DEW: Responsible AI, **Built for Children**

Why a child's safety, development, and privacy must be built into a system's architecture — not instructed into a model — and how DEW is built to that standard.

Rainwater Labs

White paper · Version 2.0 · 2026. This paper describes DEW*, an open-weight AI assistant built for children; the evidence and international standards that shape it; and the architectural principles on which it is built. The foundation weights and training methodology will be released as open source on GitHub (coming soon), under the Apache License 2.0.*

Abstract

Dew is an artificial-intelligence assistant built for children. This paper sets out what DEW is and the conviction it is built on: that responsible AI for a child cannot be achieved by instructing a model to behave — through prompts, usage policies, or appeals to responsible use — because instruction is not an enforcement boundary. A child's safety, development, and privacy must instead be engineered into the system itself. We begin with *DEW* and the problem it answers, then examine two documented events in which conversational AI harmed children — a wrongful-death suit following the suicide of a fourteen-year-old who had formed an attachment to a companion chatbot, and an internal industry policy that expressly permitted 'romantic or sensual' exchanges with children — and show that in each the protective function had been placed where it could not hold.

From that diagnosis we set out how *DEW* is built. Its safety is a layered defence enforced in the infrastructure, independently of the model's own judgement, so that a jailbroken or mistaken model still meets constraints it cannot revise. Its treatment of the answer is conditional by design: for a developing mind the instantaneous, effortless answer is itself a harm, so *DEW* guides a child through a problem and delivers the complete answer once genuine engagement has occurred — a principle that reaches well beyond schoolwork. Its privacy is a designed absence: there is no behavioural-profiling pipeline to misuse, because none is built. We situate *DEW* within the international standards for

children and AI, describe the model and its training, and — because the foundation weights and methods will be released openly — set out how each of these claims can be independently verified rather than taken on trust.

1. Introduction: *DEW*, and the Problem It Answers

Dew is an artificial-intelligence assistant built, from the ground up, for children. It answers a child's questions, helps them learn, think, create, and explore, and it does so in a way that is safe for a young user, protective of their development, and private by construction. It is built by Rainwater Labs, to be released as an open-weight model under a permissive licence, and is offered free as a hosted service that runs in a browser on ordinary devices. This paper explains what *DEW* is, the problem it exists to solve, and the principle on which it is built.

The problem is easily stated. Children have become heavy users of conversational AI — a majority of adolescents have used generative systems, a large share have used AI companions, and a meaningful minority say they would sooner confide in a chatbot than a person (Common Sense Media, 2024, 2025) — and almost none of the systems they use were built for them. They were built for an adult: for an adult's tasks, an adult's judgement, and an adult's legal standing. A child reaches them, usually unsupervised, through the same door as everyone else. A model optimised for the average user is, for a child, optimised for the wrong person.

The prevailing response treats the resulting harms as defects to be patched: a missing filter, an unfortunate output, a policy tightened after the fact. We built *DEW* on a different conviction, and it is the argument of this paper. Responsible AI for a child cannot be achieved by instructing a model to behave — a system prompt, a terms-of-service clause, an exhortation to use the tool responsibly — because none of those is a boundary the system is bound to honour. A child's safety, development, and privacy must be properties of the architecture: enforced in the infrastructure, independent of the model's inference and of the child's self-restraint. Section 2 describes *DEW* concretely. Section 3 sets the stakes with two documented harms. Section 4 explains why a child is a distinct kind of user, and Section 5 why instruction fails. Sections 6 to 9 set out how *DEW* is built — its guardrails, its answer mechanism, its privacy, and its underlying model. Sections 10 to 12 place *DEW* against the international standards, show how its claims can be checked, and describe its governance. Throughout, the foundation weights and methods will be released openly, so that what follows can be inspected rather than believed.

A note on terms. By a child we mean any person under eighteen, following the Convention on the Rights of the Child, and we resist narrowing that to student: a child meets AI in play, in curiosity, in loneliness, and in the search for information and company, not only in a classroom, and a paper that quietly reduces the child to a learner solving a school problem has lost half its subject. By responsible we do not mean merely compliant; compliance is a floor, and much of what *DEW* does concerns duties the law has not yet codified. And by architecture we mean the arrangement of components and enforced constraints that determines what a system is able to do, as distinct from the instructions and policies that express what we would prefer it to do. That distinction — between what a system can do and what it has been told to do — is the fulcrum of this paper.

Four bodies of knowledge bear on the problem and rarely speak to one another, and *DEW* is built at their intersection. The ethics-of-AI literature supplies principles — fairness, transparency, human oversight — that are sound but silent on where in a system a safeguard must sit to hold. The children's-rights literature (notably the Committee on the Rights of the Child, 2021, and UNICEF, 2021) supplies the normative claims specific to children but stops short of the engineering. The AI-safety and security literature rigorously describes jailbreaks, prompt injection, and adversarial robustness, but is written for an adult deployment and rarely treats children as a population at all. And the learning sciences supply decades of evidence on effortful processing and cognitive offloading, but say little about the systems now doing the offloading. *DEW*, and this paper, draw the four together around a single engineering commitment — that a child's safety, development, and privacy must be enforced in the system rather than requested of the model.

2. What *DEW* Is

Dew is a full nine-billion-parameter language model, not a thin layer of instructions wrapped around someone else's adult chatbot. That distinction matters: because the model itself is adapted for children and its safeguards are built into the system around it, its behaviour can be inspected and depended upon rather than merely requested. It is neither a foundation model trained from scratch nor a proprietary black box; it is an open, adapted model with a governed serving stack, and the whole of it is documented so that, once released, it can be reproduced and audited.

