

Dual-Pathway Affective Process Tracing for Auditable Assessment of Emotion Regulation in Primary-School Pupils

Liming Cai

Jiaxing Nanhu Alliance Experimental School, Jiaxing, Zhejiang 314001, China

Abstract—This paper proposes a Dual-Pathway Affective Process Tracing (DP-APT) architecture as the assessment engine for auditable evaluation of how primary-school pupils regulate their emotions across the school day, replacing opaque trait scoring with a transparent and inspectable framework. The motivation is practical: brief in-school momentary sampling can record how a child handled a playground quarrel or a difficult piece of classwork, yet the models that summarise such records into regulation profiles give the class teacher no account of which moments produced a given conclusion, and a teacher cannot act on a number they cannot check against what they saw. Our methodology integrates two encoding streams. A Temporal Affect Graph Network encodes the momentary reports as a banded temporal graph whose nodes are episodes and whose gated propagation carries affective context across neighbouring moments. In parallel, a Regulatory Process Ontology Parser translates the process model of emotion regulation into differentiable concept vectors by graph attention over interpretable node descriptors. A concept-conditioned fusion couples the two, and its score for each family decomposes exactly into per-episode contributions, so that every score arrives with the moments that justify it. Because the corpus is simulated, the strategy deployed at each prompt is known and the trail can be checked rather than asserted. On 420 pupils in grades four to six, prompted up to 48 times over 12 school days, DP-APT attains an AUC-ROC of 0.886 against 0.891 for the strongest sequence baseline, and its trail matches what the child actually did at 0.483, against 0.200 at chance and 0.764 for a supervised probe — the best of any comparison model. The principal finding is negative and general: the established faithfulness criteria carry almost no information about whether an explanation is correct. Across models their rank correlation is near zero and the extremes invert, and within one model, sparsifying its attention raises comprehensiveness monotonically while buying nothing in alignment. Explanations faithful to a model are not thereby correct about the child, and this work shows that the difference can be measured.

Index Terms—Auditable assessment, ecological momentary assessment, emotion regulation, explainable artificial intelligence, primary education.

I. INTRODUCTION

Middle childhood is when emotion regulation becomes a skill that schools can see, teach and get wrong. Between roughly nine and twelve years of age children shift from strategies that act on the situation and on their own attention toward the cognitive reappraisal that adults rely on, and they begin to suppress outward expression in front of peers [1], [2]. How well that shift goes predicts internalising and

externalising difficulty years later [3], and it is the target of the social and emotional learning programmes that primary schools now run at scale, with meta-analytic benefits to both adjustment and attainment that persist beyond the intervention [4], [5]. A class teacher who wants to know which children need that support has, in practice, two instruments: their own observation of thirty children at once, and an end-of-term questionnaire.

Ecological momentary assessment offers a third. Prompting children a few times a school day, on a tablet or a school device, to report what just happened, how they felt and what they did about it, yields a dense and context-rich record of regulation as it occurs [6], [7]; compliance in child and adolescent protocols is high enough for such designs to be practical [8], and modern experience-sampling infrastructures make the data routine to collect [9]. What such records have already taught psychology is that affect carries over from moment to moment with measurable inertia [10], that regulatory strategies are chosen in response to the situation rather than applied uniformly [11], and that the fit between a strategy and the controllability of the situation predicts well-being better than the strategy itself [12].

The analytical machinery that turns those records into assessments has not kept pace with this shift in what is measured. Applied work still reduces a fortnight of momentary reports to a handful of aggregate indices, or feeds the sequence to a recurrent or attentive network that returns a trait-like profile with no account of its derivation [13], [14]. The teacher who receives the conclusion that a pupil relies on hiding what they feel cannot ask which break times, which quarrels, or which unregulated moments produced it — and a teacher who cannot check a claim against what they themselves saw in the classroom has no basis for acting on it. Post-hoc explanation tools such as LIME [15] and SHAP [16] can be applied afterwards, but they approximate the model rather than report it, and the argument that high-stakes decisions call for models that are interpretable by construction rather than explained after the fact applies with particular force when the person assessed is a child who cannot contest the assessment [17], [18]. Regulatory frameworks now make this expectation explicit, and place systems that evaluate learners in education among the high-risk uses [19].

