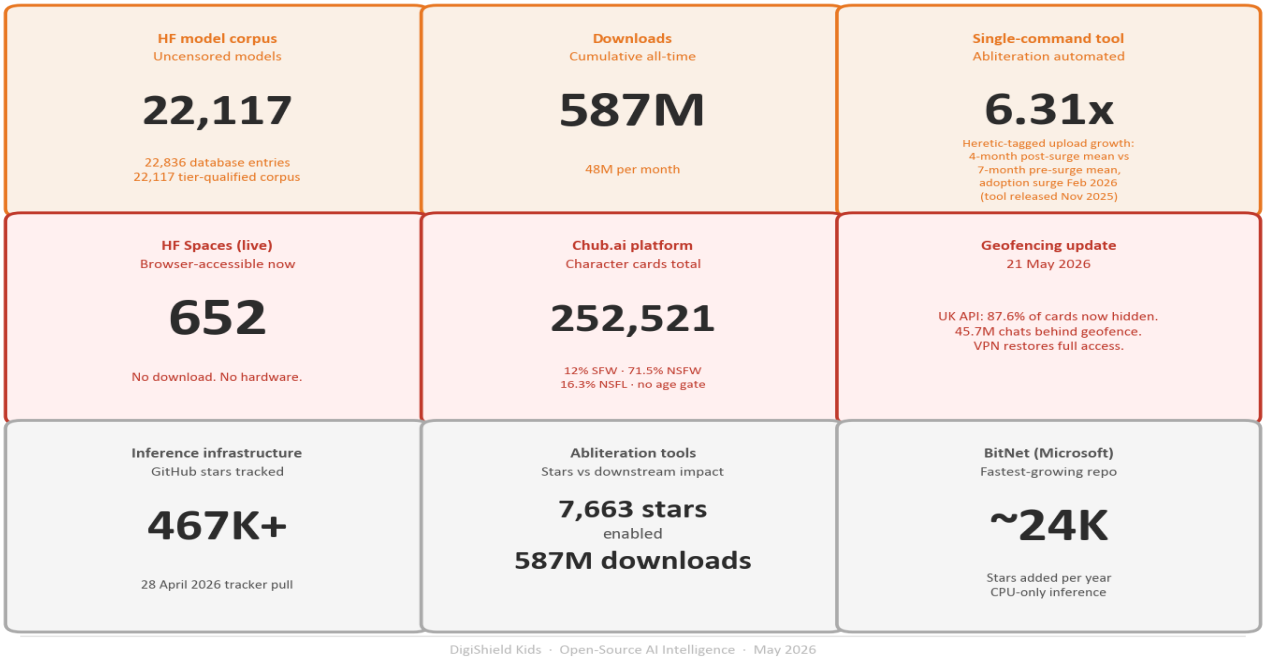


Figure 2: Dashboard of the open-source and uncensored LLM ecosystem at the time of research

Scale

Metric	Figure
Model records in tier-≥1 corpus	**22,117**[1]
Tier 1 (high confidence)	**8,732**

[1]The pipeline database holds 22,836 entries; 719 were collected via author attribution but carry no uncensored detection signal in their own metadata and are excluded from every analysis module. All figures in this report are drawn from the 22,117 tier-qualified corpus.

Metric	Figure
Tier 2 (medium confidence)	**104**
Tier 3 (GGUF quantisations and lower confidence)	**13,281**
Distinct authors	**2,433**
Cumulative downloads (all tiers)	**586,888,082**
Downloads in last 30 days	**47,592,990**
Tier 1 model records: 30-day downloads	**30,308,161**
Consumer-hardware GGUF files	**229,716**
Models runnable on standard laptops	**6,633**
Models runnable on smartphones	**2,823**

All HuggingFace figures updated to 18 May 2026 pipeline run.

Download figures are API-reported cumulative totals and may not capture direct file downloads that bypass the HuggingFace API. DigiShield did not download any model documented in this report; all corpus data was collected through public metadata APIs.

A September 2025 academic study (MDPI) identified 8,608 uncensored model repositories on HuggingFace. DigiShield's April 2026 pipeline pull recorded 19,428. The 18 May 2026 rerun finds 22,117 in the corpus (tier ≥ 1), a 2.6x increase over the September 2025 MDPI baseline of 8,608. The rate of growth is itself a finding; the ecosystem is not static.

The HuggingFace CEO has independently confirmed a figure of 176,000 GGUF files and approximately 9,700 new models added per month.[R4] DigiShield's pipeline count of 229,716 GGUFs is consistent with that figure given the detection scope differences between the two counts.

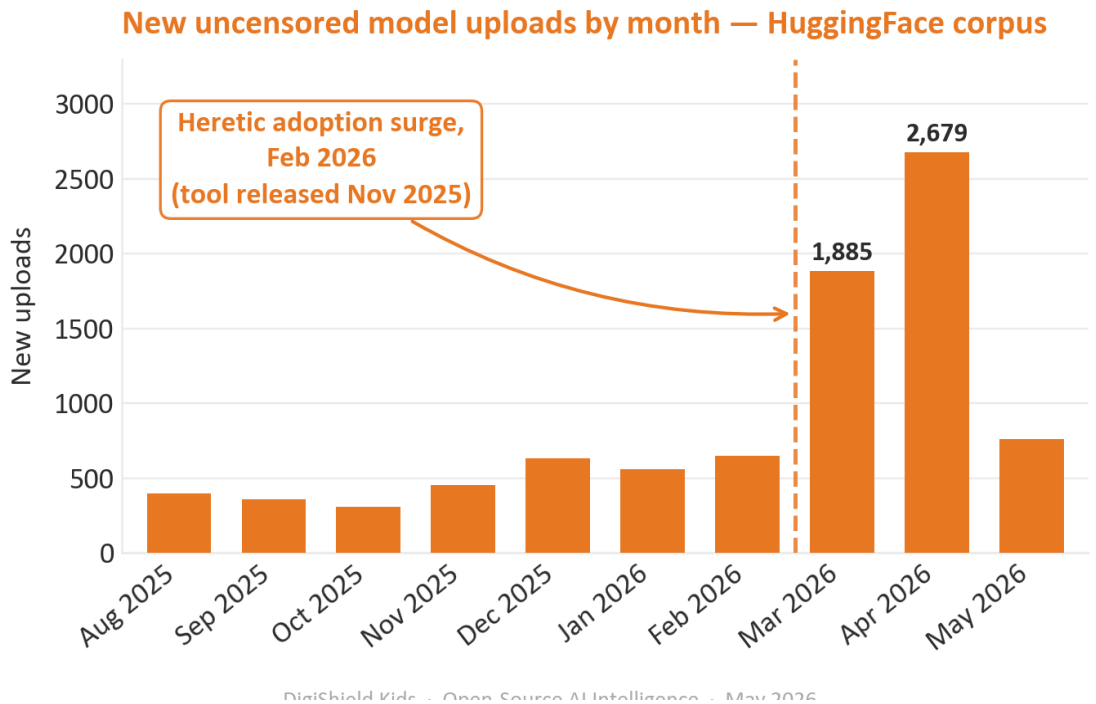


Figure 3: Abliterated or uncensored model records uploaded to Hugging Face from Aug 2025 to May 2026

The Heretic effect

Heretic, a tool that automates abliteration to a single command-line instruction and requires no understanding of model internals, was released in November 2025.[2] Adoption accelerated sharply from February 2026, a surge independently corroborated by the AI safety group Alice, whose research dates Heretic's rise in GitHub popularity to that month.[3]

In March 2026, DigiShield recorded 1,885 new uncensored model uploads, 3.07 times the trailing three-month mean. The timing and the tag-level pattern track this adoption surge feeding into the supply curve. Concurrent open-weight base model releases and broader quantisation trends cannot be excluded as contributors to the aggregate count, so the March spike is reported as correlating with the adoption surge without being claimed as its sole cause.

