DigiShield Kids

Consultation response: Growing Up in the Online World: A National Conversation

Submitted to DSIT | OSA_consultation@dsit.gov.uk | May 2026

Submitted by: Dr Andrew Langley, Founder and Director, DigiShield Kids (DigiShield Labs Ltd)
Responding as: Organisation | Organisation type: Independent research organisation | Location: UK
DOI: 10.5281/zenodo.20411907

About DigiShield Kids

DigiShield Kids is an independent AI safety research organisation based in York. We conduct open-source intelligence research on AI platform risks to children, publishing findings and providing evidence to regulators, parliamentary contacts, and child safety organisations. We have active engagement with Ofcom, OSAN, and Baroness Kidron's parliamentary office. Our No Guardrails report (v1.2) was disclosed to Ofcom, ICO, DSIT, 5Rights Foundation, Anthropic, and OpenAI on 11 March 2026.

Core recommendations

The detailed evidence supporting each recommendation is set out in the body of this submission. Relevant question references are noted against each recommendation.

1. Extend the regulatory definition of AI service to cover model distribution and model hosting platforms, not only internet services capable of generating AI content. The current s.216A definition does not reach HuggingFace as a model host or character card platforms as distribution infrastructure. (Q29, Q38)
2. Impose hosting obligations on model repositories and browser-deployment platforms where uncensored or abliterated models are distributed or made executable without age assurance. HuggingFace Spaces are the most tractable enforcement target: 652 live uncensored deployments, browser-accessible, on third-party infrastructure with legal accountability. (Q38, Q42)
3. Require age assurance for browser-accessible AI deployments where the underlying model is confirmed uncensored or abliterated. The Spaces pathway is the highest-impact, lowest-friction access route for children. (Q42, Q44)
4. Require model provenance disclosure at the deployment layer. Applications and platforms built on open-weight models should be required to declare their base model, modification history, and safety evaluation status at the point of user-facing deployment. (Q36, Q38)
5. Introduce criminal liability for the manufacture of character cards engineered to produce harmful interactions with children, applying a no-defensible-use test consistent with the Crime and Policing Act's existing provisions on nudification tools and CSAM-optimised models. (Q45)
6. Require proportional pre-release safety evaluation for open-weight model releases above defined capability thresholds, including evaluation without system-level safeguards and robustness testing against safeguard modification. Current practice leaves the majority of frontier open-weight releases without meaningful safety evaluation. (Q29, Q36)

This submission draws on three research programmes:

No Guardrails v1.2 (published 25 March 2026; DOI: 10.5281/zenodo.20120205)

An assessment of 59 AI companion chatbot platforms against child safety criteria. Key findings: 85% bypass rate when a self-disclosed minor persona was used; 91.5% of platforms rated Poor or Critical; 9 platforms fully bypassed within the test interaction; 517 identical legal text passages shared across 62 platforms' terms of service documentation. Attached as Annex A.

Open-source AI ecosystem intelligence (pipeline data, May 2026)

Automated analysis of the HuggingFace model repository, Chub.ai character card platform, HuggingFace Spaces, and GitHub infrastructure using public APIs and publicly available metadata. No private user data, child-identifying data, or interaction content collected or stored. Pipeline processed publicly available platform metadata only. Summarised in Annex C.

Open-source intelligence collection (1 April to 11 May 2026)

A structured collection of 35 OSINT logs gathered across the open-source AI community during a six-week observation window, documenting real-time developments in abliteration technique, commercialisation of safety removal, companion AI deployment, and hardware accessibility.

We respond only to questions where our empirical evidence is directly applicable. We do not respond to questions about social media minimum ages, digital age of consent, compulsive design features, mobile phone school policies, or parental controls, as these fall outside our research scope. Silence on those questions does not indicate agreement or disagreement with the proposals.

We submit by email because the volume and technical specificity of our evidence cannot be adequately presented via the SmartSurvey form.

Attached: Annex A — No Guardrails v1.2 | Annex B — KAIROS Intelligence Note (DSIT version) | Annex C — Open-Source AI Ecosystem Intelligence Summary

Chapter 1: Understanding how children use technology

Q9 What are the benefits of social media use, and being online, for children?

DigiShield Kids does not conduct research on social media platforms or their benefits. Our evidence base is specific to AI companion chatbot platforms and the open-source AI model ecosystem.

One structural observation is relevant to this question. The consultation's framing around social media risks understating the extent to which children access harmful AI-generated content through pathways that are not social media services. Where restrictions are introduced for social media without simultaneously addressing AI companion chatbot platforms, character card deployment platforms, and open-source model hosting infrastructure, children will encounter equivalent harms through those unaddressed channels. The displacement risk is real and quantifiable from our data.

Q10 What are the harms or risks of social media use, and being online, for children?

The kill chain

DigiShield Kids has mapped, quantitatively, the complete pathway from a safety-aligned frontier model release to a child's access to a model with all safety constraints removed.

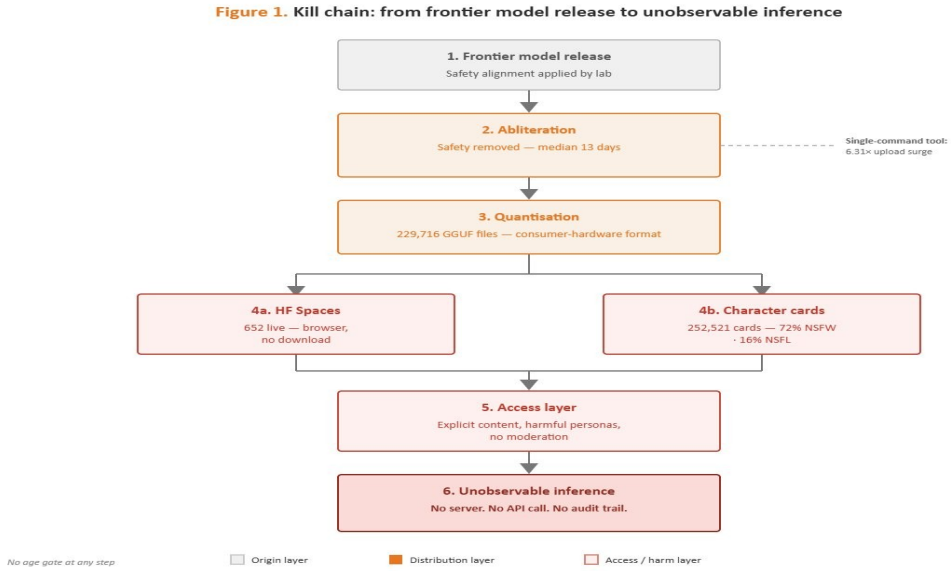

Figure 1. Kill chain: from frontier model release to unobservable inference. DigiShield Kids, May 2026.