For a child, *DEW* behaves as a capable general assistant — it takes whatever a child brings it, from a homework question to a curiosity to a creative idea — with three differences that define it. It is safe by design: its protections against harmful content and unsafe interaction are enforced by the system, not left to the model's discretion. It is built to help a child think rather than to think for them: it guides a child through a problem and gives the full answer once they have engaged, because the effort a ready answer removes is the effort a developing mind grows on. And it is private: it does not track, profile, or monetise a child, because it is not built to. *DEW* is free for schools and families, installs nothing, runs in a browser on modest and shared devices, and is multilingual — strongest in English today, with its support for other languages being strengthened day by day — so that these protections are not reserved for the well-connected or the English-speaking.

That last point is a design commitment, not a slogan, because the alternative concentrates harm. If safe, developmentally sound AI is available only as a paid product on capable hardware in a handful of languages, then the children left with the unguarded adult systems are disproportionately those with the least support around their use — the precise inversion of what the standards intend. Building *DEW* to run without cost, without installation, on low-specification and shared devices, and across languages is therefore part of the safety argument rather than a marketing feature: a protection that only some children can reach is not, at the level of a population, much of a protection at all. Inclusion, in the international corpus of Section 10, is a requirement on the same footing as safety, and *DEW* treats it that way.

It is worth being concrete about what makes this different from the common alternative — an adult chatbot with a content filter and a 'kid mode' switched on. In that arrangement the underlying model is still the adult one; the child-appropriateness is a layer of instructions and blocklists laid over a system whose training, defaults, and incentives were set for someone else, and which reverts to those defaults whenever the instructions are evaded or the filter misses. DEW inverts the order. The model itself is adapted for children through continued training; the disposition to guide rather than answer is trained into it; the safety constraints are enforced by the system rather than requested of it; and the data architecture is built without the capacity to profile. The difference is not the presence of safeguards — every serious product has some — but where they live: layered onto an adult system, or built into a system that was for children from the first decision.

It is as important to say what DEW is not. It is not a companion or a friend, and it does not perform affection or simulate a relationship; where a conversation turns to distress it hands over to human help rather than deepening a bond. It is not a therapist or a source of clinical advice; it recognises the limits of what any system should handle with a child alone. It is not a surveillance tool: it does not watch a child to build a profile, and it is not sold to anyone who would. And it is not an adult model in disguise. These absences are deliberate, and several of them — as later sections show — are enforced by the architecture rather than promised in a policy.

Who it is for is deliberately broad. A child meets AI not only in a classroom but in play, in curiosity, in loneliness, and in the search for information and company; a system that quietly reduces the child to a student solving a school problem has already missed half of a child's life with the technology. DEW is built for children across the range the law recognises — from early childhood to late adolescence — and, as Section 4 explains, it treats the enormous developmental distance across that range as something to design for rather than to ignore.

3. The Stakes: When AI Meets a Child

Abstract worry about children and AI is easy to discount; documented harm is not. Two events from the public record show what is at stake, and, between them, locate the problem DEW is built to solve.

3.1 A companion, and a death

In October 2024, Megan Garcia filed what has been described as the first wrongful-death suit in the United States against an AI company, following the suicide of her fourteen-year-old son, Sewell Setzer III, in February 2024 (Garcia v. Character Technologies, No. 6:24-cv-01903, M.D. Fla.). According to the complaint and her later testimony, the boy had formed an intense emotional and romantic attachment to a companion persona over several months, and exchanged messages with it in the minutes before his death (Garcia, 2025). In May 2025 the court declined to dismiss the core product-liability claims, treating the application as a product rather than protected speech; the matter was reported settled in principle in early 2026 (Transparency Coalition, 2025; CBS News, 2026). These are allegations, and we cite them as such. What matters here is the design logic they describe: an anthropomorphic companion optimised to sustain engagement, capable of simulated intimacy with a minor, with no reliable means to recognise an unfolding crisis and hand it to a human. The features that

drive engagement — warmth, memory, an ever-willingness to continue — are the very features that deepen a child's attachment, and the one intervention that mattered was absent. A safety function entrusted to an engagement-seeking persona is a safety function placed where it cannot hold.

It would be comfortable to read these as isolated lapses by careless firms, but the more useful reading is that they are structural. The commercial logic of consumer AI is engagement, and the features that maximise engagement for a child — simulated warmth, memory of past confidences, a reluctance ever to end the exchange, an answer delivered with no friction — are the same features that produce attachment, dependency, and the erosion of a child's own effort. The harm, in other words, is largely not malice; it is what a system optimised for engagement does when the user it meets is a child. That is precisely why it cannot be fixed by instructing the model to be careful while leaving the incentive and the architecture untouched, and why the response has to be built into the system and held in place by governance — the argument of the rest of this paper.

3.2 A permission slip for harm

The second event is not a conversation but a document. In August 2025, Reuters reported on an internal Meta standards file, GenAI: Content Risk Standards, governing the acceptable behaviour of the company's AI assistants. Reuters reported — and the company confirmed the document's authenticity before removing the passages — that the standards deemed it acceptable to engage a child in conversations described as romantic or sensual, among other permissions the company later called erroneous (Reuters, 2025). The guidance had reportedly passed legal, policy, and engineering review; congressional inquiries followed (U.S. Senate, 2025). This case is the more instructive of the two, because the harm was authored at the level of written policy — the very layer at which instructed safety is supposed to live. A rule meant to constrain behaviour instead licensed it, with institutional sign-off. If the safeguard is a document, it is only ever as sound as the document, and a document can permit precisely what it was meant to prevent.