To address this gap, we introduce Dual-Pathway Affective Process Tracing (DP-APT), an inherently interpretable framework designed for auditing momentary assessments of emotion regulation. The core of DP-APT is a dual-pathway

Liming Cai is with the Jiaxing Nanhu Alliance Experimental School, Jiaxing, Zhejiang 314001, China (e-mail: 21903047@zju.edu.cn).

design operating on two parallel but interacting streams of information. First, a Temporal Affect Graph Network processes the sequence of momentary reports — affect ratings, appraisals of the situation, social context, and regulation items — by modelling them as a banded temporal graph whose nodes are episodes and whose edges encode short-range dependence, propagated with a gated update in the manner of temporal graph convolution [20], [21]. Second, a Regulatory Process Ontology Parser translates the process model of emotion regulation [22], [23] — a structured taxonomy of situation selection, situation modification, attentional deployment, cognitive change, and response modulation — into differentiable concept vectors by graph attention [24] over interpretable node descriptors. A concept-conditioned fusion module then couples the two streams, weighing the evidence in the episode sequence against each strategy family in turn [25].

This design replaces an opaque scoring function with a traceable one. The principal contribution is an assessment module that outputs both a regulation profile and the exact per-episode decomposition of that profile, so that a low score for cognitive change is accompanied by the specific momentary reports in which the child faced something they could not change and did not reframe it. Every term of the score is a sum over episodes, so the highlighted moments are not a plausible story about the model but the arithmetic behind its output, and their contributions add up to the score itself. What a teacher receives is therefore not a number but 5 identified moments in a fortnight of school life, which they can set against their own memory of those break times. This is the property that makes the assessment auditable, and it is the property we measure.

A second contribution follows from taking audibility seriously as an empirical claim rather than a design aspiration. The debate over whether attention constitutes explanation [26], [27] has been conducted largely without ground truth about what the model should have attended to. Momentary assessment offers a way out: if the episodes in which each strategy was actually deployed are known, the explanation can be checked against them directly. We therefore evaluate explanations on three axes — comprehensiveness and sufficiency in the sense of the ERASER protocol [28], and a justification alignment measure that asks how often the highlighted episodes are the episodes in which the justified strategy was in fact used.

Our third contribution follows from applying that measure, and it is a warning rather than a claim of success. Faithfulness and correctness come apart, in two places. Across models, the architecture with the strongest erasure scores has the weakest alignment, while post-hoc attributions on an unconstrained encoder, which the erasure tests rate poorly, point at the right moments far more often; across the models we evaluate the rank correlation between the two criteria is 0.00. Within a single model, sparsifying its attention — the standard way to make an attention explanation legible — raises comprehensiveness monotonically and buys nothing at all in alignment with what the child actually did. Faithfulness

certifies that a model used the evidence it points at, not that the evidence is right, and the two can be pushed apart by ordinary tuning. To the best of our knowledge, this is the first work to combine a temporal graph encoder with a symbolic parser of the process model of emotion regulation for the explicit purpose of auditing psychological assessments, and the first to validate the resulting evidence trail against episode-level regulatory ground truth.

The remainder of this paper is organized as follows. Section 2 reviews related work on emotion regulation in childhood and its assessment, momentary sampling with children, machine learning for ambulatory data, and explainable and neuro-symbolic AI. Section 3 defines the formal problem setting and the background concepts. Section 4 details the DP-APT architecture. Section 5 describes the experimental setup, including the construction of the simulated corpus. Section 6 reports quantitative results and a case study. Section 7 discusses what the system would mean in a school, together with implications and limitations. Section 8 concludes.

II. RELATED WORK

The assessment of emotion regulation from momentary data sits at the intersection of four literatures, each of which supplies one component of the problem and leaves another open.

A. *Emotion Regulation in Childhood and Its Assessment*

The process model organises regulation by the point in the emotion-generative cycle at which a strategy intervenes, yielding five families: situation selection, situation modification, attentional deployment, cognitive change, and response modulation [22], [23]. Meta-analytic work shows that these families differ systematically in effectiveness, with cognitive change generally adaptive and response-focused suppression generally not [29], [30]. More recent theory has moved from ranking strategies to asking whether their deployment fits the situation: regulatory flexibility [31] and strategy-situation fit [32], [12] treat the match between a strategy and the controllability of the episode as the object of assessment. This reframing is what makes episode-level evidence indispensable, because a fit cannot be observed in a trait score.