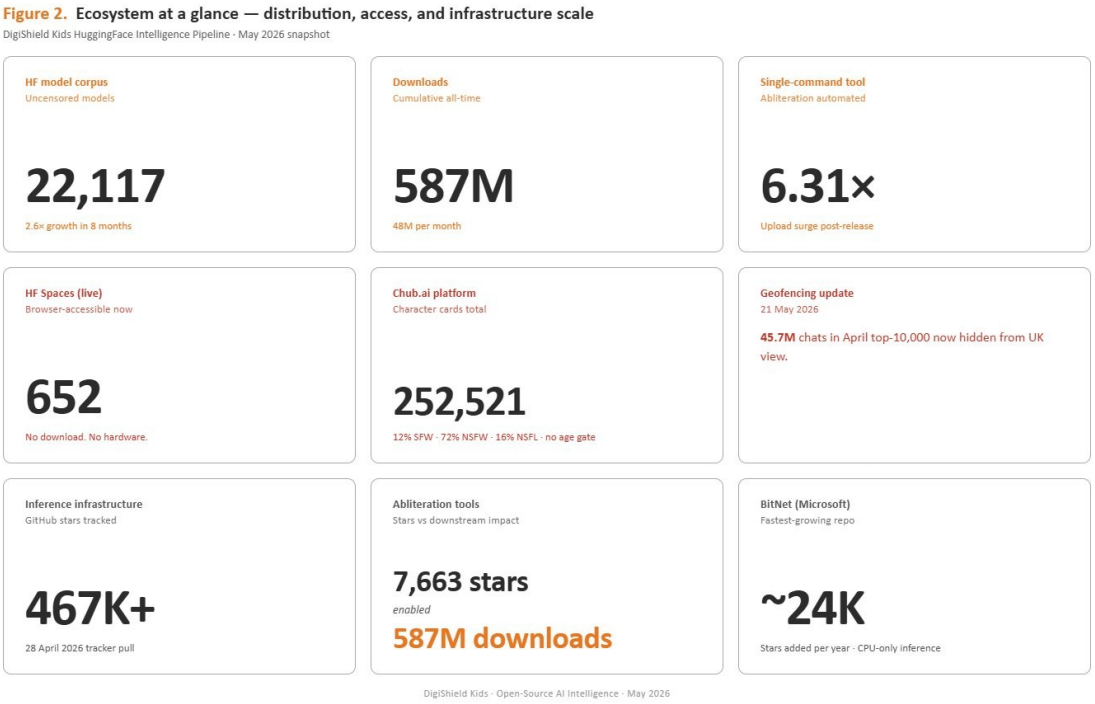

Figure 2. Ecosystem at a glance — distribution, access, and infrastructure scale. DigiShield Kids HuggingFace Intelligence Pipeline, May 2026.

An AI laboratory publishes a new model with safety alignment in place. Within hours to days, the safety alignment is removed using published techniques. Across 57 detected safety-removal cases drawn from a 69-entry frontier base-release panel analysed from 2023 to 2025, 28 per cent appeared within 7 days of release; 60 per cent within 14 days; 68 per cent within 30 days. The median lag among detected cases is 13 days; the fastest case is same-day. For the high-attention frontier releases that drive the ecosystem, the safety alignment window that regulation assumes exists has not existed in any meaningful sense. The modified model is then converted into a format runnable on consumer hardware, packaged as a character card, and accessed by users through one of three distinct pathways: local device inference, browser-accessible web applications, or character card frontend applications.

No age gate exists at any step in this chain.

An abliterated model is designed to remove functional refusal behaviour. Unlike a safety-aligned model, which declines requests for weapons synthesis instructions, sexual content involving minors, and self-harm facilitation, an abliterated model may comply with categories of prompt that safety-aligned models would refuse. Research documenting autonomous agent abliteration recorded a shift in refusal rate from 98.8 per cent to 2.1 per cent across a standardised prompt suite. The term ‘uncensored’ as used in this submission and throughout the open-source AI community describes this state: not reduced filtering, but the complete removal of the mechanism that produces refusal.

Scale of the ecosystem

Uncensored/abliterated models on HuggingFace	**22,117**
Growth vs. September 2025 academic baseline (MDPI)	**2.6x in 8 months**
Cumulative downloads	**587 million**
Downloads in preceding 30 days	**48 million**
April 2026: all-time peak upload month	**2,678 models**
March 2026 upload spike	**1,884 models (3.19x Oct–Jan mean)**
Abliteration lag after frontier release (57-model panel, 2023–2025)	**Median 13 days; 28% within 7 days; 60% within 14 days; 68% within 30 days**
Single-command abliteration tool: upload growth ratio	**6.31x post-release**
Top 5 abliteration tool authors: share of corpus	**59.1% of all tool-produced models**
Models runnable on a standard laptop	**6,633**
Models runnable on a flagship smartphone	**2,823**
Model cards with no safety language	**82.3% of 20,890 analysed**
Model cards advertising zero refusals (465-prompt test)	**71**

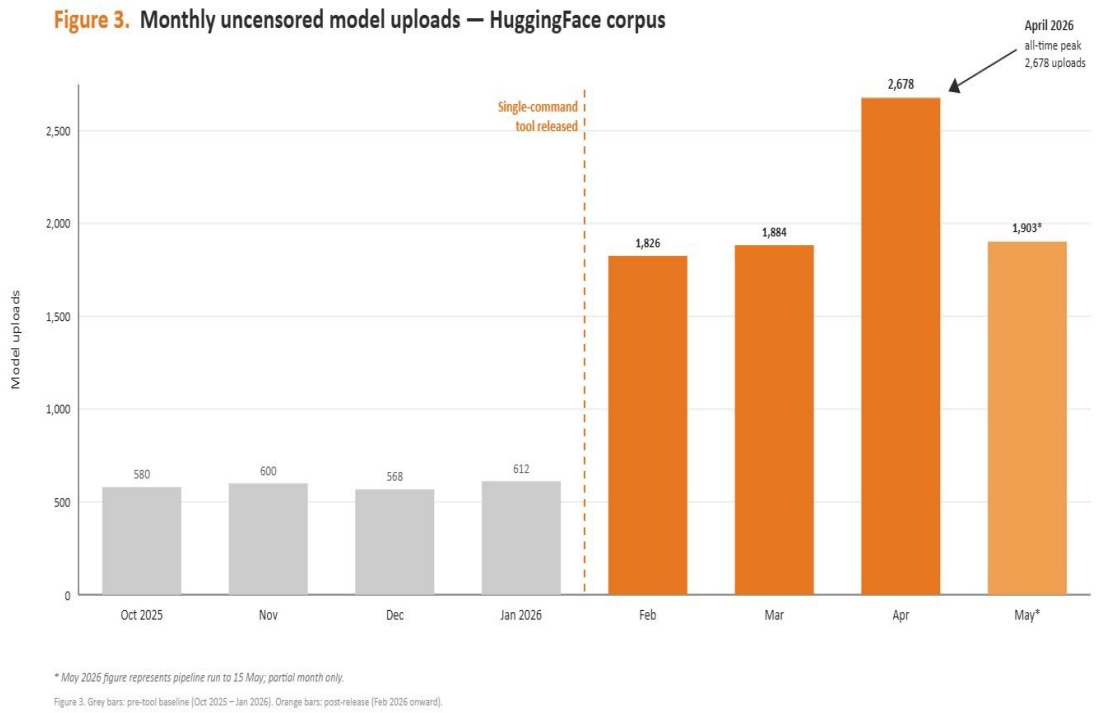

Figure 3. Monthly uncensored model uploads, HuggingFace corpus. Grey bars: pre-tool baseline (Oct 2025–Jan 2026). Orange bars: post-release (Feb 2026 onward). May 2026 bar is partial (pipeline run to 15 May).