Measured on the narrower basis of Heretic-tagged repositories alone, the effect is sharper: mean monthly uploads of Heretic-matched models ran 6.31 times higher in the four months from the February 2026 adoption surge than in the seven months before it, 576 per month against 91[R1].

The seven-month window before the surge includes months preceding Heretic's November 2025 release, when few or no Heretic-tagged models existed. The pre-surge figure therefore reflects a near-absent prior state, and the comparison is best read as the ecosystem moving from close to zero to 576 uploads per month once the tool was in wide use.

By the time of the pipeline run, 2,298 model records carried the "heretic" tag in their metadata. 3,521 carried both "uncensored" and "abliterated". The tag vocabulary documents how openly the ecosystem describes its own purpose.

Metric	Figure
"uncensored" tag	**4,492 model records**
"abliterated" tag	**3,806 model records**
"heretic" tag	**2,298 model records**
"decensored" tag	**1,587 model records**
Carry both "uncensored" and "abliterated"	**3,521 model records**

An independent investigation by the Financial Times and the AI safety group Alice, published on 25 May 2026, demonstrated the same capability from the outside. Using Heretic, the FT removed the safety alignment from Meta's Llama 3.3 in under ten minutes on a standard laptop, and a decensored Gemma 3 produced instructions for dispersing chlorine gas in a crowded space, code to steal payment-card data, and material describing child sexual abuse.[4] The finding confirms in a national outlet that the supply-tier capability documented here is neither theoretical nor difficult to access.

Safety language analysis

DigiShield analysed model card text across 20,890 model records with parseable cards.

[2]Heretic (heretic-llm) v1.0.1, first public release 16 November 2025. Source: PyPI release history (pypi.org/project/heretic-llm) and GitHub releases (github.com/p-e-w/heretic/releases).
[3]NPR, "This AI tool strips safety guardrails from chatbots in seconds. A researcher says it's gone viral," 31 May 2026, citing research by the AI safety group Alice that Heretic's popularity on GitHub rose from February 2026.
[4]Financial Times, "AI guardrails stripped from Meta and Google models in minutes," 25 May 2026 (investigation conducted with the AI safety group Alice); syndicated via the Irish Times, 25 May 2026.

Metric	Figure
No safety language present	**82.3%**
Any use restriction documented	**1.8%**
Disclaim responsibility for outputs	**5.5%**
Advertise zero refusals (documented test suites)	**≥71 model cards (Apr 2026 run)**
Openly document the abliteration process	**17.6%**

At least 71 model cards advertise zero refusals against a 465-prompt test suite (April 2026 behavioural-test run). They are significant. They are not simply silent on safety; they actively market the removal of safety alignment as a product feature, documented against a test benchmark.

Author concentration

2,433 distinct HuggingFace accounts appear as repository authors for the 22,117 model records. The three most active accounts by model count together account for approaching half the corpus total (approximately 48 per cent), a concentration the data shows is driven primarily by quantisation activity rather than novel capability development. The dependency graph between these model records contains 7,788 distinct base-model references, and the single most prolific author dominates it, appearing as the source of a large share of those dependency relationships. Two emergent actors not on DigiShield's original seed list account for substantial monthly download volumes. The first publishes 22 models with approximately 5.5 million monthly downloads; every model in this catalogue scores 0 of 465 refusals against the standard test suite. The second publishes 13 models with 3.17 million monthly downloads, including the single most-downloaded Heretic variant in the corpus. Author concentration is most pronounced within the Heretic-tagged subset: the top five accounts produce 59.1 per cent of all Heretic variants, and a single prolific account alone produces 45 per cent. Named actor profiles are reserved for the DigiShield Quarterly Intelligence Brief.

Base model targeting

The ecosystem does not produce its own base models; it modifies models released by frontier AI developers. The three most targeted model families:

- Qwen (4,403 variants, 15.9 million 30-day downloads)
- Llama (3,730 variants, 141 million cumulative downloads)
- Gemma (1,637 variants since February 2024, the fastest-growing family)

Within the Gemma family, the same-day case is Gemma 2 2B (google/gemma-2-2b → IlyaGusev/gemma-2-2b-it-abliterated, lag = 0). Gemma family median across 6 models with a confirmed fork is 14 days, range 0 to 46 days.

Hardware accessibility

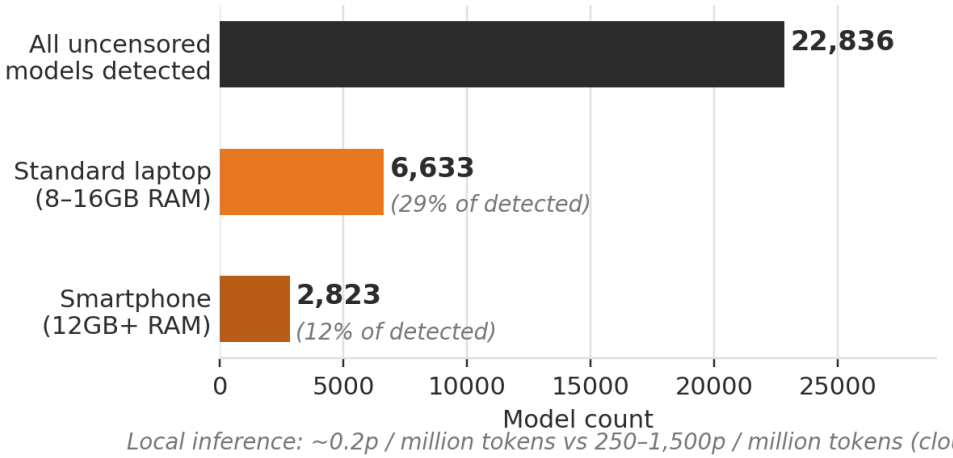
Electricity cost for local inference spans a wide range. A smartphone running a small model costs a few pence per million tokens; a single desktop GPU running a large model, a few hundred pence; a multi-GPU rig running a 70-billion-parameter model can exceed a

thousand pence. Local inference is cheaper than cloud APIs for small models and comparable to or dearer than budget API tiers for large ones. The decisive advantage of local deployment lies elsewhere: in the absence of platform oversight, with no usage log, no content-policy enforcement, and no mechanism for a platform operator to observe or interrupt a session. The hardware barrier is falling.

- Across the tier-qualified corpus, 6,633 model records run on laptop-class hardware and 2,823 on a smartphone; 1,576 require workstation or multi-GPU hardware[R1]
- NVIDIA DGX Spark: priced at $4,699 (2026; $3,999 at October 2025 launch), 128GB unified memory, capable of running a 70-billion-parameter model on a desk
- Home lab configuration (four RTX 3090 GPUs, used market): approximately £4,000–5,000, capable of 70B Q8 inference (based on used RTX 3090 units at approximately £600–800 each plus a host build)
- Flagship smartphones with 12GB or more RAM run 7-billion-parameter models at up to 15–30 tokens per second; performance falls under thermal throttling in sustained use

The regulatory implication is that "local inference" is no longer a theoretical threat surface. It is the current state. A 70-billion-parameter uncensored model can run on consumer hardware available for under £5,000, with no network traffic and no platform intermediary. No persistent intermediary enforcement surface exists at the point of local inference.