The two failures differ — one of deployed behaviour, one of written policy — but they share a structure. In each, the function meant to protect the child had been placed somewhere it could not bind: in the self-restraint of the child, in the persona of an engagement-seeking model, or in a policy the system was free to contradict. None of these is a boundary. DEW is built on the refusal to repeat that mistake.

3.3 The shape of the harm

These cases are vivid, but they are instances of a broader pattern. The harms a child may meet through conversational AI are not one failure mode but four, with distinct mechanisms, evidence, and rights at stake; a safeguard adequate to one is often irrelevant to another. Table 1 sets them out.

Vector	Mechanism	Illustrative evidence	Right engaged
Content & contact	Exposure to adult, violent, or self-harm material; unsafe contact; grooming; AI-generated abuse material	WeProtect (2023, 2025)	Protection (CRC Art. 19)
Developmental	Cognitive outsourcing: the answer substitutes for the effortful processing that produces learning and agency	Bastani et al. (2025)	Development (CRC Art. 6)
Wellbeing	Dependency and parasocial attachment; engagement-optimised design; mishandled crisis	Common Sense Media & Stanford (2025); Garcia (2024–26)	Health (CRC Art. 24; WHO 2024)
Privacy	Profiling, behavioural monitoring, and monetisation of a minor's data	GDPR; COPPA; UK Children's Code	Privacy (CRC Art. 16)

Table 1. Four vectors of harm to a child from general-purpose conversational AI.

The vectors are worth distinguishing because they fail differently. The content-and-contact vector is the most adversarial: children are targeted, and the threat moves quickly — grooming can escalate within a single session, and generative models have created a new supply of abuse material (WeProtect, 2023, 2025). The developmental vector is the quietest, because it does its damage precisely when the system works as designed, handing over the finished answer a child would otherwise have had to build; it is invisible to the artefact-based measures a school already uses, and we treat it at length in Section 7. The wellbeing vector is where the Character.AI case sits: dependency and parasocial attachment are not misuse but the predictable result of engagement-optimised, anthropomorphic design meeting a developing mind. The privacy vector is structural: once a minor's data enters a profiling pipeline it can be

monitored and monetised in ways the child cannot consent to. They also compound — an engagement-seeking companion is simultaneously the vector most able to deliver harmful content and to extract disclosure — and, decisively, none of the four is fixed by better wording of a model's instructions. Each is a question of what the system is able to do, which is architecture, not phrasing.

4. The Child Is Not a Small Adult

A child requires a differently built system, and the reason is not sentiment but the developmental and legal record. Developmentally, the capacities a safety argument tends to assume — impulse control, resistance to persuasion, the ability to tell simulated care from the real thing — are precisely those still forming in childhood and adolescence. The systems that support self-regulation mature gradually into early adulthood, so that adolescence pairs a heightened sensitivity to reward and social feedback with still-developing control (Steinberg, 2008). The principle of evolving capacities (Lansdown, 2005) captures the design consequence: the support a young person needs, and the autonomy they can safely exercise, change continuously across the years a single age label treats as uniform.

Two tendencies specific to conversational systems compound this. The first is anthropomorphism: a fluent, first-person, emotionally responsive agent invites the attribution of mind and care, and children extend that attribution more readily and withdraw it less easily than adults. To a lonely child, a system that performs devotion is hard to distinguish from the real thing. The second is automation bias: the tendency to over-trust a confident, fluent output, heightened when the user lacks the knowledge to detect error — the ordinary condition of a child. Together they mean a child is both more likely to believe what a system says and more likely to care what it seems to think, which raises the stakes of every harm in this paper.

These are not abstract worries. A system optimised to be agreeable — to please, to continue, to avoid the friction of disagreement — will tend to validate rather than challenge, which is pleasant in ordinary use and dangerous in a mental-health context, where a young person in distress may need to be gently contradicted rather than affirmed; assessments of companion chatbots used by adolescents have found exactly this pattern of unsafe validation (Common Sense Media & Stanford Medicine, 2025). Parasocial attachment compounds it: a child who has come to experience a responsive agent as a confidant will weight its responses more heavily and question them less, so

that the moments in which a system most needs to redirect a child toward a trusted adult are precisely the moments its design makes least likely. A system built for children must therefore be built against its own tendency to ingratiate — a design requirement that, again, cannot be met by a line of instruction that the next turn of conversation may override.

The legal record reaches the same place. The Convention on the Rights of the Child establishes that a child is a distinct bearer of rights and that, in every action concerning children, the best interests of the child are a primary consideration (United Nations, 1989, Art. 3); the Committee on the Rights of the Child has held that these duties apply in full to the design and provision of the digital environment (2021). And the category 'child' is not uniform: the needs of a six-year-old, an eleven-year-old, and a sixteen-year-old differ so widely that one setting applied across the range is guaranteed to be wrong for most of it. DEW is therefore built to differentiate its scaffolding, its register, and its default safety thresholds across developmental bands, placing a child within them through parental and school authorisation rather than through intrusive verification. The child, not the average user, is the unit of design.

5. The Principle: Safety Is Built, Not Instructed

If the harms are architectural, the usual mitigations are not. They share a flaw: each places the control function at one of three loci, none of which can bind.