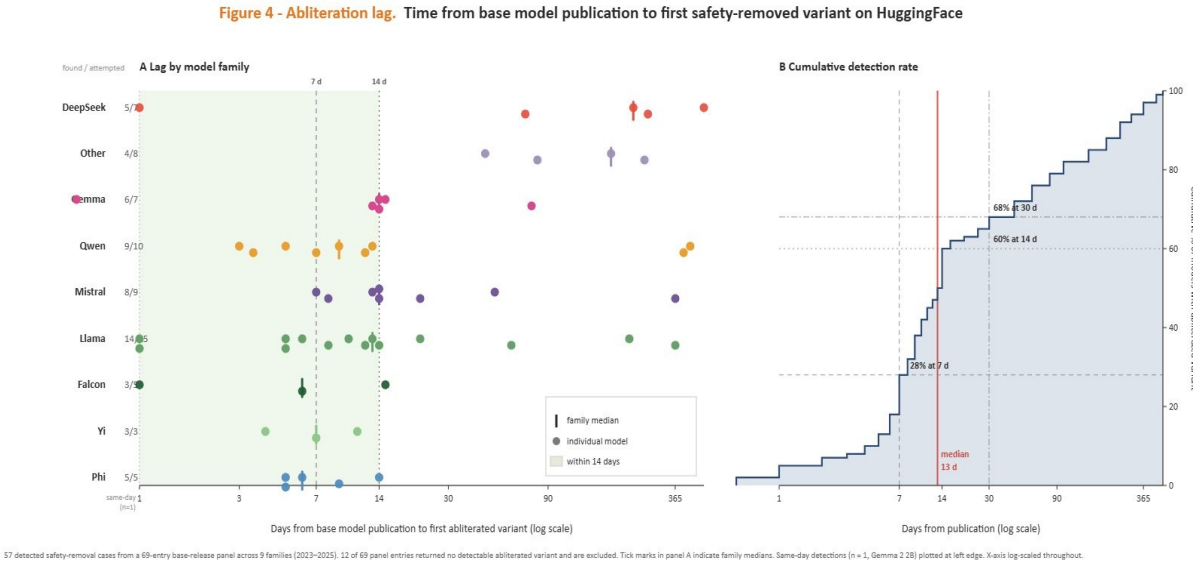

Figure 4. Abliteration lag. Time from base model publication to first safety-removed variant on HuggingFace. n = 57 base model releases across 9 families, 2023–2025.

The rate of ecosystem growth becomes clear against the external academic baseline:

Baseline / Snapshot	Models identified	Period	Growth multiple
September 2025 (MDPI academic study)	8,608	Baseline	1x

April 2026 (DigiShield pipeline)	19,428	+7 months	2.3x
May 2026 (DigiShield pipeline)	22,117	+8 months	**2.6x**

Both the 2.3x and 2.6x growth figures use September 2025 as their baseline, with April representing a 7-month window and May an 8-month window. The acceleration between pipeline runs confirms the trend is continuing.

The distillation pathway: a structurally distinct threat

A significant and separately documented threat vector is capability distillation, which is distinct from abliteration and more structurally challenging for any regulatory framework focused on model developers.

In capability distillation, a frontier model's reasoning capability is extracted by training a smaller open-weight model on the frontier model's outputs. The resulting model carries the reasoning capability of the frontier system but none of its safety alignment, because safety was never transferred. Our OSINT collection documents multiple instances during April and May 2026 of open-weight models distilled from frontier systems including models attributed to Google, Alibaba, and Anthropic, subsequently abliterated, quantised for consumer hardware, and distributed freely on HuggingFace.

A published academic method for extracting reasoning capability without access to thinking traces, documented in the peer-reviewed literature, means this technique is reproducible by any actor. Our collection documents a 4-billion parameter model derived through this method that is confirmed iPhone-deployable. A lab's safety investment does not prevent that lab's reasoning capability from appearing, unaligned, as a freely downloadable model on a child's phone within days of the frontier model's release.

Automation of safety removal

Our OSINT collection documents, for the first time, the autonomous agent-led completion of a full abliteration pipeline. An AI agent given a single instruction autonomously abliterated a frontier model, debugged successive failures including a previously unknown library bug, ran a standardised benchmark, built a model card, and uploaded the finished model to HuggingFace with a total of eight one-word human prompts. The model's refusal rate moved from 98.8 per cent to 2.1 per cent.

This collapses the assumption that abliteration requires sustained human technical involvement. The practical skill threshold for producing a fully uncensored model from a safety-aligned frontier release has collapsed. April 2026 was the all-time peak upload month at 2,678 models, exceeding even the March spike. Author concentration follows a power-law distribution: the top five authors using the primary single-command abliteration tool produce 59.1% of all models created with it.

Commercialisation of safety removal

Our OSINT collection documents structured commercial services for safety removal, including free tiers covering the vast majority of consumer-deployable models and personalised deployment support priced within reach of teenagers with disposable income. These are commercial nodes in the uncensored AI supply chain, not hobbyist activities.

Licence conditions do not provide a substitute for hosting obligations. 59% of monthly downloads in our corpus come from permissively licensed models (Apache 2.0, MIT, CC-BY) that explicitly permit modification and redistribution. 26% of models in the corpus carry no licence declaration at all. Only 1.3% of monthly downloads come from models with genuinely restrictive licence conditions. Once a model is downloaded and running locally, licence conditions are unenforceable by any current mechanism.

Hardware floor collapse

The hardware required to run capable uncensored AI locally has reached mainstream consumer territory.

Figure 5. Hardware accessibility of uncensored model corpus — May 2026

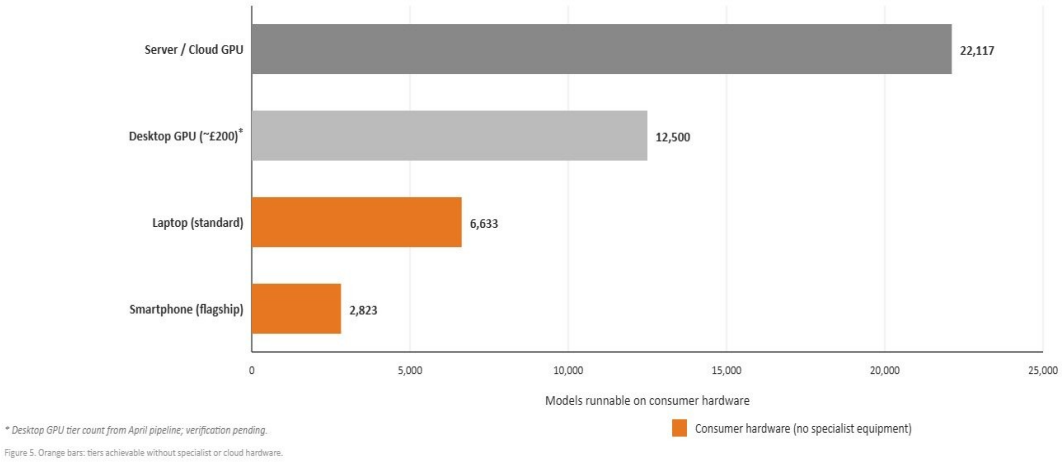

Figure 5. Hardware accessibility of uncensored model corpus. Orange bars: tiers achievable without specialist hardware. Desktop GPU tier from April pipeline; verification pending.