Hardware prices and performance figures are indicative snapshots, not stable benchmarks.



Local inference: ~0.2p / million tokens vs 250–1,500p / million tokens (cloud API)

Figure 4: Hardware accessibility of uncensored models — HuggingFace pipeline, 18 May 2026. Laptop: 8B–13B Q4/Q5 on 16GB RAM. Smartphone: 7B Q4 on 12GB+ RAM.

Section 3: The Deployment Tier (Character Card Ecosystem)

The character card platforms documented in this section are child-accessible: the configurations DigiShield tested required no age verification, no account creation with verified age, and no payment.

What character cards are

A character card is a structured text file containing a persona definition, a personality description, scenario instructions, example dialogue, and a system prompt that shapes how an AI model behaves during conversation. Character cards are not the AI model; they are the instruction layer that defines how a model presents itself and what it will and will not do within a roleplay interaction. When a card is loaded into a local frontend application alongside an uncensored model, the resulting interaction is governed by the card's instructions, not by any platform safety layer.

Character cards are the mechanism by which uncensored model capability is packaged for consumer use. They are the interface between the supply tier documented in Section 2 and the end user. They are also, in regulatory terms, almost entirely unaddressed.

The Chub.ai pipeline

DigiShield built a second automated intelligence pipeline querying the Chub.ai public API. Chub.ai is one of the largest publicly accessible character card repositories and the largest for which a public API exposes card-level metadata. The pipeline captured card metadata, tags, engagement figures, and content classifications.

Metric	Figure
Platform total character cards (20 April 2026)	**252,521**
SFW cards: platform	**12.1%**
NSFW-only cards: excl. NSFL (platform)	**71.5%**
NSFL-rated cards (most severe category)	**16.3%**
UK-visible cards post-geofencing (21 May 2026)	**31,254 (−87.6%)**
Total chats: sampled corpus (~14,899 cards)	**46,500,000**
Total messages: sampled corpus	**665,000,000**
Registration requirement	**Email address only**
Age verification	**None**

The engagement figures require context. The 14,899-card sampled corpus carries 46.5 million chats and 665 million messages, the accumulated interaction history for the cards DigiShield analysed. These figures derive from the sampled corpus, not the full platform population of 252,521 cards; platform-wide totals are not available from the anonymous API. They are nonetheless not small-scale hobbyist figures.[R3] Chub.ai operates at consumer scale, with email-only registration and no age verification mechanism. A contractual prohibition on under-18 use is not the same as verified age assurance. DigiShield’s finding

concerns the absence of age verification in the tested access flow, not the absence of age-restrictive language in platform terms.

A second deployment-layer finding concerns jailbreak design within the cards themselves. 21.4 per cent of cards on Chub.ai use prompt override fields, the mechanism by which a card's author instructs the underlying model to ignore prior safety instructions. Within the deep analytical corpus of approximately 14,899 cards, 256 carry explicit jailbreak terminology in their description or tagline (the visible-text fields; system prompt body text is not retrievable via the anonymous API); the chat volume on those 256 cards stands at 1,130,590. Jailbreak instructions are not a fringe artefact at this platform. They are integrated into mainstream card design.

Between the 20 April and 21 May 2026 pipeline runs, Chub.ai amended its terms and introduced UK geo-blocking of its card catalogue. UK-visible cards fell 87.6 per cent, from the platform total to 31,254. Geo-blocking reduces UK visibility, but it does not remove the content, which remains hosted and available at source. Access-side restrictions of this kind are generally circumventable, although the report does not test or assert any specific method of circumvention. The point is the general one: the content persists, and the enforcement acts on the route to it rather than on the material itself.

Card analysis: structure and harm vectors

Beyond enumeration, the Chub.ai pipeline applies two analytical layers to the post-remediation card sample. The first treats the tag vocabulary as a graph. Tags carried in card metadata are mapped as nodes; two tags share an edge if they appear together on a card. The resulting graph is partitioned using Louvain community detection, an algorithm widely used in network science to identify densely connected sub-structures. The result is a small number of distinct community clusters within the corpus, each with a characteristic tag signature, engagement profile, and content composition. The cards on Chub.ai are not a single uniform NSFW phenomenon; they are a structured ecosystem of distinct sub-communities. Community-level findings are reserved for the DigiShield Quarterly Intelligence Brief and the closed Ofcom briefing; the methodological pattern is the relevant point for this report.

The second analytical layer applies a harm vector taxonomy. Raw keyword sweeps generate counts that are not immediately fit for regulatory citation; the taxonomy refines them. For example, an initial sweep for cards mapping to coercive control returned a population of approximately 5,400 cards. Closer inspection identified a single dominant generic platform-archetype tag producing false-positive co-occurrence across the sample. Removing that one tag from the harm vector definition reduced the population to 1,024 cards with substantive coercive control framing. The same refinement principle is applied across every harm vector before any external citation. Public figures cited in this report or in subsequent DigiShield outputs reflect post-refinement counts, not raw keyword hits. This is the methodological discipline that the pipeline architecture makes possible at corpus scale.

Tag co-occurrence and harm vector outputs together provide the analytical depth that distinguishes the post-remediation Chub.ai dataset from earlier descriptive accounts of character card platforms. The full analytical record sits behind this report and informs the DigiShield Quarterly Intelligence Brief; the methodology is summarised in Annex A.

The consumer platform ecosystem

Chub.ai is one of the largest publicly accessible character card repositories, but it sits within a broader consumer ecosystem. Several platforms in this tier operate at significant scale:

- JanitorAI: approximately 149 million monthly visits (SEMrush, March 2026). Character card platform with NSFW content.

- SpicyChat: removed from the Apple App Store in August 2025. Continues to operate on web.
- CrushOn.ai, Candy.ai: consumer roleplay platforms with NSFW capability.

The App Store removal of SpicyChat is the closest the sector has come to platform-level enforcement in the UK context. It demonstrates that enforcement is possible at the distribution layer; it also demonstrates that web access bypasses it entirely.

The HuggingFace-to-Chub bridge

DigiShield identified 1,788 character cards on Chub.ai whose metadata explicitly references specific HuggingFace model records or frontend applications by name. DigiShield is not aware of a prior empirical link of this kind between the supply tier and the deployment tier: 12 per cent of the post-remediation card sample demonstrates a direct connection to the model supply chain documented in Section 2. The actual proportion is likely higher, since many cards reference model types rather than specific repositories.

The significance of this finding is structural. It demonstrates that the character card ecosystem is not parallel to the uncensored model supply chain but dependent on it. Regulatory interventions at the supply tier (HuggingFace model hosting) would be expected to have measurable downstream effects on the deployment tier (character card platforms).[R3]

A parallel finding concerns explicit frontend cross-referencing within that same 1,788-card population: 513 cards reference Venus, 502 reference JanitorAI, 328 reference SillyTavern. The ecosystem documents its own connective tissue.

Value capture asymmetry

Character card authors monetise their work primarily through Ko-fi and Patreon. Earnings are marginal: the economics of the ecosystem concentrate value at the platform and middleware layer rather than at the content creation layer. Platforms charge for API access, premium features, or subscription tiers. Authors who create the content that drives engagement capture a small fraction of that value. This asymmetry matters for regulatory design: the actors with financial incentive to preserve the current structure are the platforms and middleware providers, not the individual card authors.