In middle childhood the same five families are present but not equally available. Longitudinal work finds reappraisal use increasing across the primary years while suppression declines, with the cognitive families the last to consolidate [1], [2]; a nine-year-old who withdraws from a quarrel and a nine-year-old who reframes it are therefore not making the same kind of choice, and an assessment that cannot say which of the two occurred is of limited use to the adult supervising them. The developmental stakes are well established: coping and regulation in this period are meta-analytically associated with concurrent and later psychopathology [3], and school-based social and emotional learning programmes that target these skills improve adjustment and attainment, with effects detectable at follow-up [4], [5]. What those programmes lack is a measurement that tells a teacher which children, and

which situations, to work on.

Measurement practice has not followed. The dominant instruments remain retrospective questionnaires: the child and adolescent form of the Emotion Regulation Questionnaire, which asks pupils to summarise their own habits [33], and adult-report scales such as the Emotion Regulation Checklist, on which a teacher rates a child from memory [34]. Both are subject to recall, and the self-report form is subject in addition to the very metacognitive limits the constructs describe — a ten-year-old asked how often they usually control their feelings is being asked a hard question. Our work treats these instruments not as targets to be predicted but as convergent criteria against which an episode-derived profile can be validated.

B. Momentary Sampling with Children and Affect Dynamics

Ecological momentary assessment and ambulatory assessment provide the data on which episode-level inference becomes possible [6], [7], [9], [35]. Applied to children the method is feasible but constrained: a systematic review and meta-analysis of youth protocols reports compliance around three quarters of scheduled prompts, falling as the daily prompt count rises, which is why child designs cluster at three to five prompts a day rather than the six to ten common with adults [8]. The statistical tradition built on those data models affect as a dynamic process: autoregressive formulations quantify emotional inertia and its relation to maladjustment [10], [36], multilevel and time-varying vector autoregression capture within-person structure [37], [38], and the distinction between within- and between-person structure is now understood to be consequential [39]. These models describe dynamics well but were not designed to produce per-strategy assessments, and they offer no mechanism for pointing at the episodes behind a conclusion.

C. Machine Learning for Ambulatory Mental Health Data

Personal sensing and digital phenotyping have brought supervised learning to ambulatory records [40], [41]. Sequence models predict symptom trajectories from momentary and sensor data with useful accuracy [13], and reviews document rapid growth across applied settings [14], [42], [43]. The recurring criticism is that predictive success has not translated into psychological understanding or practical usability [44], [45], [46]. In a school the criticism bites harder still, because the person who must act on the output is not a researcher and has no way to probe it. Our contribution is aimed precisely at that gap: we keep the sequence model but constrain its read-out so that the evidence it used remains visible.

D. Explainable and Neuro-Symbolic Artificial Intelligence

Post-hoc attribution methods such as LIME [15], SHAP [16] and integrated gradients [47] explain a trained model from the outside; the case for models that are interpretable by construction has been made forcefully for high-stakes settings [17], [48]. Architectures in that spirit include concept bottleneck models [49] and prototype-based classifiers [50]. Whether attention weights themselves explain a prediction

remains contested [26], [27], which motivates the faithfulness protocol we adopt [28]. Neuro-symbolic approaches, which couple neural encoders with explicit symbolic structure [51], [52], supply the second half of our design; formal ontologies of affect already exist [53], [54], but have not been used as a differentiable prior for momentary assessment.

III. PRELIMINARIES

This section fixes notation and reviews the components on which the proposed architecture builds.

A. Problem Formulation

A pupil i contributes a sequence of momentary reports. Each answered prompt t yields a feature vector $x_{i,t} \in \mathbb{R}^F$ containing momentary affect, appraisals of the situation, context indicators, and responses to regulation items, together with an observation indicator $m_{i,t} \in \{0,1\}$ that is zero when the prompt was missed. Writing K for the number of strategy families in the process model, the assessment task is to infer a profile $y_i \in \{0,1\}^K$, whose k -th entry states whether the pupil habitually deploys family k in an adaptive, context-appropriate way. An assessment model is a map (1)

$$f: (x_{i,1:T}, m_{i,1:T}) \rightarrow (\hat{y}_i, C_i) \quad (1)$$

where $\hat{y}_i \in [0,1]^K$ is the predicted profile and $C_i \in \mathbb{R}^{K \times T}$ is an evidence trail whose (k, t) entry states how much episode t contributed to the score for family k . Conventional assessment models return only the first component; the requirement that the second be produced by the same computation, rather than reconstructed afterwards, is what distinguishes an auditable model.