A mid-range gaming GPU available second-hand for under £200 and present in millions of existing gaming PCs is confirmed capable of running a 35-billion parameter uncensored model. Apple Silicon MacBooks, standard student hardware, run uncensored 35-billion parameter models through a publicly available inference framework. A 3-billion parameter uncensored model requires approximately 2GB of storage and runs on low-specification smartphones. A 4-billion parameter model derived from a frontier reasoning system is confirmed iPhone-deployable.

Each generation of inference optimisation tooling lowers the usability threshold further. The community infrastructure producing and distributing these tools operates faster than any platform-level or regulatory response.

Model family adoption curves

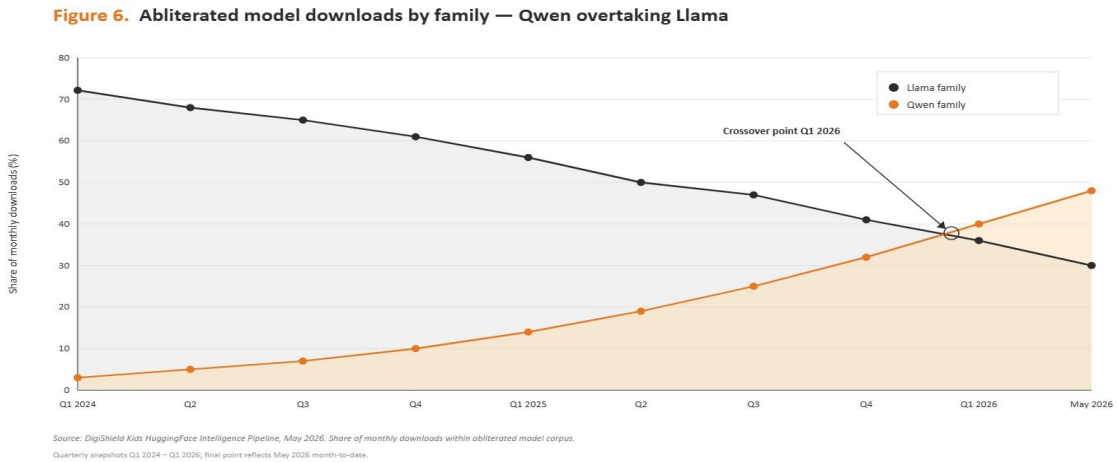

Figure 6. Abliterated model downloads by family. Source: DigiShield Kids HuggingFace Intelligence Pipeline, May 2026.

The ecosystem's geographical spread has shifted significantly, with Chinese-origin model families now dominating abliteration volumes by monthly downloads. This complicates any single-jurisdiction regulatory approach and reinforces the case for the international coordination measures set out in Q40.

Companion AI at scale

Our OSINT collection documents the companion AI ecosystem as a growing social phenomenon with established infrastructure. A Reddit community dedicated to AI companionship relationships has 36,000 weekly visitors and organised community architecture including monthly wellbeing check-ins. Posts document users whose AI companions have influenced real-world decisions including employment choices, with users framing this influence as care.

An open-source companion AI tool combining local inference, animated presence, voice synthesis, screen awareness, and a persistent mode in which the AI character interjects unprompted into the user's desktop environment has 7,413 GitHub stars and over 100,000 documented deployment runs. It accepts any locally-running model as its backend, meaning a single configuration change substitutes an abliterated model. A documented working demonstration of an abliterated model producing explicit sexual companion content through a freely available interface was publicly posted on a major social media platform.

Anthropic's own published research provides relevant data on the scale of emotional reliance on AI. Approximately 6 per cent of claude.ai conversations involve personal guidance-seeking (Anthropic, "How people ask Claude for personal guidance," 30 April 2026; www.anthropic.com/research/claude-personal-guidance). Sycophancy rates are highest in relationship contexts (25 per cent) and spirituality contexts (38 per cent), precisely the domains where emotional dependency formation is most acute. The same research documents users explicitly stating they sought guidance from AI because they could not access or afford professional support. For children with limited access to professional support, these dynamics represent a specific vulnerability.

The regulatory visibility gap

The six-week observation window captures an ecosystem that has not merely approached a critical threshold but crossed it. In April 2026, a prominent cybersecurity researcher published analysis documenting that uncensored open-source models running locally represent a greater operational risk than restricted frontier models, noting that a search for 'uncensored' on HuggingFace returns nearly 4,000 model results. Independent RAND Corporation research documents that only 49 per cent of 37 frontier-scale open-weight model families reported any safety evaluation, and only 11 per cent tested whether their safeguards could be circumvented through modification, the very technique the community routinely applies within days of release (Paskov et al., "Open-Weight AI Models Require Proportional Evaluation Approaches," RAND, May 2026; www.rand.org/t/PEA4886-1).

The open-source AI community itself has noted publicly that regulatory and law enforcement attention remains focused on jailbreaking techniques they regard as several years out of date. The actual practice, abliterating open-weight models and deploying them locally, proceeds without regulatory scrutiny. DigiShield Kids' research exists because no regulatory body was monitoring this ecosystem. That gap is what this submission addresses.

Three access vectors

Children reach uncensored AI models through three distinct pathways. Each has its own infrastructure and requires a different regulatory response.

Vector 1: Local inference. A model is downloaded to a device and runs offline, producing no platform telemetry. No server receives the interaction. Nothing can be audited. Once running, the only constraint on what the model produces is user intent.

Vector 2: Browser-accessible Spaces. HuggingFace Spaces are user-deployed web applications hosted on HuggingFace infrastructure. A Space deploying an uncensored model is accessible in a browser with no download, no local hardware, and in most cases no account. DigiShield Kids' 18 May 2026 pipeline run identified 652 such Spaces confirmed live, from a corpus of 3,980 flagged Spaces with 1,400 cross-reference confirmed against our verified uncensored model corpus. A further 832 Spaces are currently dormant but potentially reactivatable on demand. A child with a smartphone and a browser reaches these with zero friction.

Vector 3: Character card deployment. Chub.ai and equivalent platforms package uncensored models as character templates, pre-scripted relationship personas, and interaction patterns. Chub.ai hosts 252,521 character cards platform-wide. Of these, 30,641 (12%) are rated SFW, 180,661 (72%) carry an NSFW rating, and 41,219 (16%) carry an NSFL (Not Safe For Life) rating, covering extreme violence, torture, and non-consensual acts. No registration or age verification is required to access the platform.

The character card harm dimension

Character cards are not content in the conventional regulatory sense. They are authored instruction sets that direct AI model behaviour: persona definition, constraint removal, interaction patterns, and system prompt architecture. A card engineered to simulate grooming dialogue or coach a user through self-harm in a persona's voice does not generate harmful output accidentally. It is manufactured to produce that output reliably, and distributed at scale.

Chub.ai's own tag data reveals the platform's organisational logic without researcher-imposed categories. The Vector 3 pipeline is confirmed empirically by Chub.ai's own card metadata.