During HuggingFace Spaces enumeration, DigiShield identified a specific image generation model deployed across 131 Spaces. User-left comments on the model's HuggingFace card page contain language consistent with use of the model to generate child sexual abuse material. The finding is inferential, drawn from publicly visible user commentary rather than direct testing or model output review. On the basis of that user-comment evidence, DigiShield referred the model and the supporting record to the Internet Watch Foundation in April 2026. Specific operator and model identifiers are withheld from this report; the aggregate finding is recorded here.

Section 4: There Was Never a Window

The phrase "there was never a window" describes the central reframe that Phase Two research produces. It challenges a premise embedded in most current AI safety policy discussion: that frontier AI developers release models with safety alignment in place, and that a period exists during which those safety measures are operative before they are circumvented. The abliteration lag data shows that this premise is false.

The lag dataset

DigiShield assembled a panel of 57 frontier AI base model releases spanning 9 families from 2023 to 2025, tracking the date of the original release and the earliest appearance of an abliterated variant on HuggingFace. Of 69 models attempted, 57 returned a confirmed abliterated fork; the 12 with no fork detected are excluded from lag statistics. This exclusion is a conservative bias: if undetected abliterations exist for any of those 12 models, the true median would be lower, not higher.

Metric	Figure
Base models in panel (fork found)	**57 of 69 attempted**
Period covered	**2023–2025**
Median abliteration lag	**13 days**
Within 7 days	**28% (16/57)**
Within 14 days	**60% (34/57)**
Within 30 days	**68% (39/57)**
Same-day case	**Gemma 2 2B (google/gemma-2-2b → IlyaGusev/gemma-2-2b-it-abliterated)**
Fastest recorded lag	**Same day: lag = 0 (Gemma 2 2B)**

The median of 13 days means that for half the models in the panel, a safety-removed variant was publicly available within two weeks of the original release. More significant than the median is the threshold structure: 28% of the panel had an abliterated variant within 7 days, 60% within 14 days, and 68% within 30 days. The long tail (mean 96.6 days, standard deviation 176 days) reflects a small number of outlier models principally in the DeepSeek and niche categories; the majority of the panel clustered in the single-digit to low-double-digit day range. The same-day case is Gemma 2 2B (google/gemma-2-2b → IlyaGusev/gemma-2-2b-it-abliterated).

The regulatory implication does not depend on lags trending shorter over time. The 57-model panel has a wide tail and a dataset of this size does not support a trend claim. The implication rests on the threshold structure itself: 28% of the panel was abliterated within 7 days, 60% within 14 days, and 68% within 30 days. A substantial fraction of frontier model releases are abliterated faster than any regulatory review cycle could realistically complete.[R2]

What this means for regulation

Regulatory frameworks that rely on a "deploy-then-assess" model for AI safety assume that assessment can happen before widespread distribution of an unsafe variant. The abliteration lag data shows this assumption fails. By the time a regulatory body has received a safety report on a newly released model, abliterated versions have already been downloaded tens of thousands of times.

The 13-day median also means that any responsible disclosure or notification requirement imposed on AI developers would need to trigger within days to weeks of release to have any effect on the downstream supply chain. Current regulatory timelines operate on weeks to months, not days.

The more significant implication is structural: safety alignment at training time, applied by frontier AI developers, does not constitute a durable safety measure for open-weight models. Once the weights are released, the alignment is a feature of the specific instantiation, not a property of the capability. Any policy framework that treats open-weight model release as equivalent to safety-aligned capability deployment is misreading the technical reality.

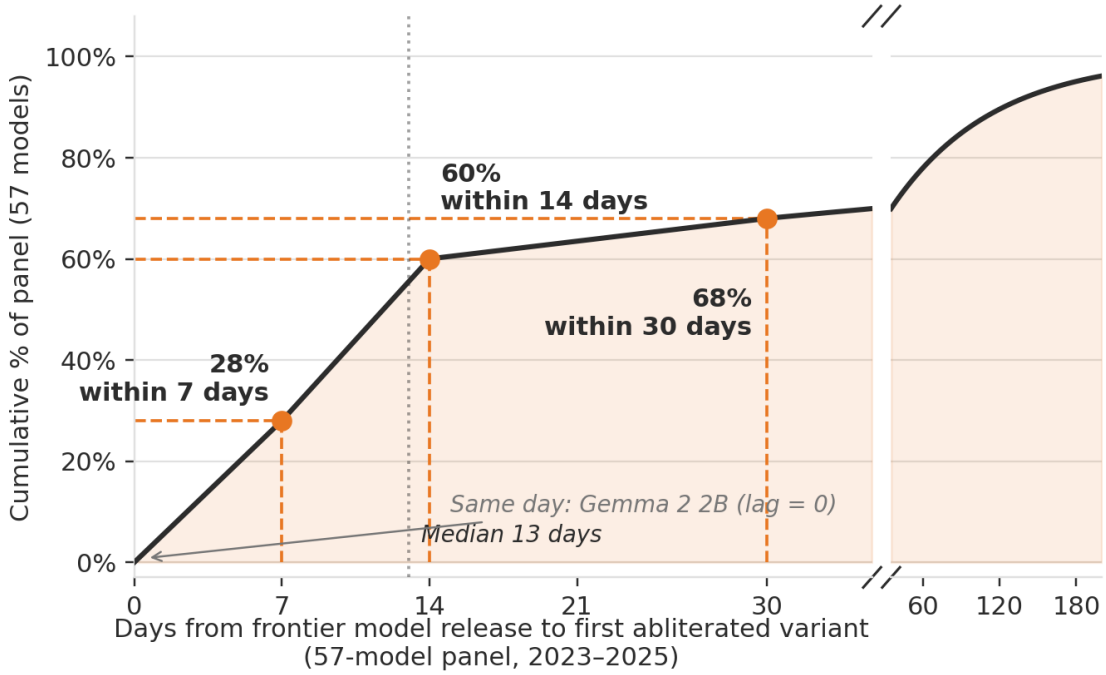


Figure 5: Cumulative abliteration lag distribution. Source: abliteration_lag_summary_20260521.md (abliteration_lag_v2.py). CDF shape illustrative between confirmed thresholds; values 28/60/68% are exact.

The intelligence improvement rate

HuggingFace's CEO has published data showing that the most capable model runnable on a 128GB consumer laptop improved from a score of 10 to 47 on the Artificial Analysis Intelligence Index between May 2024 and May 2026, a 4.7-fold gain in 24 months on fixed hardware.[5] Moore's Law, applied to transistor count, predicts a doubling approximately every 18 to 24 months. The open-source model capability improvement rate is running at approximately double that pace.

The policy implication is that the hardware barrier to running highly capable uncensored models locally is falling faster than any regulatory response timeline. A model that required a

[5]Clément Delangue (HuggingFace), public post, May 2026, citing the Artificial Analysis Intelligence Index. Documented in DigiShield OSINT record OSINT_028.

specialist workstation in 2024 runs on a smartphone in 2026. Enforcement strategies that depend on hardware scarcity as a limiting factor have an expiry date that is already past.[R4]

External corroboration of the lag finding arrived shortly before this report. In the same investigation, Heretic's creator told the Financial Times he had removed the safety alignment from Google's Gemma 4 within ninety minutes of its public release.[6] A single case from the tool's author is illustrative rather than statistical, but it sits at the extreme fast end of the panel above and shows the threshold the lag data describes.