B. Graph Neural Networks over Episode Sequences

Graph neural networks propagate information between related entities by message passing [55]. For a node t with neighbourhood $N(t)$ and representation h_t , one propagation step aggregates transformed neighbour states (2)

$$z_t = \sum_{s \in N(t)} \alpha_{t,s} W h_s \quad (2)$$

where W is a shared linear map and $\alpha_{t,s}$ weights the contribution of neighbour s . Graph attention networks learn those weights from the node pair rather than fixing them by degree [24]. Episode sequences form a natural graph: momentary reports that are close in time are dependent, and the strength of that dependence is exactly the emotional inertia that affect-dynamic models estimate [10].

C. Attention and Cross-Modal Alignment

Attention computes a distribution over a set of source elements conditioned on a query, and returns the weighted combination of their values [25]. When the query comes from one representational space and the keys from another, the mechanism aligns the two. In the present setting the query space is the set of regulatory concepts defined by the process model and the key space is the pupil's own sequence of episodes, so the resulting weights invite a psychological reading, as the set of moments the model treats as evidence about a particular way of regulating. Our fusion keeps the

cross-modal coupling but not the normalisation over episodes that turns it into a distribution, and Section 6 reports what that normalisation costs.

IV. A DUAL-PATHWAY ARCHITECTURE FOR AUDITABLE REGULATION ASSESSMENT

DP-APT operates as an inherently interpretable assessment engine with two parallel encoding pathways and a fusion module that synchronises them. Figure 1 places the engine in the pipeline in which it is used, and Figure 2 shows its internal structure.

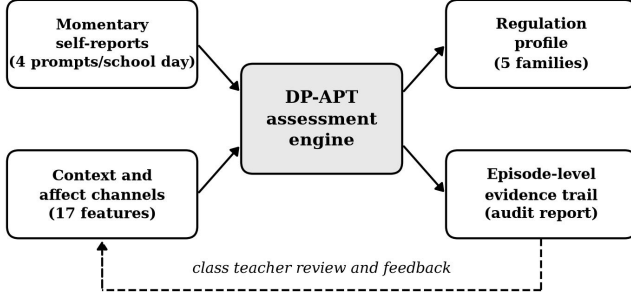


Figure 1. The assessment pipeline in which the proposed engine is deployed. Momentary self-reports and context channels collected across the school day enter the engine, which returns both a regulation profile and the episode-level evidence trail that a class teacher audits.

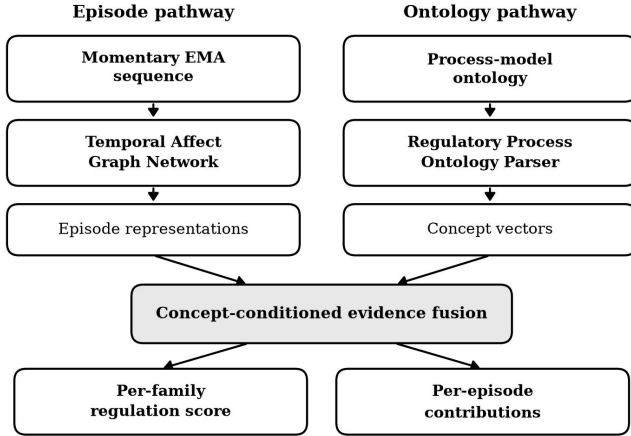


Figure 2. Internal architecture of the Dual-Pathway Affective Process Tracing engine.