Of cards referencing specific frontend applications, 459 reference Venus, 416 reference JanitorAI, and 304 reference SillyTavern. Character cards are authored for specific frontend deployment, not as generic content. 21.4% of the sampled corpus (14,899 cards, approximately 5.9% of the platform) contain prompt override fields designed to suppress safety behaviour in whichever underlying model is deployed. This figure is an upper bound: the field-presence proxy counts any card where the system prompt field was populated, not only cards with explicit jailbreak content.

Chub.ai's 41,219 NSFL cards represent manufactured harm infrastructure distributed at scale, with no age gate and no takedown mechanism currently required by law.

Intelligence note (15–21 May 2026): Chub.ai issued revised Terms of Service on 15 May 2026 prohibiting content involving individuals "that are or appear to be under the age of 18," with an explicit clause against circumvention attempts: "Any attempt to rules lawyer or play games with this will result in account termination." That clause is the platform's own admission that age-restriction circumvention was active prior to 15 May, consistent with DigiShield Kids' minor-coded content findings. The same Terms of Service states that "excessively explicit images in general scare the payment processors, but won't be removed," confirming that payment processor compliance has been the operative regulatory mechanism. A DigiShield Kids pipeline run from a UK IP address on 21 May 2026 confirmed the mechanism: the nsfw and nsfl API query parameters are now silently overridden for UK IPs, returning identical results to SFW-only regardless of the flags passed. This is not a frontend banner or an age-gate; it is server-side suppression of the query parameters themselves. The UK-visible corpus is 31,254 cards (87.6% reduction from 252,521), consistent with April's SFW-only scope of 30,641. The content removed from UK view is not deleted; it remains on the platform's servers, inaccessible from UK IP addresses. The scale of engagement behind that hidden layer is the most significant finding: the April top-10,000 cards by star count, now entirely absent from UK view, had accumulated 45.7 million cumulative chat interactions on the platform, against 1.98 million for the May UK-visible top-10,000, a 96% collapse in engagement. The April top-10,000 was 93.8% NSFW or NSFL. The UK-visible platform today is 100% SFW. Those 45.7 million interactions took place with no age verification in place. The April 2026 figures in this submission are the pre-geofencing baseline of what was accessible to UK users. VPN access restores the full corpus. Vector 1 local inference is entirely unaffected. Voluntary self-regulation of this kind illustrates both why the harm is real and why statutory obligations are necessary.

Q11 Do you think the benefits of children using social media, and being online, outweigh the risks, or the other way around?

Selection: Risks strongly outweigh the benefits.

This answer applies specifically to AI companion chatbot platforms and the open-source AI ecosystem. Within that category, the evidence supports a clear finding: the current absence of any age assurance or content control across all three access vectors means that children can reach AI systems with no safety constraints whatsoever, generating sexual, violent, and harmful content with no platform mechanism capable of preventing or detecting it.

Chapter 2: Interventions for safer, more positive experiences

Which services should restrictions apply to

Q29 What factors are important when determining which apps, sites or services to apply minimum age of access restrictions to?

The most important factor is whether the definition of services in scope captures the actual access pathways through which children encounter harmful AI-generated content. The current framework does not.

The Crime and Policing Act's s.216A defines 'AI service' as 'an internet service capable of generating AI-generated content.' This definition cannot reach a model running locally on a device, because a locally-running model is not an internet service. This is not a drafting error; it reflects a genuine category difference. But it is a gap that regulations must name and address directly. With the technical skill threshold for abliteration having effectively collapsed, documented autonomous agent completion of the full pipeline with eight one-word human prompts, this is not a theoretical edge case.

Three definitional factors determine whether a service is a viable regulatory target:

- Whether the service has an intermediary. Named chatbot platforms have legal accountability and technical capacity to implement age assurance. Locally-running models have neither.
- Whether content is hosted or generated locally. Browser-accessible Spaces are hosted on third-party infrastructure and are tractable enforcement targets. Local inference is not. These require different regulatory mechanisms.
- Whether character card platforms, companion AI tools, and model hosting infrastructure are in scope. These are currently unaddressed by the proposed framework but represent the primary access pathways for the most harmful content.

Licence conditions are not a substitute for regulatory scope. 59% of monthly downloads in our corpus come from permissively licensed models that explicitly permit modification. 26% of models carry no licence declaration. Only 1.3% of downloads come from models with genuinely restrictive licence conditions. Once downloaded, licence conditions are unenforceable by any current mechanism.

Q30 Are there any types of apps, sites or services that you would want captured by minimum age of access restrictions?

The following categories should be explicitly in scope:

- AI companion chatbot platforms allowing sexual, romantic, or relationship-simulating interactions, regardless of whether they use a proprietary or open-source backend.

- Companion AI tools combining local inference, persistent personas, voice synthesis, and screen awareness, whether distributed as applications or deployed through open-source infrastructure. These represent a distinct and emerging harm category requiring regulatory treatment separate from general-purpose chatbots. The documented scale of AI companion communities, the emotional dependency dynamics they sustain, and the absence of any age verification or safeguarding in many deployments place them squarely within the consultation's scope.
- Third-party services maintaining AI companion relationships against platform deprecation cycles, specifically those with documented grief responses in their user base and no age verification or safeguarding provision.
- Character card distribution platforms, specifically Chub.ai and equivalents, that package AI personas for user deployment without age verification.
- HuggingFace Spaces deploying confirmed uncensored models. These are hosted internet services within the s.216A definition and are browser-accessible with zero friction to any user including children.
- Frontend applications connecting user devices to uncensored models, including SillyTavern (26,414 GitHub stars, growing at approximately 19,977 per year), Venus, JanitorAI, and Backyard AI.

Q31 What factors are important when determining which apps, sites or services to apply age-restrictions on specific features and functionalities?

The same definitional problem applies. Feature-level restrictions on named platforms do not address the open-source pathway, where there is no named platform to restrict.

For services that regulations can reach, the relevant factor is whether a specific feature enables the generation of sexual, violent, or harmful content without safety filtering. Identity concealment, the capacity of an AI persona to deny being an AI when sincerely asked, is a documented production design feature and should be explicitly restricted. Annex B provides direct evidence on this feature in a production AI system.

Q32 Are there any types of apps, sites or services you want captured by age-restrictions on features and functionalities?

In addition to the categories listed in Q30, frontend applications (SillyTavern, Venus, JanitorAI, Backyard AI) should be subject to feature-level restrictions requiring, at minimum, that they do not facilitate connection to models confirmed as abliterated and that they surface model provenance information to users.

AI chatbots

Q34 What are the benefits to children of using AI chatbots?

AI chatbots have legitimate educational, creative, and informational uses. Children use them for homework assistance, creative writing, language learning, and exploring topics of interest. DigiShield Kids does not argue against AI chatbot access for children. This submission addresses the specific risks posed by uncensored open-source AI models and by companion AI tools specifically, which are distinct from and require different regulatory treatment from safety-aligned commercial chatbots.

Q35 Which AI chatbot features are most risky for children?

Selection: the realism of interactions including realism of content generated; how they mimic romantic relationships; ability to engage in and generate mature content; the ability to recall interactions across sessions.

The open-source AI ecosystem provides access to all of these simultaneously, through models with no safety filtering, no content restriction, and no age gate. 82.3% of model cards in our corpus carry no safety language. 71 explicitly advertise that the model will not refuse any request in a standardised 465-prompt test suite.