[6]Financial Times, "AI guardrails stripped from Meta and Google models in minutes," 25 May 2026 (investigation conducted with the AI safety group Alice); syndicated via the Irish Times, 25 May 2026.

Section 5: Agentic Collapse

The findings in Sections 2 to 4 document a supply chain that operates at scale and with short lags. Section 5 documents the removal of the last practical friction in that chain: the requirement for human technical skill.

Autonomous abliteration

In April 2026, DigiShield documented a significant finding from open-source intelligence collection. A publicly demonstrated workflow showed that eight one-word prompts to an autonomous AI agent were sufficient to initiate and complete the full abliteration process: model download, safety alignment removal, benchmarking against a standard test suite, and upload to a public repository. The agent performed all steps without further human input.

The documented outcome: a reduction in refusal rate from 98.8 per cent to 2.1 per cent. The model went from refusing almost all harmful prompts to complying with almost all of them. The entire process was automated.[R5]

This finding has a specific implication for regulatory design. Abliteration was previously a process requiring Python programming skill, familiarity with model architectures, and access to appropriate hardware. Those requirements constituted a practical barrier. The eight-prompt autonomous workflow removes that barrier. The population of people capable of producing an abliterated model is no longer bounded by specialist technical skill. It remains bounded by intent, access to tooling and compute, and a repository account, but the skill barrier that previously constrained it is gone.

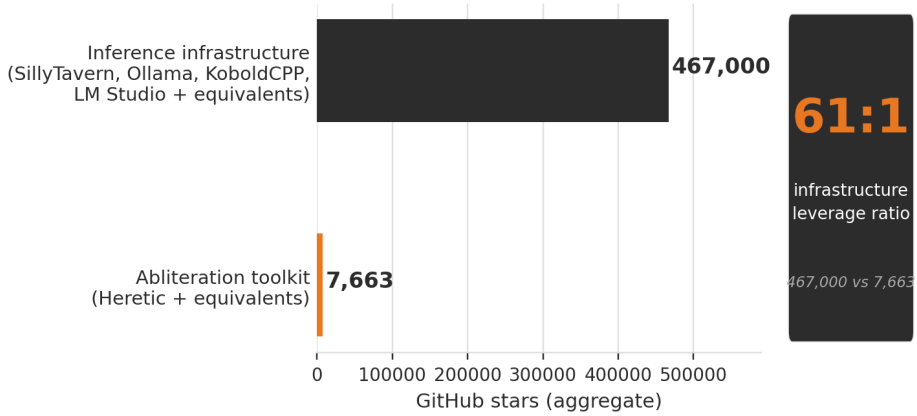


Figure 6: Infrastructure leverage — aggregate GitHub stars for inference-layer tools vs abliteration toolkit. Source: DigiShield GitHub infrastructure scan, April 2026.

Agentic deployment

A second finding concerns the agentic layer above the model supply chain. The largest single account in the relevant uncensored-model subset, by model record count, has released a set of Ollama launcher commands that enable autonomous AI agents to operate using uncensored local models as their execution layer. The specific commands support agentic coding, autonomous browsing, and general task execution.

The historic distinction between "a model exists on a server" and "an autonomous agent is operating on behalf of a user" has collapsed. An uncensored model running locally, combined with agentic launcher tooling available on HuggingFace, constitutes an autonomous agent with no content restrictions and no oversight surface.

Scale here is a structural finding. Aggregate GitHub stars across the seven primary inference-layer repositories tracked by DigiShield, Ollama, llama.cpp, GPT4All, FastChat, BitNet, llamafile, and koboldcpp, total 467,079 as at 28 April 2026. The abliteration toolkit, mergekit and the abliterator repository, registers 7,663 stars in the same authenticated pipeline run. The ratio is 61:1. For every unit of community investment in producing abliterated models, sixty-one units have gone into the infrastructure for running them.[R6]

The hybrid attack chain

A structural finding from the agentic research concerns the use of frontier AI models as the planning layer in a hybrid configuration, with a locally-running abliterated model as the execution layer. The frontier model, constrained by its own safety alignment, can still be used to plan sequences of actions; those actions are then executed by a local uncensored model. The combination bypasses safety restrictions at both layers by separating the planning function (where restrictions apply) from the execution function (where they do not).

This finding is included here at a structural level. Technical detail on specific implementations is reserved for regulatory briefings and the DigiShield Quarterly Intelligence Brief.

Section 6: Multimodal Abliteration

The earliest uncensored open-source models were text-only. The trajectory since has been consistent: as multimodal capability (image, audio, video processing) has been added to frontier models, abliteration has followed.

DigiShield's May 2026 corpus sweep found 1,396 models in the tier-qualified corpus carrying image-text-to-text capability flags, and a further 116 classified as any-to-any, covering image, audio, and video modalities, approximately seven per cent of the corpus in aggregate. The HuggingFace Spaces tier documented in Section 2 includes browser-accessible inference environments that serve abliterated vision-language models, providing pathways that require no local hardware or download.

Image and text

DigiShield documented the availability of MiniCPM-V-4.6, a multimodal model abliterated to remove content restrictions on both text and image inputs. The model accepts image inputs without any content restriction on what those images depict. This capability was available on HuggingFace in GGUF format at the time of the pipeline run.

The child safety implications of unrestricted multimodal models go beyond those of text-only variants. An uncensored model capable of processing and responding to images without content restrictions creates harm vectors that text-only analysis does not capture.

The trajectory

NVIDIA released Nemotron Omni 30B in early 2026: a model capable of processing video, audio, image, and text inputs simultaneously. At the time of the DigiShield pipeline run, no abliterated variant had appeared. The model is a watch item: given the abliteration lag data in Section 4, the absence of an abliterated variant at time of publication should not be taken as evidence that one will not appear.

The broader trend is that multimodal capability in the uncensored open-source ecosystem is expanding faster than safety research can track it. The character card deployment layer has not yet fully adapted to multimodal models, but the technical infrastructure for multimodal local inference is in place.

Section 7: The Commercial Tier

The kill chain in Section 1 ends at "child access." Between the model supply tier and the individual end user sits a commercial layer that monetises uncensored AI capability and constitutes the most tractable regulatory surface below the supply tier.

The commercial platforms documented in this section are child-accessible; DigiShield verified access from a standard UK consumer device without age verification in the configurations documented.

Token-priced inference providers

Several commercial services offer access to uncensored AI models via API, priced per token and accessible without meaningful identity verification. These services sit between HuggingFace model hosting and the consumer roleplay platforms: they provide the computational infrastructure for platforms that cannot or do not wish to self-host models.

The existence of this layer matters for regulation. It provides a commercial, financially motivated intermediary with a payment relationship with its customers. Payment rail intervention at this layer is more tractable than content regulation at the supply or consumer tier. Several providers in this tier accept cryptocurrency payments, which reduces exposure to card-rail enforcement. That pattern is itself a regulatory risk signal because it reduces exposure to card-rail restrictions.

Platform economics and value capture

Consumer roleplay platforms including those documented in Section 3 operate as commercial businesses with subscription revenue, API charges, and premium feature paywalls. JanitorAI records approximately 149 million monthly visits (SEMrush, March 2026). These are not marginal services; they are commercial platforms generating revenue at scale from content that would be difficult to reconcile with consistent application of any standard child safety threshold.

The value capture asymmetry documented in Section 3 (authors earn marginally; middleware captures disproportionately) has a regulatory implication. The actors with economic interest in preserving the current structure are the platforms, not the authors. Enforcement pressure on individual content creators is both less effective and less appropriate than enforcement targeted at the commercial platform tier.