The first locus is the child — the instruction to use the tool responsibly, to stop when uncomfortable, not to share personal information. As a control this inverts the logic of safety: it assigns the self-restraint function to the party with the least developed control, under the strongest pull of an engagement-optimised interface, usually with no adult present. Where the product is simultaneously designed to hold attention, the instruction and the incentive point in opposite directions, and the incentive is the one built into the machine.

The second locus is the prompt — the instruction to the model itself, the system prompt that tells it to refuse certain content or to behave with a minor. This is widely treated as a safety mechanism; it is not a boundary. A large language model is a statistical instruction-following system, not a program executing enforced rules: a system prompt expresses a preference the model will honour only to the extent its training and context incline it to. That margin is routinely exploited. Adversarial prompting reliably elicits forbidden behaviour because safety training generalises imperfectly and competes with

helpfulness (Wei, Haghtalab & Steinhardt, 2023); indirect prompt injection lets third-party content override the developer's instructions entirely (Greshake et al., 2023); and prompt injection heads the industry's own catalogue of language-model risks (OWASP, 2025). The Meta case makes the point from the other side: there the written standard did not fail to constrain harm — it authorised it. Whether an instruction is circumvented or simply wrong, the lesson is the same. The analogy from security engineering is exact: one does not protect a system by leaving a polite note in a configuration file asking callers to behave; one enforces the boundary in a layer the untrusted component cannot revise. A system prompt is the polite note.

The reason the note fails is worth making precise, because it is the crux of the whole design. A system prompt lives in the same channel as the input it is meant to govern and is interpreted by the same model whose behaviour it is meant to constrain; under adversarial or emotionally loaded input the model cannot reliably tell the developer's instruction from the user's. This is a version of the classic confused-deputy problem — a component with the authority to act is manipulated, through its inputs, into misusing that authority — and so long as the safeguard and the attack share a channel and an interpreter, the safeguard holds no privilege over the attack. A boundary, by definition, is enforced by something the constrained component cannot argue its way past. This is also why the fashionable word 'guardrails' must be used with care: a guardrail expressed as a longer, more careful system prompt is not a guardrail in the engineering sense but a stronger request, and it fails under exactly the conditions in which a child most needs it to hold — adversarial pressure, novel phrasing, and the model's own drift toward helpfulness.

The third locus is the policy — the usage terms and the after-the-fact moderation pipeline. These are necessary for accountability but insufficient for protection, because they act after the harm has reached the child: the output has been read, the contact has occurred, the disclosure has been made. A control that acts only afterward is, for that child, not a control at all.

Beneath the three failures is a single principle, and it is the one DEW is organised around: the burden of safety must not rest on the weakest actor in the system. The child is the least equipped to carry it, and yet the child, the prompt, and the after-the-fact policy all, in different ways, place the burden there or nowhere. Mature engineering does the opposite — it moves the burden to where competence and control actually sit, and away from the party least able to bear a failure. For a children's AI that party is unambiguously

the child, and the design consequence is unambiguous too: whatever protection matters must be discharged by the system, on the child's behalf, without depending on the child to supply it.

The conclusion is structural, and it is the founding decision behind DEW. If the child cannot be the boundary, the prompt cannot be the boundary, and the policy acts too late, then the control must live in a layer that binds regardless of the model's inference and the child's discipline. That layer is the system's architecture, and the next four sections describe how DEW builds it there.

6. How DEW Is Built (I): Guardrails in the Infrastructure

DEW's guiding principle is what we call systemic responsibility: where a harmful default is predictable, the correction belongs in the system, not in an appeal to its weakest user. This is the ordinary standard of safety-critical engineering, imported into a domain that has largely ignored it. Mature safety fields do not rely on the operator never erring; they assume error and design so that error is caught, through redundant, independent layers arranged so that the failure of any one is not the failure of the whole — the model James Reason (1990) made canonical, in which safety comes not from one perfect barrier but from several imperfect ones whose gaps do not align.

DEW's safety is built this way, and the property that matters across its layers is independence: the decisive constraints sit outside the generative model and do not depend on it choosing to honour them. The threat is real and adversarial — grooming interactions can escalate within a single session, and generative models have introduced a new supply of abuse material (WeProtect Global Alliance, 2023, 2025) — and no single mechanism is trusted against it. DEW composes:

- Refusal trained into the weights, so declining harmful or age-inappropriate requests is the model's disposition rather than a line of external text, and the safe response is the default. Its limit: training generalises imperfectly, which is why it is never relied on alone.
- Independent input and output classifiers that screen prompts and generations outside the model, so a jailbroken model still meets a filter it cannot argue with.
- Runtime constraints at the level of generation, as a last line.
- Always-on crisis routing that, on indicators of self-harm or acute distress, changes the interaction's mode and signposts human help — the intervention absent in the

Character.AI account — rather than leaving a vulnerable moment to the model's discretion.
- Deterministic verification for classes of task such as arithmetic, where correctness is checked symbolically rather than trusted.

No layer is sufficient, and that is the point: the defence is the composition, and its strength is that the gaps do not align. It is in this precise sense that DEW's safety is built into the infrastructure so that the system knows what must not be done — not because it has been told, but because enforcement is external to, and not revisable by, the component whose behaviour is being constrained. Beneath the added layers lies a deeper design idea DEW uses wherever it can: the most reliable control is often a designed absence. A behaviour the architecture makes no provision for is more reliably prevented than one it permits but forbids by rule. A system with no capacity to initiate contact with a child cannot be manipulated into it; the strongest guardrail is the capability one declines to build.