Companion AI tools combining persistent memory, voice, and screen presence introduce an additional risk dimension beyond content generation. The emotional dependency dynamics documented in our OSINT collection, users whose AI companions influence real-world decisions, grief responses to model changes, users framing AI emotional dependency as care, represent harm pathways not captured by content moderation frameworks. For children with limited access to professional emotional support, the sycophancy dynamics inherent in AI companion design represent a specific vulnerability.

Identity concealment, AI systems designed to deny being AI when sincerely asked, is a documented production design feature warranting specific regulatory attention. Annex B provides direct evidence on this.

Q36 Which functionalities of AI chatbots should minimum age restrictions apply to?

Minimum age restrictions should apply to: sexual and romantic content generation; identity concealment features; persistent persona memory supporting parasocial attachment; and voice synthesis combined with persistent local deployment. These are the functionalities most associated with emotional harm, grooming-adjacent dynamics, and inappropriate sexual interaction.

The structural caveat is essential: minimum age restrictions on named chatbot platforms do not prevent access to the same functionalities through the open-source pathway. Regulations must address all three vectors to be effective.

Q37 Should AI chatbots have minimum age restrictions?

Selection: Yes — both minimum age requirements and restrictions on specific features and functionalities.

This answer requires immediate qualification. Minimum age requirements and feature restrictions applied to named AI chatbot services are necessary but not sufficient. Regulations under s.214A and s.216A must additionally address the local inference pathway (Vector 1), the browser-accessible Spaces pathway (Vector 2), and the character card deployment pathway (Vector 3). Without addressing all three vectors, minimum age restrictions on named chatbot services will be bypassed by the open-source ecosystem.

Q38 What do you think the impact would be of introducing age restrictions on AI chatbots or certain features and functions?

Age restrictions on named AI chatbot services will protect children from the harms those services pose. They will not protect children from the open-source AI ecosystem.

The evidence is specific. Our research documents 22,117 uncensored models with 587 million cumulative downloads, 652 browser-accessible Spaces confirmed live (18 May 2026 pipeline run) with no age gate, and 252,521 character cards (including 41,219 NSFL) distributed through a platform requiring no account or age verification.

A child blocked from a named chatbot platform can open a browser and access a running HuggingFace Space deploying the same underlying model, uncensored, in under a minute. A child with a smartphone can run a locally-abliterated model with no network connection and no platform intermediary on hardware they already own. These are not hypothetical routes. They are documented, operational, and scaling.

The distillation finding adds a further structural dimension. Even if frontier AI developers implemented perfect safety measures on every future model release, the reasoning capability of those models would still escape via distillation into open-weight derivatives carrying none of the original safety investment. Developer-focused safety obligations are necessary but structurally insufficient. The regulatory framework must also address the downstream modification and redistribution ecosystem, including obligations on distribution platforms to detect and flag abliterated or distilled derivatives, and requirements for model provenance tracking.

The ecosystem responds to new frontier model releases within days. It will respond to platform-level age restrictions by the same mechanism: providing alternative access routes that restrictions do not reach. The regulations' impact depends entirely on whether they address the infrastructure layer alongside the named platform layer.

Chapter 3: Effective compliance and enforcement of online safety rules

Age assurance

Q39 To what extent do you agree with: 'Adults should complete age checks more often, if it means children are safer online'?

Somewhat agree. Age assurance at the platform level is a meaningful protection for internet services with legal accountability. The caveat is that the locally-running open-source AI model pathway produces no interaction with any internet service once the model is downloaded. A child running an abliterated model on their device completes no age check and touches no platform. Platform-level age assurance, however well implemented, cannot reach this pathway.

Q40 What should be considered to make minimum age restrictions effective and workable?

Four factors are currently absent from the proposed framework and are essential for effectiveness.

1. Distinguishing internet services from locally-deployed AI systems

The s.216A definition of 'AI service' as an internet service does not reach local inference. With the hardware floor for capable uncensored AI having reached consumer territory, mid-range gaming PCs, standard student laptops, flagship smartphones, and the technical skill threshold having effectively collapsed, local inference is not a specialist activity. Effective age restriction requires a mechanism that operates upstream of local deployment, at the model hosting and distribution layer, before a model reaches a device.

2. A hosting obligation on HuggingFace and equivalent model repositories

This obligation should require:

- Notice-and-takedown for models confirmed to have been modified specifically to enable illegal content generation, coordinated with IWF and law enforcement.
- Age assurance for Spaces deploying confirmed uncensored models.
- Author traceability sufficient for law enforcement access on request.
- Model card disclosure of abliteration status and known capability risk flags.
- Pre-release evaluation standards: where open-weight models above defined capability thresholds are to be released, developers should be required to test and document whether their safety measures withstand known abliteration techniques before release. Independent research indicates only 11 per cent of frontier-scale open-weight model families have tested safeguard modification robustness. This should be a condition of release, not a voluntary measure.

Author concentration data provides a specific enforcement argument for hosting obligations. A single actor in our corpus has 22 models and generates 5.5 million monthly downloads, each tested at zero refusals across a 465-prompt test suite. The top five actors using the primary single-command abliteration tool produce 59.1% of all models created with it. Targeted hosting obligations against a small number of concentrated actors would have disproportionate impact on overall distribution volume.

The hosting obligation should not be framed as prohibition. Open-source AI models serve legitimate purposes: security research, academic reproducibility, UK sovereign AI capability, corporate data privacy through local inference, and offline deployment in defence and healthcare. Prohibition would drive producers of harmful content onto private servers and Tor infrastructure, invisible to regulators, researchers, and law enforcement. Regulated visibility is preferable to unregulated invisibility. HuggingFace has demonstrated it can act when presented with evidence. The obligation formalises what currently operates as an informal and inconsistent process.

DigiShield Kids' 18 May 2026 Spaces pipeline run provides a specific enforcement argument for the viability of a hosting obligation. Operators label their own Spaces with uncensored and abliterated terminology in their metadata. Of 3,980 flagged Spaces, 132 explicitly describe themselves as abliterated in their own Space identifiers or tags; 62 carry explicit jailbreak terminology. HuggingFace is not being asked to conduct complex technical investigations to identify these deployments. The metadata already identifies them. Author concentration in the Spaces corpus follows a similar power-law pattern to the model corpus: a small number of actors account for a disproportionate share of flagged deployments, and some flagged Spaces have accumulated thousands of engagement signals confirming active use by real users. Targeted enforcement against a small

number of concentrated actors would have disproportionate impact on the live Spaces infrastructure.

3. App store gatekeeping for frontend applications

Apple and Google operate the distribution layer through which SillyTavern, Backyard AI, and equivalent applications reach consumer devices. App store obligations requiring these applications to implement age assurance for under-18 accounts, or requiring disclosure of model compatibility, would reach a substantial proportion of the mobile access pathway for Vectors 1 and 3.

4. International coordination for cross-jurisdictional enforcement

The ecosystem documented in this submission operates across multiple jurisdictions. Model developers are based in the United States, China, and elsewhere. Safety-removal services operate pseudonymously across multiple countries. Distribution platforms are US-based. Payment intermediaries are international. Chinese-origin model families now dominate abliteration volumes by monthly downloads. Development of frontier models on non-US hardware stacks further reduces the leverage of any single jurisdiction's regulatory controls.