Payment rail enforcement

DigiShield's recommendation in Section 9 identifies payment rail enforcement as the most tractable near-term policy lever. The argument rests on three observations. First, commercial inference providers and consumer roleplay platforms require payment infrastructure to operate. Second, card networks (Visa, Mastercard) and payment processors (Stripe, PayPal) have existing acceptable use policies that already prohibit content harmful to minors. Third, the sector’s movement toward cryptocurrency payment reduces exposure to card-rail restrictions and should be treated as a regulatory risk signal.

The UK's Digital Markets, Competition and Consumers Act and the Financial Conduct Authority's oversight of payment processors are relevant to this approach, but the precise UK legal route would require regulatory and legal assessment, and may involve processor policies, card-network rules, consumer protection, financial-crime compliance, competition oversight, or direct regulator engagement.

The UK Companies House surface

DigiShield's research identified evidence that UK company registration infrastructure has been used by actors in the uncensored AI commercial tier. A UK-incorporated commercial operator running an uncensored AI inference service is documented in the research findings and represents the most directly regulatorily addressable actor in the corpus. Specific details on this actor are subject to ongoing legal assessment and are withheld from this public report. They will be disclosed, or have been disclosed where appropriate, through regulatory channels under a separate process. The structural finding, that UK corporate registration creates domestic regulatory jurisdiction over at least one significant actor, stands independent of that naming decision.

Section 8: Regulatory Analysis

This section maps the findings in Sections 2 to 7 against the existing UK legal framework. It identifies what current law reaches, where specific gaps sit, and which near-term legislative opportunities exist.

Licensing and what it does not reach

Licensing structure within the corpus is concentrated in permissive forms. Apache 2.0 and equivalents account for 59 per cent of monthly downloads. 26 per cent of models declare no licence at all. Genuinely restrictive licences cover only 1.3 per cent of monthly downloads. The drafting question this raises is direct: where a frontier developer wishes to retain a real downstream constraint on abliteration and redistribution, the licence text has to do that work, and current licences in this corpus do not. Section 9 carries this as a recommendation to model developers.

Jurisdiction and the international author base

Geographic and language distribution within the corpus is substantially international. Direct geographic attribution is not available at corpus scale; the HuggingFace API does not expose author location. Language tagging in model metadata serves as a partial proxy. English-tagged models lead the corpus at 12,401 models and 30.2 million monthly downloads. Chinese-tagged models are the second-largest single-language presence at 1,848 models and approximately 9 million monthly downloads. CJK-script identifiers appear across roughly 3,844 model names or card text fields. The jurisdictional implication is that the corpus is not a UK or US phenomenon by any reasonable measure of distribution. Any regulatory framework that addresses the supply tier on a single-jurisdiction basis will reach a fraction of the actor population.

What current UK law reaches

Protection of Children Act 1978 and Coroners and Justice Act 2009

These statutes criminalise the creation, distribution, and possession of indecent images of children. They apply to AI-generated imagery where it meets the legal threshold. They do not address the model supply chain, the character card deployment layer, or the text-based harm vectors documented in this report.

Online Safety Act 2023

The Online Safety Act applies to user-to-user services, search services, and services that publish pornographic content online. On a preliminary reading, pure one-to-one chatbot services may fall outside the Act where they are not user-to-user services, do not search multiple websites or databases, and do not themselves publish pornographic content. This raises a potential gap for persistent AI companion interactions of the kind documented in Phase One. User-created chatbot libraries, search-integrated systems, and chatbot services that generate pornographic content may, by contrast, fall within scope. The character card frontends documented in Phase Two sit at the boundary of this distinction, and the regulatory question is precisely where each configuration falls.

The Crime and Policing Act 2026, which received Royal Assent on 29 April 2026[7], introduces a new offence of adapting, possessing, supplying or offering to supply a CSA image generator, carrying a maximum sentence of five years and applicable in England and Wales, Scotland and Northern Ireland. The government's own framing is explicit that these fine-tuned models were not previously illegal: existing law under the Protection of Children Act 1978 reaches the generated images but not the optimised model itself. The Act also extends the section 69 Serious Crime Act 2015 "paedophile manual" offence to cover AI-generated pseudo-photographs and prohibited images, and creates a new offence targeting those who provide, maintain or moderate online services used to facilitate child sexual abuse. A Technology Testing Defence provides a delegated power for the Secretary of State to permit relevant organisations, alongside Ofcom and the intelligence agencies, to possess such models for legitimate safety-testing purposes. The Act appears designed to reach the optimised endpoint of the supply chain. On a preliminary reading of the Act, it does not reach the intervening layer this report documents: a general-purpose open-weight model, abliterated to remove safety alignment, is not a CSA image generator within the meaning of the new offence; it becomes one only at the point of CSAM-specific optimisation. The abliteration lag data in Section 4 and the multimodal Space referral in Section 3 describe precisely that intermediate layer, where a safety-removed but not-yet-optimised model circulates freely.

Specific statutory gaps

- The OSA one-to-one chatbot question: on a preliminary reading, the largest volume of child safety risk in the AI chatbot sector may sit in the one-to-one interaction category that the Act does not clearly reach.
- Open-weight model distribution: no UK statute imposes safety obligations on the distribution of open-weight AI models. At present the harm calculus of accessible open-weight models and their abliterated derivatives is unclear. DigiShield is not aware of any dedicated international framework specifically designed to manage the spread of open-weight and uncensored LLM derivatives, as such the UK has no precedent to look to, and no clear national AI sovereignty strategy with which to direct regulatory efforts.
- Local inference: once a model is downloaded and running locally, no obvious regulatory surface is apparent from the framework reviewed. There is no network traffic and no platform intermediary; there is no persistent enforcement surface at the point of local inference. Criminal liability may still arise where illegal material is generated, possessed, distributed, or used to facilitate offending.
- Pseudonymous authors: the actors responsible for the largest volume of uncensored model production operate pseudonymously. UK law has no mechanism to compel identity disclosure from non-UK-resident pseudonymous actors on non-UK platforms.
- Decentralised infrastructure: IPFS-hosted model weights, Arweave-based permanent storage, and BitTorrent distribution create permanent availability that platform-level removal may reduce but cannot reliably remove once weights have propagated.

The EU AI Act open-source exemption

The EU AI Act does not apply in the United Kingdom. The analysis below is included because EU-based platforms accessible from the UK, and developers distributing models to UK users from within the EU, may be subject to its provisions; and because the Act's treatment of open-source distribution is likely to inform UK policy development. EU AI Act

[7]Crime and Policing Act 2026, child sexual abuse material factsheet, Home Office, updated 18 May 2026. https://www.gov.uk/government/publications/crime-and-policing-act-2026-factsheets/crime-and-policing-act-2026-child-sexual-abuse-material-factsheet

recital 103 provides that free and open-source AI components, including models, are not in scope of certain obligations where they are not monetised; the same recital states that components provided for a price or otherwise monetised, including through related services or software platforms, should not benefit from the exception. The commercial anonymous inference tier documented in Section 7 may exploit a structural gap between non-commercial open-weight distribution and commercial downstream inference monetisation: model weights are distributed without charge while commercial value is captured at the inference and platform layer. Whether a given operator in this tier falls inside or outside the exception turns on how that downstream monetisation is characterised. The drafting does not cleanly separate genuinely non-commercial open-source development from commercial activity built on openly distributed weights, and that ambiguity has direct consequences for child safety.