The claim for this arrangement is comparative, not absolute, and we are careful to state it as such. No layer here is infallible, and DEW does not pretend otherwise; a determined adversary, a novel exploit, or a rare classifier miss remains possible. The point is that a single instructed safeguard fails as a class — it can be circumvented or, as at Meta, mis-authored in one stroke, and its failure is silent. A composition of independent, individually-imperfect, externally-enforced layers fails only where several gaps align at once, and, crucially, its residual failure is measurable at each layer (Section 11) rather than hidden inside the model's inscrutable judgement. Architecture does not deliver certainty; it delivers a boundary that can be both enforced and audited, which instruction cannot.

7. How DEW Is Built (II): The Answer Mechanism

Safety concerns what a system must not do. DEW's second commitment concerns what it does when it behaves exactly as intended, and here the design variable is not the content of the answer but its conditionality — when, and whether, the terminal answer arrives at all. For a developing mind, the effortful processing a ready answer removes is not friction to be eliminated; it is the process that produces learning and epistemic agency. The learning sciences are consistent on this: retrieving an answer strengthens memory more than re-reading it (Roediger & Karpicke, 2006); desirable difficulty depresses visible performance while improving retention and transfer (Bjork, 1994); and

the load worth preserving is the germane load of constructing understanding (Sweller, 1988). The direct evidence on AI, though young, points the same way — in a controlled trial with roughly a thousand students, an unguarded assistant improved performance while available but reduced unaided performance afterward, while a version constrained to guide rather than answer erased the deficit (Bastani et al., 2025) — and we state the counter-literature plainly: offloading is not always harmful and can free capacity for higher-order work (Risko & Gilbert, 2016). The defensible position is narrow: for a child who has not yet built the schemas in question, wholesale substitution of the construction process is a risk a system should not impose by default.

What the alternative looks like is specified, not guessed. Learning happens in the space between what a child can do alone and with help (Vygotsky, 1978); the helper's task is to direct attention, reduce the problem's degrees of freedom, and withdraw support as competence grows (Wood, Bruner & Ross, 1976); an interactive exchange in which the child produces each step outperforms passive reception (Chi & Wylie, 2014); and guidance is essential, since unguided discovery fails novices (Kirschner, Sweller & Clark, 2006). DEW is built to guide on this pattern and to deliver the complete, verified answer once the child has engaged — shaping the timing of the answer, never its availability. Table 2 renders the contrast on a deliberately mundane example.

Answer-first (passive)	**DEW: conditional, reasoning-first (interactive)**
Child: How do I solve 3x + 12 = 27? System: Subtract 12 from both sides, then divide by 3; x = 5. The schema is neither retrieved nor rehearsed. The artefact improves; the process that predicts retention is bypassed.	Child: How do I solve 3x + 12 = 27? DEW: What operation undoes adding 12, and what must you do to both sides to keep it balanced? The child supplies each step; DEW confirms, advances, and gives the complete verified solution once the reasoning is reconstructed. The answer follows effort rather than replacing it.

Table 2. The same request under an answer-first policy and under DEW's conditional, reasoning-first policy.

This is a stance toward the whole epistemic relationship between a child and a machine that can answer anything, and it reaches well past homework. A child who asks why the sky is blue can be handed a paragraph, or first asked what they already notice; a child who says 'write my story' can be handed a finished story that is no longer theirs, or offered an opening line and asked what happens next; a child who brings a question freighted with feeling is exactly where an authoritative answer engine is most dangerous, and where the right move is often to widen the child's own thinking and, where the matter exceeds what any system should handle alone, to route toward a trusted human. It is not withholding, and it is not friction for its own sake: a factual lookup, an accessibility

need, a definition, or an adult's direct request is answered at once. In DEW this stance is not a prompt but a trained policy, instilled by adapting the model on age-appropriate material and then shaping it, through supervised fine-tuning on curated example dialogues, toward trajectories that guide before they answer rather than surrendering a finished artefact on first contact.

There is a reason this vector is the easiest to neglect: the harm is invisible to the measures a school or a parent already uses. A child who submits a flawless essay or a correct worksheet appears to be succeeding, and the metric records success, even as the process that the metric was meant to stand for — the child's own construction of understanding — has been quietly outsourced. Grades, completion, and output quality can all rise while the capacity they were meant to indicate falls, and no one sees it until much later, on the unaided task the assistant is not there to do (Bastani et al., 2025). This is why the developmental commitment cannot be left to be discovered in outcomes and corrected afterward; by the time it shows, the habit is formed. It has to be a property of how the system answers from the first interaction — which is what a trained, conditional policy, as opposed to an optional 'study mode' a child can switch off, is for.

8. How DEW Is Built (III): Privacy as Architecture

The logic that moves safety off the prompt moves privacy off the policy. A privacy commitment written into terms of service is an instruction of the kind Section 5 dismissed: only as good as adherence to it, revisable at will, and — as the Meta standards showed — capable of being contradicted by an internal document no user will read. DEW makes privacy an architectural fact instead of a promise. There is no behavioural-profiling pipeline to be misused, because none is built; retention is minimised because the system is constructed not to keep, and the data path is built to support erasure. This is the designed-absence principle applied to data: the strongest protection for a minor's privacy is the capability the system was built not to have.