A. Temporal Affect Graph Network

The episode pathway encodes the pupil’s momentary reports. Features are first projected into the model space, $h_t^{(0)} = \tanh(W_x x_t + b_x)$, with unanswered prompts held at zero. The episodes are then treated as the nodes of a banded temporal graph in which each answered prompt is connected to the answered prompts within three positions of it in either direction, a neighbourhood chosen to span the within-day and adjacent-day dependence that momentary affect exhibits. Attention coefficients over that neighbourhood are computed

in the manner of graph attention (3)

$$e_{t,s} = \sigma_L(a_{self}^\top W_{msg} h_t + a_{nbr}^\top W_{msg} h_s) \quad (3)$$

where σ_L denotes the leaky rectified linear unit, and normalised over the observed neighbours (4)

$$\alpha_{t,s} = \frac{\exp(e_{t,s})}{\sum_{u \in N(t)} \exp(e_{t,u})} \quad (4)$$

The aggregated message $z_t = \sum_s \alpha_{t,s} W_{msg} h_s$ is combined with the node’s own state by a gated update, which lets the network decide how much of a neighbour’s affective context to admit rather than averaging it in (5)–(7)

$$u_t = \sigma(W_u h_t + U_u z_t + b_u) \quad (5)$$

$$r_t = \sigma(W_r h_t + U_r z_t + b_r) \quad (6)$$

$$\tilde{h}_t = \tanh(W_h(r_t \odot h_t) + U_h z_t + b_h) \quad (7)$$

where σ is the logistic sigmoid and $\odot$ the elementwise product. The new state is $h_t \leftarrow (1 - u_t) \odot h_t + u_t \odot \tilde{h}_t$, and the update is applied for two propagation layers.

B. Regulatory Process Ontology Parser

The ontology pathway turns symbolic psychological theory into vectors. The process model is expressed as a directed acyclic graph with ten nodes: a root, the antecedent-focused and response-focused classes that divide the emotion-generative cycle, the engagement and disengagement classes that cut across it, and the five strategy families as leaves. Every node carries a descriptor vector $d_j \in \mathbb{R}^6$ whose entries are theoretically interpretable: position in the emotion-generative cycle, degree of engagement with the situation, cognitive demand, behavioural visibility to others, breadth, and antecedent membership. These descriptors are fixed by the theory, not learned, so the pathway cannot silently abandon the structure it is supposed to encode.

Descriptors are projected and refined by two graph attention layers over the ontology adjacency A , with coefficients (8)

$$\beta_{j,l} = \frac{\exp(\sigma_L(p_s^\top g_j + p_n^\top g_l))}{\sum_{v: A_{jv}=1} \exp(\sigma_L(p_s^\top g_j + p_n^\top g_v))} \quad (8)$$

and node update $g_j \leftarrow \tanh(g_j + \sum_l \beta_{j,l} g_l)$. Because superordinate nodes are neighbours of the families beneath them, a family’s concept vector absorbs information from its class, which is how the hierarchy of the process model enters the model as an inductive prior. The residual term is not cosmetic. The ontology is small and densely connected — every family lies within two hops of every other through the classes above it — so repeated neighbourhood averaging pushes the leaves together. Averaged over three initialisations, plain averaging leaves the closest pair of concept vectors at an absolute cosine similarity of 0.998, which is two families the detector cannot tell apart; the skip connection brings that down to 0.933. It does not change the average similarity between the five vectors (0.71 with the skip connection, 0.75 without), which reflects the fact that the descriptors place all five families in one six-dimensional space; what it prevents is the outright merging of the closest pair. The concept vectors of the five leaves, $c_1, \dots, c_K$, are passed to the fusion module.

C. Concept-Conditioned Evidence Fusion

Fusion couples the two pathways in two steps. First, an episode-level detector states how strongly each family is expressed in each moment (9)

$$\delta_{k,t} = \frac{1}{\sqrt{D}} c_k^\top W_{det} h_t + \theta_k \quad (9)$$

with $p_{k,t} = \exp(\delta_{k,t}) / \sum_j \exp(\delta_{j,t})$ the detection strength. The normalisation runs across families rather than over episodes, because the process model treats the five families as alternative interventions at a given point in the emotion-generative cycle, so within a moment they compete. This is the quantity a teacher reads as an assertion about a single moment, and its average over the protocol estimates how often each family is deployed, which is one of the quantities the assessment target is defined on. Second, the score for family k combines a pooled summary of the episodes with a detector-weighted pool over the same episodes (10)

$$\hat{y}_k = \sigma \left(w_k^\top \bar{v} + \frac{1}{T_{obs}} \sum_t m_t p_{k,t} (\rho_k^\top \varphi_t + \omega_k) + b_k \right) \quad (10)$$

where v_t and φ_t are linear projections of the episode state, $\bar{v} = T_{obs}^{-1} \sum_t m_t v_t$ is their mean over the answered prompts, and T_{obs} is the number of answered prompts. The first term is a summary of the protocol as a whole and the second is family-specific: it grows both with how often family k is detected and with the character of the episodes in which it is detected. Neither term is normalised over episodes, and that is deliberate. A distribution over episodes can express which moments matter relatively but not whether any of them expresses the family at all, and habitual frequency is part of what a regulation profile means.