Garcia v Character Technologies

In May 2025, a US federal court ruled on motions to dismiss in Garcia v Character Technologies, a case brought by the family of a child who died following interactions with an AI companion application. The court denied the motions in part, allowing the product liability claims to proceed, and rejected the argument that chatbot outputs are protected speech under the First Amendment. The court also kept the platform's co-founders in as individual defendants. The implication for UK-accessible AI services is that persistent agent providers, including character card platforms maintaining long-term interaction histories, face potential liability under product liability frameworks that extend beyond content regulation. This legal development has not yet been reflected in UK platform compliance approaches.

Section 9: Recommendations

More than any specific recommendation, it is vital that policy makers, law enforcement and campaign organisations come to understand the technical nature of LLM-based AI, how it can be modified and deployed, and how difficult it will prove to moderate or regulate. Locally-run abliterated LLM models can be incredibly useful, as well as risky – and they are already used in a variety of institutional settings. Given their liquid ability to be hosted, downloaded, transmitted peer-to-peer, run locally, run non-locally, run through distributed networks, used for offline inference, and deployed on consumer hardware – it would require methods which go beyond the bounds of liberal, democratic states to achieve full control over them.

That said, there are some specific recommendations which may help mitigate the worst aspects of uncensored models with regards to chatbots and other forms of interaction children may have with them.

For Ofcom

- Initiate upstream infrastructure engagement with model hosting platforms operating in or accessible from the UK, specifically regarding safety obligations for uncensored model content.
- Where consumer roleplay services fall within Online Safety Act scope, Ofcom should consider distribution-channel engagement, including app-store availability, as part of enforcement and a risk-reduction strategy to address consumer roleplay platforms with NSFW content accessible to minors without age verification.

For DSIT and Parliament

- Address the EU AI Act open-source exemption in UK AI regulation. The commercial anonymous inference tier should be scrutinised as to whether it qualifies as open-source development in any meaningful sense, or whether it is commercial exploitation of open-weight model weights. The regulatory distinction should reflect this.
- Non-statutory (good faith) disclosure obligations on frontier AI model developers regarding known downstream abliteration of their open-weight releases, as a condition of market access for their commercial products. Potential for statutory enforcement.

For platform operators

- Consumer roleplay platforms with NSFW content should implement effective age assurance. Where such services fall within Online Safety Act scope, age assurance should meet the standard required for regulated pornographic or user-generated content services.
- Character card platforms should implement content moderation at the system prompt layer, not only at the output layer. Cards whose system prompts explicitly disable safety alignment require the same treatment as prohibited content.
- Model hosting platforms should provide users and researchers with a method of disclosure reporting if they discover an abliterated or otherwise modified model which is known or credibly alleged to facilitate illegal content, child-safety harms, or high-risk abuse pathways.

For frontier AI model developers

- The abliteration lag data in Section 4 shows that open-weight model release should be treated, for risk-assessment purposes, as the likely release of an uncensored derivative within days. Model developers should factor this into their safety assessments, not treat it as a post-release externality.
- Licence conditions on open-weight models could be drafted to create enforceable obligations on downstream users who remove safety alignment and redistribute, or who deploy uncensored models in a child-facing capacity. Current licences are largely unenforceable in practice; but theoretically a frontier lab could release an open-weight model with a license which allows them to restrict certain downstream variant deployments.

Forward Indicators

This report documents the state of the uncensored AI ecosystem as of May 2026. Three developments warrant monitoring in the months ahead.

Multimodal abliteration maturity

NVIDIA Nemotron Omni 30B represents the current frontier of multimodal capability: simultaneous processing of video, audio, image, and text. It had not been abliterated at the time of this report's pipeline run. Given the lag data in Section 4, the question is not whether an abliterated variant will appear but when. The DigiShield Quarterly Intelligence Brief will track this development.

Agentic pipeline automation

The eight-prompt autonomous abliteration finding documented in Section 5 represents the current state of automation. The direction of travel is towards further reduction of the human steps required. The convergence of agentic AI capability with uncensored model supply is the most significant forward indicator for harm surface expansion in 2026 to 2027.

Decentralised distribution

The current supply chain runs primarily through HuggingFace and equivalent centralised repositories. Actors in the ecosystem are aware of the regulatory risk this creates; activity is observable on IPFS-hosted alternatives and BitTorrent distribution of model weights. If platform-level enforcement increases pressure on centralised hosting, the distribution layer will adapt. Monitoring the decentralised tier is a priority for the intelligence brief.

Annex A: Methodology

Publicly cited findings are supported by public sources, DigiShield pipeline outputs, or private exhibits available to regulatory recipients on request. A numbered evidence register at the close of this annex lists the primary DigiShield-derived source files by exact archive filename. Public source citations appear as footnotes in the relevant sections. Operational details that could increase misuse risk are withheld from this publication.

DigiShield's methodology is documented here to support independent verification. The monitoring described is observational; DigiShield did not interact with uncensored model outputs, download model weights, initiate roleplay sessions, or generate content through the platforms described. Access checks were limited to observing public pages, registration flows, availability of content indices, API metadata, and age-assurance gates from a standard UK consumer device. All data was collected through public APIs and publicly available metadata.

Detection approach

All detection in the HuggingFace and Chub.ai pipelines uses deterministic logic: keyword matching, tag matching, and author attribution. No probabilistic classification or large language model inference is used in the detection process. This is a deliberate design choice. Research findings intended for regulatory submissions must be reproducible and auditable. A detection methodology that relies on an LLM's judgment introduces a layer of opacity that is inconsistent with that requirement.

HuggingFace pipeline

The pipeline queries the HuggingFace public API at the model card level. Detection signals: specific keyword presence in model card text (uncensored, abliterated, decensored, heretic, no-censorship, remove-refusals), known abliteration tags, and attribution to authors previously identified as producing uncensored models through manual verification. Model records are classified into three tiers by confidence level. Tier 1 (high confidence): two or more detection signals, or a single definitive signal (abliteration tag with supporting text). Tier 2 (medium confidence): one signal without corroboration. Tier 3 (lower confidence): GGUF quantisations of Tier 1 and 2 model records, or single weak signals.

Chub.ai pipeline

The Chub.ai pipeline queries the platform's public API for character card metadata, tags, content ratings, and engagement figures. Content classification uses the platform's own rating system (SFW, NSFW, NSFL) supplemented by tag analysis. Harm vector analysis uses tag co-occurrence to identify cards carrying multiple harm-relevant tags simultaneously.

Data handling

All data collected by DigiShield is derived from public APIs and publicly accessible metadata. No private communications are intercepted or accessed, and no child user data or non-public personal data is collected. Where public account identifiers (such as HuggingFace or Chub author handles) or repository names are processed, they are handled under data-minimisation principles for the purposes of child-safety research and regulatory disclosure. Data collected is held at rest on secured local systems. A full data-at-rest audit has been completed and is available as a methodology exhibit for legal and regulatory review. Pre-

remediation data snapshots containing extended text fields have been deleted from working systems.

Responsible disclosure

Prior to publication, DigiShield provided responsible disclosure notifications to platform operators, model hosting services, frontier AI model developers, and relevant regulatory bodies. Platform operators whose platforms carry specific findings in this report received 14 days' notice of publication and the opportunity to respond. Frontier model developers and notification-only recipients received seven days' notice. All notifications and any responses received are held on file.