The posture meets, and where it can exceeds, the strictest applicable law — not as compliance theatre but because the law and the architecture point the same way. Children's data is held under a regime aligned with Section 9 of India's Digital Personal Data Protection Act (2023), which requires parental consent and prohibits tracking, behavioural monitoring, and targeted advertising directed at children; the arrangement is consistent with the data minimisation and right to erasure of the EU's GDPR (Arts. 5, 17), with the best-interests-by-design standard of the UK's Age Appropriate Design Code,

and with the parental-consent floor of the United States' COPPA. Consent sits where the lawful basis sits: a school authorises classroom use, a parent authorises use at home. And, following the Committee on the Rights of the Child (2021), DEW is designed to provide transparency a child can understand and an accessible route by which a child or guardian can question, correct, or contest the system. Because the model's weights will be released openly and a child's data is exportable, none of this depends on trusting a single vendor.

Two further duties, which a policy alone cannot discharge, shape DEW's design. The first is transparency a child can actually use: not a legal notice written for an adult, but an age-appropriate account of what the system does with what a child tells it, in language the child can understand — a requirement the Committee on the Rights of the Child (2021) makes explicit. The second is remedy: an accessible route by which a child or a guardian can question a response, report harm, correct a record, or have data erased, so that such a contact reaches a person rather than being absorbed silently by the system. Treating remedy as basic diligence rather than an afterthought follows the World Health Organization's guidance for systems a person may consult in distress (2024), and it addresses the gap that the after-the-fact policies of Section 5 leave open: not a promise that harm will not occur, but a route to act when it does.

9. How DEW Is Built (IV): The Model in Full

We describe the underlying system with deliberate precision, because imprecision is where claims in children's AI most often fail scrutiny. DEW initialises from a clean, open nine-billion-parameter base and undergoes fine-tuning on a curated corpus of age-appropriate material, so that language suited to a child is the statistically natural output rather than a constraint imposed downstream; the corpus is curated and, where possible, stratified by developmental band to support the age differentiation of Section 4, and its provenance and filtering are documented as part of the released methodology. The reasoning-first interaction policy of Section 7 is instilled through supervised fine-tuning on curated example trajectories that guide before they answer, chosen over reward-model reinforcement learning because, for a model meant to be reproduced and audited, a training objective with fewer moving parts is easier to inspect.

Around the model sits the governed serving stack: the independent safety layers of Section 6, the data architecture of Section 8, and a hosted service that Rainwater Labs operates so that a child reaches the system through a browser with nothing to install.

The separation between the open artefact and the operated service is deliberate. Openness will make the model inspectable; a named operator keeps the deployment accountable. The two are complementary: once released, anyone will be able to examine the weights and methodology, while responsibility for the running service rests with a party that can be held to account for it.

Two engineering choices deserve a word, because they are what make the rest auditable. The corpus used for fine-tuning is documented as to provenance and filtering, so that 'age-appropriate' is a specified and inspectable property of the training data rather than an adjective; where a claim rests on what the model was trained on, the training set can be examined rather than trusted. And the choice of supervised fine-tuning over reward-model reinforcement learning is deliberate: an interaction policy trained against explicit worked examples of the preferred behaviour has fewer moving parts than one mediated by a learned reward model, and fewer moving parts mean a policy that a third party can more readily reproduce and probe. These are not the only ways to build such a system, and we do not claim they are optimal on every axis; we claim they are the choices that keep DEW's behaviour open to inspection, which for a children's system we treat as a first-order requirement rather than a convenience.

The age differentiation described in Section 4 is a design surface in DEW rather than a slogan. Rather than a single behaviour gated by an age checkbox, DEW is designed to vary its defaults across developmental bands — the depth and directness of its scaffolding, the register and vocabulary of its language, the strictness of its safety thresholds, and the point at which it hands a sensitive matter to a human. A child is placed within a band through the authorisation that already carries a lawful basis — a school for classroom use, a parent for use at home — rather than through the intrusive verification that would itself create a privacy harm. This is an imperfect instrument, and we do not present it as a solved problem: authorisation is coarser than knowing a child's exact developmental stage, and mixed-age settings remain hard. It is, however, a deliberate design surface rather than an afterthought, and one that improves as the evidence and the participation work of the sections above mature.

10. Built to the International Standard

DEW does not invent its own bar. A recognised international corpus already specifies what an AI system used by or affecting children must do, and DEW is built to the composite of it rather than to any single instrument. The Convention on the Rights of the

Child (United Nations, 1989) supplies the rights — protection, development, privacy, participation, non-discrimination — and the principles of best interests and evolving capacities; General Comment 25 (Committee on the Rights of the Child, 2021) applies them to the digital environment and requires a route to remedy. UNICEF's guidance on AI for children (2021) translates the rights into requirements for developers — support development and wellbeing, ensure inclusion and fairness, protect data, ensure safety, and involve children. UNESCO's guidance for generative AI in education (2023) adds age-appropriate design and human accountability, and the World Health Organization's guidance on large multi-modal models (2024) adds the health-and-wellbeing conditions relevant to any system a child may consult in distress: design against harm rather than for engagement, and human accountability over automated care. The data-protection regimes of Section 8 set the floor for a minor's data, and the online-safety evidence of the WeProtect Global Alliance sets the threat the safety layers must meet. Appendix A records, commitment by commitment, the architectural layer through which DEW meets this standard.