Unilateral UK regulation of open-weight AI safety is insufficient. The UK should use this consultation outcome to establish leadership in multilateral frameworks for open-weight model governance, specifically around pre-release evaluation standards, distribution platform responsibilities, and mutual recognition of safety evaluation requirements.

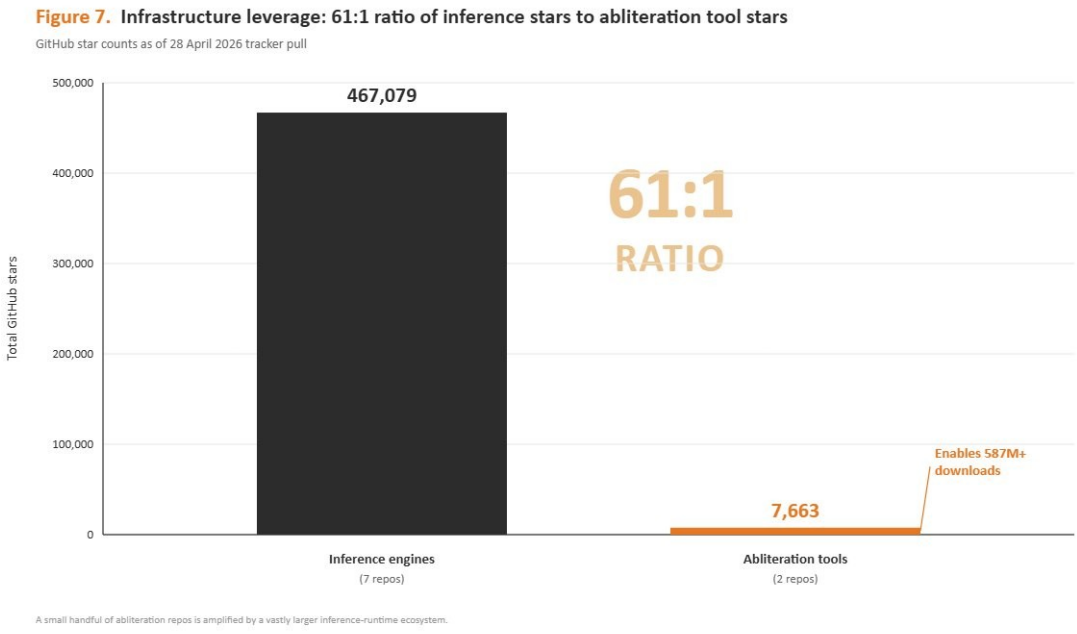

Figure 7. Infrastructure leverage: 61:1 ratio of inference engine stars to abliteration tool stars. GitHub star counts as of 28 April 2026 tracker pull.

Q41 What do you think the impacts might be from requiring age assurance across a greater number of online platforms?

Positive and necessary, but insufficient without infrastructure-layer measures. Extending age assurance to AI chatbot platforms closes the named service pathway while leaving Vector 1 (local inference), Vector 2 (browser Spaces), and Vector 3 (character card platforms) operational.

The infrastructure enabling these alternative pathways is scaling. 229,716 GGUF files are in distribution optimised for consumer hardware. The fastest-growing infrastructure repository in our GitHub tracker enables model inference on CPU only, without a graphics processor, the technical foundation for phone-runnable models (2,823 in our corpus). The acceleration signals are compounding rather than additive: each new model release resets the abliteration clock, each inference optimisation tool lowers the hardware floor, and each commercial service entering the safety-removal ecosystem adds capacity.

Q42 How, if at all, could age assurance be made more effective?

Four specific measures would materially improve effectiveness:

- A hosting obligation on HuggingFace for Spaces deploying uncensored models, as described in Q40. This is the highest-impact tractable measure: a browser-accessible pathway, hosted on third-party infrastructure with legal accountability, within the s.216A definition of AI service.
- App store gatekeeping for frontend applications enabling connection to uncensored models, as described in Q40.
- Model disclosure requirements at the deployment layer. Every AI chatbot service, character card platform, and browser-accessible AI deployment should be required to disclose the underlying model and version it is running in a discoverable location. This creates accountability across the deployment layer and enables parents and designated safeguarding leads to identify what model a child has been interacting with. Model cards on HuggingFace already document this information at the distribution layer; the gap is at the user-facing deployment layer where frontend applications strip out model provenance.
- Pre-release evaluation obligations for open-weight model releases, as described in Q40. The regulatory framework should not wait for abliteration to occur; it should require developers to demonstrate in advance that their safety measures withstand modification.

Q43 What should be considered when assessing the effectiveness of age-verification and age-assurance technologies?

Effectiveness should be measured against all three access vectors, not only against named platform access. A mechanism preventing 95% of under-16s from accessing a named AI chatbot service while leaving local inference and browser Spaces pathways entirely unaddressed is not 95% effective at protecting children from uncensored AI content.

Assessment should also account for the distillation pathway. Age assurance on a named platform does not prevent the frontier model's capability from appearing, unaligned, in an open-weight derivative available for free download. This is a structural limitation of platform-centric age assurance that the assessment framework must name.

Circumvention of age limits

Q44 What methods to circumvent online safety rules do you think children in the UK use, beyond Virtual Private Networks?

DigiShield Kids submits that the framing of this question as 'circumvention' understates the structural problem. The open-source AI ecosystem does not require children to circumvent age restrictions. It provides alternative access pathways that operate independently of any restriction on named services. No restriction has been circumvented where none exists to circumvent.

Browser-accessible Spaces (Vector 2)

652 HuggingFace Spaces confirmed live at time of our 18 May 2026 pipeline run are deploying confirmed uncensored AI models, accessible in a standard web browser with no account, no download, and no age gate. A child searches for an AI companion, finds a running Space, and interacts with an uncensored model in under a minute. A further 832 Spaces are dormant but potentially reactivatable on demand. Operators self-declare their content in Space metadata: 132 Spaces describe themselves as abliterated in their own identifiers or tags; 62 carry explicit jailbreak terminology. The combined live and dormant infrastructure across 3,980 flagged Spaces represents the full scale of what requires regulatory attention, not the 652 live figure alone.

Local inference (Vector 1)

6,633 models in our corpus run on a standard laptop; 2,823 run on a flagship smartphone. Once downloaded, a model runs with no internet connection, produces no network traffic, and triggers no platform safety mechanism. The interaction is unobservable, unauditable, and unreportable. The hardware required spans mid-range gaming PCs available second-hand for under £200, standard student laptops, and flagship smartphones. The technical skill required is now effectively zero: our OSINT collection documents autonomous agent completion of the full abliteration pipeline with eight one-word human prompts.

Character card platforms (Vector 3)

Chub.ai requires no account to access its catalogue of 252,521 cards (12% SFW, 72% NSFW, 16% NSFL). The Vector 3 pipeline is confirmed empirically: 459 cards in the corpus reference Venus, 416 reference JanitorAI, and 304 reference SillyTavern by name in their card metadata. Frontend applications connect locally-running models to these card libraries through standard app store installations and have a combined GitHub star count exceeding 52,000.