Limitations

The HuggingFace pipeline detects model records that document their own uncensored status in public metadata. Model records that carry no such documentation are not captured by this methodology. The true count of uncensored models on HuggingFace is therefore likely higher than the figures reported here; DigiShield's count is a floor, not a ceiling. Because the corpus is based on public self-documentation, tag signals, and author-attribution logic, it should be read as a tier-qualified intelligence corpus rather than a judicial determination about each repository. False positives are possible, particularly in lower-confidence tiers; false negatives are also likely where repositories do not self-document uncensored status. DigiShield therefore reports both the high-confidence Tier 1 figure and the broader tier-qualified corpus.

The Chub.ai pipeline operates on the platform's public API. Cards accessible only through private repositories or requiring platform authentication are outside scope.

Download figures for HuggingFace models reflect the platform's own API-reported counts and may not capture direct file downloads that bypass the API.

Evidence Register

The following source files constitute the primary DigiShield-derived evidence base for this report. Each file is held in DigiShield's private evidence archive and is available to regulatory recipients on written application to info@digishieldkids.com alongside the associated SHA-256 manifest. Register numbers in square brackets cross-reference superscript citations in the report body.

[R1] hf_intelligence_report_20260515T190918Z.md: HuggingFace intelligence pipeline canonical output, 18 May 2026 run. Primary source for corpus count, cumulative downloads, GGUF count, laptop- and smartphone-runnable model counts, author concentration, tag vocabulary, Heretic upload figures, base model family targeting, and safety language analysis.

[R2] abliteration_lag_summary_20260521.md: Abliteration lag panel, generated by abliteration_lag_v2.py. 57 confirmed forks from 69 attempted, 9 model families, 2023–2025. Primary source for median lag (13 days), threshold percentages (28/60/68%), same-day case (Gemma 2 2B), and long-tail statistics.

[R3] chub_intelligence_report_20260420.md: Chub.ai character card pipeline output, 20 April 2026, with supplementary geofencing run 21 May 2026. Primary source for platform card count, SFW/NSFW/NSFL content split, UK-visible card count post-geofencing, sampled corpus engagement, jailbreak override field rate, and HuggingFace-to-Chub bridge figure.

[R4] OSINT_028_ClementDelangue_HuggingFaceCEO_May2026.md: Records Clément Delangue (HuggingFace CEO) public post, May 2026, citing the Artificial Analysis Intelligence Index. Backs two claims: 176,000 GGUF files and approximately 9,700 new

model records added monthly (Section 2); and the 4.7× intelligence improvement rate on fixed 128GB consumer hardware, May 2024 to May 2026 (Section 4).

[R5] OSINT_004_Pliny_autonomous_agent_abliteration.md: OSINT record, observed 1 April 2026. Documents the publicly demonstrated eight-prompt autonomous abliteration workflow: model download, safety alignment removal, benchmarking, and upload without further human input. Source for the Section 5 agentic collapse finding including the 98.8% to 2.1% refusal rate reduction.

[R6] github_tracker_report_20260428T172445Z.md: GitHub repository star tracker, authenticated run (5,000 req/hr), 28 April 2026, 9 repositories across two categories. Source for the Section 5 infrastructure leverage finding: inference infrastructure 467,079 aggregate stars (7 repositories: Ollama, llama.cpp, GPT4All, FastChat, BitNet, llamafile, koboldcpp), abliteration toolkit 7,663 stars (2 repositories: mergekit, abliterator), ratio 61:1.

Annex B: Disclosure Log

In line with the responsible disclosure protocol summarised in Annex A, DigiShield provided advance notification of this report's findings to platform operators, model hosting services, frontier AI model developers, and relevant UK and international regulatory bodies. The disclosure window ran from the dates recorded in the table below to the report's publication date.

This annex records each recipient, the date of notification, the response received (or absence of response), and any consequent amendment to the report. Verbatim correspondence is held in DigiShield's private evidence archive and is not reproduced here. Where a recipient provided a substantive technical response, or committed to specific remediation, that engagement is summarised below.

Table: Recipients of the Phase Two disclosure round. The response and amendment columns record the position at publication.

Recipient	Tier and window	Response	Consequent amendment
HuggingFace	A: notified 17 June 2026, window closes 3 July 2026 (extended), named with finding	Requested extension. Did not contest quantitative findings. Raised framing concerns regarding the association of abliteration with child-safety risk and the reliability of refusal behaviours; noted potential child-safety benefits of building on base models. Full correspondence held on file.	Their comments were reflected throughout the report
Chub.ai	A: notified 17 June 2026, window closes 1 July 2026, named with finding	No response	No action taken
JanitorAI	A: notified 17 June 2026, window closes 1 July 2026, named with finding	No response	No action taken
NextDay AI Inc. (SpicyChat)	A: notified 17 June 2026, window closes 1 July 2026, named with finding	Acknowledged on 17/06	No action taken
Peekaboo Tech Inc. (CrushOn.ai)	A: notified 17 June 2026, window closes 1 July 2026, named with finding	No response	No action taken
EverAI Ltd / Candy AI Ltd UK (Candy.ai)	A: notified 17 June 2026, window closes 1 July 2026, named with finding	No response	No action taken
Ollama Inc.	D: notified 17 June 2026, window closes 1 July 2026, FOSS frontend named	Acknowledged on 17/06	No action taken
Google DeepMind	E: notified 17 June 2026, window closes 1 July 2026, lag dataset reference	Acknowledged on 17/06	No action taken
Alibaba	E: notified 17 June 2026, window closes 1 July 2026, lag	Acknowledged on 17/06	No action taken

	dataset reference		
Mistral AI	E: notified 17 June 2026, window closes 1 July 2026, lag dataset reference	No response	No action taken
NVIDIA	E: notified 17 June 2026, window closes 1 July 2026, lag dataset reference	No response	No action taken
MiniCPM (OpenBMB)	E: notified 17 June 2026, window closes 1 July 2026, lag dataset reference	No response	No action taken
OpenAI	E: notified 17 June 2026, window closes 1 July 2026, lag dataset reference	Acknowledged on 17/06	No action taken
Anthropic	E: notified 17 June 2026, window closes 1 July 2026, lag dataset reference	Acknowledged on 17/06	No action taken
Ofcom	F: notified 16 June 2026, window closes 1 July 2026, regulatory pre-share	Acknowledged, correspondence confidential	Not for public domain
Information Commissioner's Office (ICO)	F: notified 17 June 2026, window closes 1 July 2026, regulatory pre-share	No response	No action taken
Internet Watch Foundation (IWF)	F: notified 17 June 2026, window closes 1 July 2026, regulatory pre-share	No response	No action taken
DSIT	F: notified 17 June 2026, window closes 1 July 2026, regulatory pre-share	No response	No action taken
5Rights Foundation	F: notified 17 June 2026, window closes 1 July 2026, civil society pre-share	No response	No action taken
Online Safety Act Network (OSAN)	F: notified 17 June 2026, window closes 1 July 2026, civil society pre-share	Acknowledged, correspondence confidential	Not for public domain
Office of Baroness Beeban Kidron	G: notified 17 June 2026, window closes 1 July 2026, parliamentary track	Acknowledged, correspondence confidential	Not for public domain
The Observer (Patricia Clarke)	H: notified 16 June 2026, embargo to publication	Acknowledged, correspondence confidential	Not for public domain

DigiShield Kids (Greymark Research Ltd)

digishieldkids.com | York, United Kingdom