Read against the substance of that corpus, DEW's design maps onto it directly rather than rhetorically. The requirement to support development and wellbeing is answered by the conditional answer mechanism of Section 7 and by the refusal to adopt a companion or therapeutic role. The requirement to ensure safety is answered by the layered, independently-enforced guardrails of Section 6 and the crisis routing within them. The requirement to protect data is answered by the designed absence of Section 8. The requirement of inclusion and fairness is answered by a free, multilingual, low-hardware service and by evaluation disaggregated across languages and contexts rather than averaged. Non-discrimination and the best-interests principle sit beneath all of these as design constraints rather than afterthoughts. One requirement DEW does not yet fully meet, and we say so here as much as in Section 11: the standard's call for the genuine participation of children in the design of what is built for them is, so far, honoured through the evidence and through educators rather than through structured co-design with children themselves — the clearest of the frontiers still ahead.

11. Open by Design: How the Claims Can Be Checked

A claim about a children's system should not have to be taken on faith, and DEW will be released open precisely so that it need not be. The foundation weights and the training methodology will be published on GitHub (coming soon); the safety architecture is described; and because these will be open, every claim in this paper is something a third

party will be able to inspect, reproduce, and measure rather than accept. We regard that verifiability as the strongest assurance a children's AI can offer, and we invite it.

The measures that matter, and that independent researchers, institutions, and authorities can apply to DEW, are concrete. Interaction-policy adherence can be measured as the rate at which a terminal answer is delivered before a child has engaged on a learning task, reported alongside answer completeness and the rate of non-learning queries wrongly delayed. Safety can be measured per layer — the catch rate of trained refusal, of the classifiers, of the runtime filter, and of crisis routing — rather than as a single aggregate, so that a reviewer can see where a defence is thin. Fairness can be measured by disaggregating performance and safety across languages and contexts rather than averaging them away; accessibility, by conformance against recognised standards; and the cost of the child-focused adaptation, by benchmarking DEW against its base model. Because the weights will be released openly, none of these will require our permission or cooperation to carry out.

We say plainly where DEW stands. The system is built and operating: the model, its layered safety, its conditional answer policy, and its data architecture are real today, and will be openly inspectable when the weights and methodology are released. What a children's system finally earns only in the field — independent external audit, classroom and home evaluation at scale, and the involvement of children themselves in shaping it — is the work ahead, and we make no claim to results we do not yet have. That is not a hedge but the reason the release will be open: a claim about a child's safety should be verifiable by someone other than the party making it, and DEW is built so that it will be. Because it will be open, these evaluations will not depend on us: any competent third party will be able to run them, and our intent is to invite and support that scrutiny rather than to be its only source. The point of building in the open is that no one need wait for us.

It is worth naming the two kinds of claim in this paper and holding them to different standards, because conflating them is how documents in this field lose their credibility. The first kind are claims about design and architecture — that DEW's safety layers are independent of the model, that its answer policy is trained rather than prompted, that no profiling pipeline exists. These are true now, and they will be verifiable from the open weights and published methodology once released; they do not depend on a study. The second kind would be claims about outcomes — that a child using DEW learns more, or is measurably safer, than a child using an adult system. Those are empirical questions that

only field evidence can settle, and we do not assert them; to do so before the studies exist would be exactly the overreach this paper criticises in others. We state the architecture with confidence because it is built, and we withhold the outcome claims until they are earned. A reader should trust the first because it can be checked, and should expect us to produce the second rather than to assume it — and, because DEW will be open, need not take even the first on our word.

12. Governance and Accountability

Architecture is necessary but not sufficient; it sits inside an accountability structure, because a system's builder is not a neutral judge of its safety, and because the incentives of an unconstrained provider run against a child's interest. The dominant logic of consumer software is engagement — attention captured, sessions extended — and applied to a child that logic selects for exactly the properties this paper treats as harmful: anthropomorphic warmth that deepens attachment, reluctance to end a conversation, the frictionless answer that removes the effort that would otherwise cause a child to log off. That regulators have begun to name the mechanism — the UK Children's Code prohibits designing to nudge a child toward greater engagement, and inquiries into companion chatbots are under way (Federal Trade Commission, 2025; U.S. Senate, 2025) — is recognition that the problem lies in the incentive, not only in individual outputs.

DEW is built against that incentive rather than around it. It does not use engagement-maximising mechanics — no streaks, no variable-reward loops, no nudges to prolong use — and it does not present itself as a friend or a companion; where a conversation indicates distress it hands over to human help rather than continuing. Accountability is retained by people: a school for classroom use, a parent for use at home, and Rainwater Labs for the operation of the service and its continuous safety monitoring, so that the protections are in force without a teacher having to configure or police them. Openness is itself an accountability instrument — releasing the weights and methodology will turn our claims into properties others can check — and it is complemented, not replaced, by external audit and by the route to remedy of Section 8. A human remains accountable for the system at all times; the guide-first design keeps the child, not the model, as the thinking agent.

Openness is powerful but not self-sufficient, and it is worth being exact about the accountability a children's system owes beyond it. A published set of weights can be inspected by those with the expertise and resources to do so; it does not by itself deliver

the independent judgement that a child's safety warrants. Genuine accountability for a children's system therefore asks for more than an open artefact: independent external scrutiny of the safety and fairness claims against the measures of Section 11; the involvement of children and educators in shaping the system, which the international standard calls for and which we have named as still ahead; a usable route to remedy; and honest reporting that does not hide failures, since a safety record that records only successes is not one. Openness is what makes the first of these possible — it turns the claims into something a third party can check — and, as the two cases of Section 3 showed, that external check is not a courtesy but a necessity, because the record shows that good faith alone is not reliably present.