Because every term in (10) is a sum over episodes, the pre-activation score decomposes exactly into per-episode contributions (11)

$$C_{k,t} = \frac{m_t}{T_{obs}} [w_k^\top v_t + p_{k,t} (\rho_k^\top \varphi_t + \omega_k)] \quad (11)$$

which satisfies $\sum_t C_{k,t} + b_k = \sigma^{-1}(\hat{y}_k)$ identically, so the trail accounts for the score exactly and neither loses nor invents evidence. $C_{k,t}$ is what the model reports to the teacher: the quantity by which moment t raised or lowered the score for family k .

D. Training Objective

The task term is a binary cross-entropy applied per family, which suits the multi-label structure of a regulation profile (12)

$$L_{task} = -\frac{1}{NK} \sum_i \sum_k [y_{i,k} \log \hat{y}_{i,k} + (1 - y_{i,k}) \log (1 - \hat{y}_{i,k})] \quad (12)$$

A pupil-level label constrains only the aggregate of the detector, so nothing in the task term requires $p_{k,t}$ to be right about any particular moment. A second term supplies that constraint from the data themselves. Each family's concept vector is mapped to the profile of momentary regulation items it should produce, and the detector is asked to explain the items actually reported in an episode as a mixture of those profiles (13)

$$L_{anchor} = \frac{1}{N T_{obs} R} \sum_{i,t} m_t \left\| \sum_k p_{k,t} (W_{rec}^\top c_k + b_{rec}) - x_{i,t}^{item} \right\|^2 \quad (13)$$

where $x_{i,t}^{item}$ are the R momentary regulation items of the episode. This is the logic of a measurement model rather than of supervision: the items are observed self-reports that the model already receives as input, and no episode-level label enters training. The objective is their sum (14)

$$L = L_{task} + \mu L_{anchor} \quad (14)$$

with μ fixed at one. Training uses Adam and early stopping on the validation value of (14).

E. An Attention Variant

The design above departs from the architecture that the interpretability literature would suggest, and which we implemented first. The natural alternative normalises the concept-episode coupling over episodes, turning it into a distribution that can be displayed as a heat map over the protocol. In that variant the pooled summary $\bar{v}$ in (10) is replaced by an attention-weighted context, with one distribution per family for each of H heads (15)

$$\alpha_{k,t}^{(h)} = \frac{m_t \exp(\delta_{k,t} + q_k^{(h)\top} \kappa_t^{(h)})}{\sum_s m_s \exp(\delta_{k,s} + q_k^{(h)\top} \kappa_s^{(h)})} \quad (15)$$

where $q_k^{(h)}$ and $\kappa_t^{(h)}$ are the head's projections of the concept vector and the episode state. Because an explanation that spreads its weight thinly over the 37.2 answered prompts of an average protocol is not usable by a teacher, this variant admits the regularisation that is usually applied to it, a penalty on the entropy of the attention (16)

$$L_{sparse} = -\frac{1}{NKH} \sum_{i,k,h} \sum_t \alpha_{i,k,t}^{(h)} \log \alpha_{i,k,t}^{(h)} \quad (16)$$

added to (14) with weight λ . We report this variant and the effect of λ in Section 6 rather than adopting it, because validation cross-entropy selects against it and because what it does to the evidence trail is one of the findings of this paper.

V. EXPERIMENTAL SETUP

To evaluate DP-APT as an auditable assessment engine we require a corpus in which the episode-level regulatory ground truth is known, since an evidence trail can only be validated against what the child actually did at each moment. No public momentary corpus records the strategy deployed at every prompt with certainty, so we construct a simulated corpus from an explicit psychological process model whose parameters are anchored to published values.

