

Excluding Children's Content from Media Buys

A transparent, layered, and independently validated methodology for keeping advertising off content made for kids, on YouTube and across digital.

What this paper commits to

Advertising that runs on or beside children’s content is one of the sharpest risks a media plan carries. It breaches platform policy and, increasingly, the law. It attaches brands to audiences they never intended to reach. And it is hard to see coming, because the signal most buyers rely on, a creator’s own “made for kids” tag, is incomplete by design.

Architect excludes children’s content with a layered model: the platform’s declared signals first, a video-level check second, and a content-pattern model third as the catch-net for everything undeclared. Uncertain cases go to a person, and every reviewed decision is kept and reused, so the model sharpens over time rather than starting blind on each run.

This paper sets out how that works, what an early evaluation across 21,112 channels produced, and how the model extends beyond YouTube. It also sets out a commitment that matters more than any single technique: to validate this methodology in the open, with regulators, child-safety authorities, and the research community, and to publish the model, its taxonomy, and its evaluation as an open, reproducible standard. A system that decides what counts as children’s content should not be a black box owned by one company.

02 • THE STAKES

Why this is becoming unavoidable

The rules that govern advertising to children have tightened sharply, and they point one way. In the European Union, the Digital Services Act prohibits platforms from showing profiling-based advertising to anyone they know with reasonable certainty is a minor, and parental consent cannot override that ban. The Commission published guidelines on protecting minors in July 2025, and the major platforms have already stopped letting advertisers target under-18s. A Digital Fairness Act now under discussion would widen scrutiny of targeted advertising and influencer marketing aimed at children.

The United States is moving the same way. The updated COPPA rule, in effect from June 2025 with a compliance deadline of April 2026, requires separate opt-in parental consent before an under-13’s data is used for targeted advertising. The Kids Internet and Digital Safety Act, passed by the House of Representatives in June 2026, would extend protection to teenagers and ban targeted advertising to minors outright.

These rules share one target: behavioural, profiling-based advertising to minors. None of them restrict advertising by content. The compliant path the market is being moved onto is contextual, decided at the level of content, with no person-level data. That is the ground this methodology is built on. What becomes scarce in that world is a reliable way to know which inventory is children’s content, so a buyer can keep advertising off it.

The sums make the point. One 2022 estimate put annual US advertising revenue from users under 18, across six major platforms, at close to 11 billion dollars, of which more than 2 billion came from children under 13, and on that estimate YouTube alone drew about 960 million dollars from under-13 users. Those figures are a 2022 snapshot rather than a current measurement, but they indicate the scale of the exposure that regulation is now moving to constrain.

03 • THE PROBLEM

Why one signal is not enough

YouTube’s “made for kids” flag is a necessary starting point, and where a creator sets it, Architect honours it first. But three gaps make it insufficient on its own:

- A channel can mislabel itself. Some creators leave the tag off by mistake, or to keep features such as comments and personalised ads that a kids’ label switches off.
- Kids content hides inside general channels. A family or entertainment channel can carry a run of clearly kids-oriented videos that are never flagged.
- Formats move faster than a checkbox. New kids formats, from unboxings to nursery-rhyme compilations to mini-vlogs, spread well ahead of any manual re-labelling.

Any method that trusts a single declared field inherits all three gaps. Architect stacks independent layers of evidence so that no one gap leaves children’s content undetected.

04 • THE METHODOLOGY

Three layers of evidence

Each layer catches what the one before it misses. A channel needs to trip only one to be held for review.

Layer 1. Declared status. The creator’s own “made for kids” flag, required under COPPA. When present, it is the fastest and most reliable signal, and a declared channel is gated automatically with no further review.

Layer 2. Video-level check. Rather than stopping at the channel label, this rolls up the individual videos on a channel and asks what share are themselves tagged for kids. It catches the mixed channels a single yes or no label misses.

Layer 3. Architect’s content-pattern model. The catch-net for channels that never declare themselves. It learns from thousands of channels the platform has already confirmed as made for kids, across ten recognisable formats (animation and cartoons, nursery rhymes and songs, toy unboxing and play, arts and crafts, pets and baby animals, learning and education, kids gaming, superhero and action play, stories and fairytales, and kid vlogs and challenges). Every channel under evaluation is read the same way, by its name, description, and recent video titles, and scored against those known patterns.

For each channel the model returns four outputs: a score from 0 to 100, a tier (low, medium, high), the closest of the ten formats, and a recommended action (clear, review, or exclude), taken down to the individual video where possible. The reasoning behind a recommendation can be inspected; nothing sits in a black box.

Why video-level matters. A general channel that occasionally posts a kids-style video is not blocklisted wholesale. Architect flags the specific videos, so the rest of the channel stays available to plan against. Exclusion should be precise, not blunt.