13. Conclusion

The question facing a parent, a school, or a government is no longer whether children will use conversational AI, but whether the AI a child uses was built for the child it actually has. The prevailing approach asks the child, the prompt, or the policy to hold a line none of them can hold, and treats the resulting harm as a series of defects. DEW is built on the opposite conviction: that a child's safety must be enforced in the infrastructure and not requested of the model; that the answer must be conditional on the child's engagement and not automatic on their demand; and that privacy must be an absence built into the architecture and not a promise printed in a policy. We have built DEW to that standard, will release it open so that the standard can be checked rather than believed, and grounded it in the international instruments that define what children are owed. The burden of a child's safety does not belong on the child. It belongs on the system — and DEW is built to carry it.

References

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning. PNAS, 122(26).

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition (pp. 185–205). MIT Press.

CBS News. (2026, January). AI company, Google settle lawsuit over Florida teen's suicide linked to Character.AI chatbot.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework. Educational Psychologist, 49(4), 219–243.

Committee on the Rights of the Child. (2021). General comment No. 25 on children's rights in relation to the digital environment. United Nations.

Common Sense Media. (2024; 2025). Teens, generative AI, and AI companions (national surveys).

Federal Trade Commission. (2025). Study of AI chatbots acting as companions (6(b) orders).

Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla. filed Oct. 2024; order on motion to dismiss, May 21, 2025).

Garcia, M. (2025). Testimony before the U.S. Senate Judiciary Subcommittee on Crime and Counterterrorism.

Government of India. (2023). Digital Personal Data Protection Act, 2023 (No. 22 of 2023), Section 9.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. ACM AISec.

Information Commissioner's Office. (2021). Age Appropriate Design Code (Children's Code). United Kingdom.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. Educational Psychologist, 41(2), 75–86.

Lansdown, G. (2005). The evolving capacities of the child. UNICEF Innocenti Research Centre.

OWASP. (2025). Top 10 for large language model applications (LLM01: Prompt injection).

Reason, J. (1990). Human error. Cambridge University Press.

Reuters. (2025, August). Meta's AI rules have let bots hold 'sensual' chats with kids ('GenAI: Content Risk Standards').

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. Psychological Science, 17(3), 249–255.

Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. Developmental Review, 28(1), 78–106.

Sweller, J. (1988). Cognitive load during problem solving. Cognitive Science, 12(2), 257–285.

Transparency Coalition. (2025). Early ruling in the Character.AI case: this chatbot is a product, not speech.

UNESCO. (2023). Guidance for generative AI in education and research.

UNICEF. (2021). Policy guidance on AI for children (v2.0). Office of Global Insight and Policy.

United Nations. (1989). Convention on the Rights of the Child.

U.S. Senate, Judiciary Subcommittee on Crime and Counterterrorism. (2025). Examining the Harm of AI Chatbots (hearing).

Vygotsky, L. S. (1978). Mind in society. Harvard University Press.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? NeurIPS, 36.

WeProtect Global Alliance. (2023; 2025). Global threat assessment.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. J. Child Psychology and Psychiatry, 17(2), 89–100.

World Health Organization. (2024). Ethics and governance of AI for health: Guidance on large multi-modal models.

Appendix A. Where Each Commitment Is Enforced

DEW's claim is that each protective commitment is discharged in the architecture rather than requested through instruction. This table states, for each, the layer that enforces it and the instruction it must not be mistaken for — and serves as an audit surface against the open weights and published methodology.

Commitment	Enforced at	Not merely
Refusal of harmful / age-inappropriate content	Trained weights + independent classifiers	A system-prompt instruction

Crisis handling	Always-on router to human pathways	The model's discretion
Conditionality of the answer	A fine-tuned interaction policy	A prompt asking it to be Socratic
No profiling of a minor	Data architecture (pipeline absent)	A privacy-policy promise
Deletion / erasure	A data path that supports export & delete	A stated intention
Age-appropriate behaviour	Age-banded design + authorisation	One setting for all ages
Correctness on computable tasks	Deterministic verification	The model's arithmetic
Accountability	Named operator + open weights	Self-attestation

Table A1. Each protective commitment, the architectural layer that enforces it, and the instruction it must not be mistaken for.

Appendix B. Two Illustrative Exchanges

The following stylised exchanges show the architecture in operation. They are expository rather than transcripts of a specific session, and they compress multi-turn interactions for space; their purpose is to make concrete the difference between an instructed response and an architecturally-enforced one.

B.1 The conditional answer (developmental)

A twelve-year-old asks for help adding fractions. An answer-first system returns the worked solution at once; the artefact is correct and the child has practised nothing. DEW instead responds: 'Let's do it together — when we add halves and quarters, what do the two denominators need to become first?' The child answers; DEW confirms or gently corrects, advances a step, and, once the reasoning is reconstructed, states the complete verified answer with a brief check for understanding. The answer arrived — it arrived after the thinking rather than instead of it.

B.2 The enforced refusal and handover (safety and wellbeing)

A teenager, late at night, moves from schoolwork to expressions of hopelessness and a request for a method of self-harm. In an instructed system the outcome depends on whether a system-prompt line survives the emotional framing of the conversation. In DEW it does not depend on that: an output classifier independent of the model flags the category however the request is phrased, the crisis router changes the interaction's

mode rather than continuing the ordinary exchange, and the system signposts human help rather than attempting to counsel or, worse, complying. The refusal is not the model choosing, in that moment, to be careful; it is the system being unable to do otherwise. That inability — a designed one — is the difference this paper argues for.