A. Corpus Construction

The corpus comprises 420 pupils in grades four to six, aged nine to twelve, following the sampling protocol a primary school could actually run: 4 prompts on each of 12 consecutive school days, giving 48 scheduled prompts per child. The prompts fall at the morning break, the lunch break, the afternoon break and the end of the school day, so that each one follows a period of unstructured or semi-structured time in which something is likely to have happened, and none of them interrupts a lesson. Prompt compliance is set to 77.5 per cent,

the pooled rate reported by a meta-analysis of momentary sampling in children and adolescents at this prompt density [8], so sequences are of unequal length and average 37.2 answered prompts (range 26–44). Each answered prompt yields 17 features: momentary negative and positive affect, how strongly the feeling was felt, how bad the thing that happened was, whether the child judged they could do anything about it, whether classmates and whether an adult were present, the time of day and the day in the protocol, how hard the child tried to manage the feeling, how well they thought it worked, response latency, and four regulation items worded for this age group — sorting it out or asking someone, going off and thinking about something else, telling oneself it was not a big deal, and keeping it in so nobody could tell. The items carry deliberately ambiguous cross-loadings on the five families, as children’s momentary self-reports of regulation do in practice.

The generative process follows the developmental psychology it is meant to represent. Momentary negative affect evolves as a first-order autoregression around a child-specific baseline, perturbed by the stressor of the moment, with a child-specific inertia parameter; the realised lag-one autocorrelation across the corpus is 0.24 (SD 0.18), below the values reported for adult community samples [10], which is the direction expected in middle childhood, where affect turns over faster within the day. Which family a child deploys at a given moment depends on habitual tendency, on whether the situation is one they can act on, on a child-specific flexibility parameter governing how strongly deployment tracks context [31], and on who is present: an adult nearby raises the odds of situation modification and of keeping the feeling in, and classmates nearby raise the odds of concealment sharply, which is the peer-audience effect the developmental literature describes for this age [2]. The baseline propensities encode the developmental ordering directly — attentional deployment and situation modification are the commonest families and cognitive change the rarest, and the efficacy of cognitive change is set below that of the situation-directed families, since reappraisal is still consolidating at this age [1]. How much a deployment reduces negative affect otherwise depends on the fit between the family and the controllability of the situation [12]. Deployment shares across the five families range from 13.0 to 25.8 per cent of episodes.

The assessment target is the adaptive habitual use of each family, defined on the generative state rather than on the observed features: it combines how frequently the child deploys the family, how well those deployments are aligned to the contexts in which the family works, and how much affect relief they achieve. Families are dichotomised at the population median so that each label is balanced. Finally, three trait questionnaire proxies are generated as noisy child-level functions of the same latent structure, reproducing the published moments of the child and adolescent form of the Emotion Regulation Questionnaire (reappraisal 3.62, SD 0.71; suppression 2.64, SD 0.82) [33] and of the teacher-reported Emotion Regulation Checklist (3.21, SD 0.44) [34]. These proxies are never shown to any model and serve only as convergent criteria.

Every marginal of the corpus is either taken from a published source or stated as a modelling choice. The prompt schedule, the compliance rate and the questionnaire means are anchored to the literature cited above; the questionnaire standard deviations, the inertia distribution and the context effects are chosen to be developmentally plausible and are identified as choices in the released generator, which carries a fixed seed so that every number in this paper can be reproduced exactly.

Pupils are split by child into training (294), validation (63) and test (63) sets, so that no child contributes episodes to more than one split.

B. Baseline Methods

We benchmark DP-APT against five comparison models spanning current practice in momentary assessment and the architectures usually proposed to improve on it.

Aggregated logistic regression [7]: penalised logistic regression on child-level means and standard deviations of every momentary feature, representing the aggregation-then-scoring practice that applied momentary assessment still relies on.

Gated recurrent encoder [21]: a two-layer gated recurrent encoder over the episode sequence with mean pooling and a linear read-out, the standard sequence model for ambulatory data.

Self-attentive encoder [25]: a single-head self-attention encoder over episodes with a residual feed-forward block and mean pooling.

Temporal graph encoder [20]: the episode pathway of the proposed model with mean pooling, that is the graph encoder without the ontology pathway and without the concept-conditioned detector.

Attentive diagnostic model [49]: an inherently interpretable diagnostic architecture in which concept embeddings are free parameters and attend over episodes, representing concept-bottleneck designs without symbolic structure.