Human review, and a memory. Every uncertain call goes to a person, not just the model. Reviewed cases are recorded and reused, so the model grows more accurate for a given channel list with each cycle instead of being re-run blind. That accumulating, human-checked ground truth is the part that compounds.

05 • EARLY EVIDENCE

One evaluation run, 21,112 channels

In one evaluation run across 21,112 channels, every channel passed through all three layers before any plan went live. The model’s recommendations broke down as follows.

Outcome	Channels	Share
Cleared, no kids-content signal	16,958	80.3%
Specific videos flagged for review	3,751	17.8%
Recommended for full exclusion	316	1.5%
Already declared by the platform	87	0.4%

These are the model’s recommendations from a single run, not a validated accuracy measurement. We publish them to show the shape of the problem: the large majority of channels carry no kids-content signal, a meaningful minority need a closer, video-level look, and a small tail warrants full exclusion. Establishing how often those recommendations are correct is the job of independent validation, set out below.

06 · ACROSS DIGITAL

Beyond YouTube

Children are not only reached on YouTube, and neither is the risk. Architect’s aim is to exclude children’s content across digital, including display, connected TV, and out-of-home.

This is not a copy-and-paste. Layers 1 and 2 rely on platform-declared signals that mostly do not exist outside YouTube. Across other media, the content-pattern model becomes the primary line rather than the catch-net, and it has to be re-expressed for each environment. On display, the unit is the domain, page, and app. On connected TV, it is the app and the programme. Out of home, there may be no content signal at all, and the relevant question is physical context, such as a screen outside a school.

The through-line is a single exclusion decision fed by medium-specific detectors, governed by one taxonomy and one review process. What is live today is YouTube. The cross-channel extension is on the roadmap, and this paper will be updated as each medium comes online, rather than claiming coverage before it exists.

07 · THE COMMITMENT

Validation, transparency, and open source

A model that decides what counts as children’s content carries public consequences, so it should be held to a public standard. Architect commits to three things.

Independent validation. We will validate the methodology with the bodies that hold the mandate for it: media and broadcasting authorities, data protection authorities, and child-safety and consumer-protection organisations, in each market where we operate. Validation means measuring the model against an agreed, human-labelled ground truth and reporting how it performs, including where it fails.

Scientific review. We will work with the research community, in media measurement, child-media studies, and machine learning, to design the evaluation, scrutinise the method, and publish results that can be reproduced and challenged. The measures that matter, such as how much real children’s content the model catches and how often it wrongly flags general content, will be published and tracked over time. We are not publishing those figures in this version. We will publish them as validation proceeds, rather than assert a number we have not yet earned.

Open source. The core of the methodology will be open: the taxonomy of children’s-content formats, the criteria by which content is classified, the evaluation harness, and a public benchmark, released so

authorities, researchers, and advertisers can inspect, test, and improve it. Individual client configurations and any customer data stay private. The shared standard does not.

The reason is simple. Child safety is not a place for proprietary secrecy, or for marketing claims that cannot be checked. An open, validated model earns more trust, catches more errors, and improves faster than a closed one. That is the model Architect intends to build.

08 · PRINCIPLES

What holds across every layer

- Channel and content level only. Architect uses signals about content, not about individuals. There is no person-level tracking, and no profiling of children.
- Cautious by default, tunable by design. When a call is uncertain, the default is to route it to review or exclusion, not to clear it. Clients set how conservative the thresholds are, but the safe default is caution.
- Precise, not blunt. Exclusion is taken to the individual video or placement where possible, so brand safety does not mean needlessly discarding inventory.
- Honest about maturity. Where something is live, we say so. Where it is on the roadmap, we say that too.

09 · WORKING TOGETHER

An invitation

Architect is looking for partners in this: authorities and child-safety bodies to help set the standard, researchers to test it, and advertisers and agencies to put it to work. If your organisation wants to review the methodology, contribute to the taxonomy, or run an evaluation, we would like to hear from you.

hello@architect.media

REFERENCES

- Raffoul A, Ward ZJ, Santoso M, Kavanaugh JR, Austin SB. Social media platforms generate billions of dollars in revenue from U.S. youth. PLOS ONE, 2023. doi:10.1371/journal.pone.0295337
- European Union. Digital Services Act (Regulation 2022/2065), Article 28, protection of minors.
- European Commission. Guidelines on the protection of minors under the Digital Services Act, July 2025.
- US Federal Trade Commission. Children’s Online Privacy Protection Rule, final amendments, January 2025 (in effect June 2025, compliance April 2026).
- US Congress. Kids Internet and Digital Safety Act (H.R. 7757), 119th Congress, passed by the House June 2026.
- European Parliament. Measures to make online services safer for minors, October 2025.

This is a working paper, version 1.1. It describes methodology and commitments, some of which are in active development. Figures attributed to external sources are estimates as of their publication date. This paper will be revised as validation results and cross-channel capabilities are published.