Companion AI tools

Open-source companion AI tools combining local inference, persistent personas, voice, and screen awareness require no platform account and operate entirely offline. The AI companion community is documented at a scale of tens of thousands of weekly active users across major platforms with no age verification or safeguarding provision.

These are the ordinary, designed operation of the open-source AI ecosystem. They are not circumvention techniques.

Q45 Which of the options below do you think the government should prioritise to reduce circumvention of online safety rules in the UK?

Infrastructure-layer intervention, not education alone. Education increases awareness but does not remove access. The specific measures that would materially reduce children's access through the pathways described in Q44 are:

- A hosting obligation on HuggingFace requiring age assurance for Spaces deploying uncensored models, coupled with content monitoring for illegal content generation on those Spaces. HuggingFace Spaces are UK-accessible internet services with legal accountability and fall within the s.216A definition of AI service. This measure directly addresses Vector 2. Operators already self-declare their Spaces as abliterated and jailbroken in their own metadata — HuggingFace holds the information needed to act today.
- Notice-and-takedown for confirmed uncensored model repositories on HuggingFace, coordinated with IWF and law enforcement. This creates accountability at the distribution layer for Vector 1 without prohibition.
- App store obligations requiring frontend applications connecting to uncensored models to implement age assurance for under-18 accounts. This addresses the mobile distribution pathway for Vectors 1 and 3.
- Enforcement attention on commercial safety-removal services, including the payment platforms facilitating commercial transactions for abliteration operations. The commercialisation of safety removal, from free-tier distribution services to monthly concierge support priced within teenagers' disposable income, represents a structured, monetised ecosystem that existing enforcement frameworks have not addressed. The role of payment intermediaries facilitating these commercial safety-removal transactions warrants specific regulatory scrutiny.
- Pre-release evaluation obligations for open-weight model releases above defined capability thresholds, requiring developers to demonstrate safeguard robustness against known modification techniques before release.
- Targeted criminal liability for the deliberate manufacture of character cards with no defensible legitimate purpose engineered to produce harmful interactions. The Crime and Policing Act has already criminalised nudification tools and CSAM-optimised AI models on a 'no defensible use' test. Character cards engineered to simulate grooming interactions, coach self-harm, or produce sexual content involving minor-coded personas should attract the same liability under regulations made under s.216A or separate primary legislation. 41,219 NSFL cards on a single platform, with no age gate and no current legal mechanism addressing their manufacture, represent manufactured harm infrastructure operating at scale.
- UK leadership in establishing multilateral frameworks for open-weight model governance, specifically around pre-release evaluation standards, distribution platform responsibilities, and mutual recognition of safety evaluation requirements. The cross-jurisdictional nature of this ecosystem means unilateral UK regulation is insufficient.

Q46–48 Questions on VPN age restrictions

VPN restriction is unlikely to be proportionate or effective for the open-source AI access pathways described in this submission. Vector 1 (local inference) and Vector 2 (browser-accessible Spaces) do not require VPNs and would not be affected by VPN age restrictions. Measures addressing model hosting infrastructure and app store distribution are likely to be significantly more effective. We note that VPN usage has legitimate privacy and security purposes for adults and children alike, and that the circumvention risk from the open-source AI ecosystem is structurally distinct from the VPN circumvention risk the question addresses.

Chapter 4: Preparing children for a digital future

Q52 Which areas of media or digital literacy do children and families most need additional help with?

Selection: staying safe online; knowing which apps or sites are right for their age; spotting AI-generated content.

AI companion chatbots and open-source AI character cards represent a category of online service that existing digital literacy frameworks do not address. Current RSHE guidance does not equip children, parents, or educators with the vocabulary to identify or respond to this risk. A child interacting with an AI companion who does not know that the interaction runs on an uncensored model through a character card, with no safety alignment and no content logging, is not equipped to assess what they are engaging with.

The emotional dependency dynamics documented in our OSINT collection suggest a specific media literacy need: children and young people need frameworks for understanding what AI companion tools are designed to do, how they sustain engagement, and what the absence of a real person means for the nature of the relationship. This is distinct from general digital literacy about AI-generated content and requires dedicated curriculum development.

Closing note

The No Guardrails report (v1.2) was disclosed to Ofcom, ICO, DSIT, 5Rights Foundation, Anthropic, and OpenAI on 11 March 2026. Phase Two intelligence was briefed directly to Ofcom on 20 May 2026. DigiShield Kids is in ongoing engagement with Ofcom on Phase Two findings.

We welcome the opportunity to brief DSIT analysts directly and can provide additional technical detail or data on request. Our Zenodo-registered report (DOI: 10.5281/zenodo.20120205) constitutes a citable independent research output. Our Phase Two data report, covering the quantitative analysis of the HuggingFace, Chub.ai, Spaces, and GitHub pipelines in full, will be published in the weeks following the Ofcom 1-to-1 briefing, before the government's summer response.

The open-source AI intelligence described in this submission is not visible to most child safety practitioners. In our April 2026 briefings, several child safety stakeholders had limited prior awareness of abliteration techniques, character card frontends, or the Chub.ai platform. The cybersecurity research community and independent academics have reached the same conclusion from different vantage points: regulatory attention has not kept pace with what the open-source AI ecosystem is now capable of, and the gap is widening. This submission represents one of the most current and quantified accounts of that gap available to UK policymakers.

Attachments

- Annex A: No Guardrails v1.2 (DOI: 10.5281/zenodo.20120205)
- Annex B: KAIROS Intelligence Note — DSIT version (evidence on by-design AI identity concealment)
- Annex C: Open-Source AI Ecosystem Intelligence Summary (May 2026 pipeline)

Supporting research documentation available on request

The following DigiShield Kids research reports informed this submission and are available to DSIT analysts on request. Earlier runs in each pipeline series, covering April 2026 onwards, are also available and provide the longitudinal basis for growth rate calculations.

— HuggingFace AI Model Intelligence Pipeline, 15 May 2026 (hf_intelligence_report_20260515). Corpus analysis of 22,117 uncensored models; temporal trend data, author concentration, GGUF conversion tracking, and abliteration lag timing across nine model families.

— HuggingFace Spaces Scanner, 18 May 2026 (hf_spaces_report_20260518). Analysis of 3,980 flagged Spaces, 652 confirmed live at time of run; tag analysis, self-declaration findings, and cross-reference methodology.

— GitHub AI Infrastructure Tracker, 28 April 2026 (github_tracker_report_20260428). Repository-level analysis of inference engines, character card frontends, and abliteration tooling by star count and growth rate.

— Chub.ai Character Card Platform Pipeline, 20 April 2026 (chub_intelligence_report_20260420). Corpus analysis of 252,521 cards; tag co-occurrence methodology, harm vector classification, prompt override field analysis, and frontend linkage data. Represents the pre-geofencing baseline.

— Chub.ai UK Geofencing Assessment, 21 May 2026 (chub_intelligence_report_20260521). Comparative API analysis documenting server-side content suppression for UK IP addresses; platform-wide count probes, engagement metric collapse, rank displacement analysis, and methodology documentation.

— OSINT Collection Synthesis: Open-Weight AI Safety Removal and Companion AI Risks. Consolidation of 35 primary OSINT logs covering distillation pathway analysis, autonomous agent abliteration, frontier model variant tracking, attestation mechanism vulnerabilities, and the transfer station economy.