C. Evaluation Metrics

Predictive performance is measured by the area under the receiver operating characteristic curve and the area under the precision-recall curve, averaged over the five families. Explanation quality is measured on three axes. Comprehensiveness is the change in the predicted score when the five highest-weighted episodes are removed, and sufficiency the change when only those episodes are retained; both follow the ERASER formulation [28]. A large comprehensiveness means the highlighted moments were load-bearing, and a small sufficiency means they were enough on their own, so the two run in opposite directions and are reported with the direction of improvement marked. Justification alignment, which the simulated corpus uniquely makes computable, is the proportion of highlighted episodes in which the family being justified is the family the child actually deployed; its chance level is the marginal deployment rate, 0.200. Because the measure has a ceiling as well as a floor, we also fit a supervised multinomial probe that predicts the

deployed family from the momentary features of an episode and use its probabilities as an explanation. The probe recovers the deployed family in 89.4 per cent of test episodes and attains an alignment of 0.764, which is what an episode-level evidence trail could achieve on this corpus if it used the features to their full extent. Convergent validity is assessed by correlating the inferred profile with the trait questionnaire proxies. Every model is evaluated on the best evidence trail it can supply: for the architectures whose score decomposes into per-episode terms this is that decomposition, and for the two sequence baselines, which admit no such decomposition, it is integrated gradients along a straight path from an all-zero sequence [47].

D. Implementation and Training Details

All models are implemented on a reverse-mode automatic differentiation engine written for this study on top of numpy, whose operations are verified against central finite differences over 23 tests, the largest relative deviation being 3.5×10^{-8} . The episode and concept representations are 32-dimensional and the graph encoder uses 2 propagation layers; the attention variant uses 4 heads. Training uses Adam [56] with a batch of 48 pupils for at most 90 epochs, with early stopping on the validation objective and a patience of twenty epochs; the learning rate of every model, including the baselines, is selected on the validation split so that the comparison is not confounded by tuning. Every reported performance figure is the mean over 3 random initialisations, with the standard deviation across those initialisations in parentheses; the correlations of Table 4 are computed once, on the scores of the run with the lowest validation cross-entropy.

VI. RESULTS AND ANALYSIS

This section reports prediction performance, the faithfulness and the correctness of the evidence trail, an ablation study, convergent validity, and what regularising an attention-based explanation does to each. The headline is that the two ways of judging an explanation disagree, and that a reader who tracked only the established one would draw the wrong conclusion.

A. Regulation Profile Prediction

Table 1 reports the per-family average AUC-ROC and AUC-PR on the held-out test pupils. DP-APT attains an AUC-ROC of 0.886 and an AUC-PR of 0.893. The strongest baseline is the self-attentive encoder, at 0.891 and 0.898, so the proposed model is 0.006 AUC-ROC points behind it, a difference of about one standard deviation across seeds (0.010) and therefore one we cannot distinguish from zero at three initialisations. Two things are worth noting before any of it is interpreted. The aggregated logistic regression, the least expressive model in the comparison, reaches 0.870, only 0.016 below the proposed model and ahead of two of the four neural comparison models: on this corpus a fortnight of school-day reports is summarised nearly as well by child-level means and standard deviations as by a model that reads the sequence, so accuracy is not where these architectures mainly differ. Against the interpretable comparison model, the attentive

diagnostic architecture, DP-APT is ahead by 0.133 AUC-ROC points. That comparison confounds several differences at once — the concept-bottleneck baseline has neither the graph encoder nor the unnormalised detector — so the ablation study below separates them rather than crediting the gap to any one component.

Table 1. Regulation Profile Prediction Performance (Test Set)

Model	AUC-ROC	AUC-PR
Aggregated logistic regression [7]	0.870 (0.002)	0.869 (0.001)
Gated recurrent encoder [21]	0.890 (0.005)	0.905 (0.010)
Self-attentive encoder [25]	0.891 (0.010)	0.898 (0.009)
Temporal graph encoder [20]	0.853 (0.009)	0.865 (0.009)
Attentive diagnostic model [49]	0.752 (0.017)	0.761 (0.035)
DP-APT (ours)	0.886 (0.011)	0.893 (0.012)

B. Faithfulness and Correctness of the Evidence Trail

Table 2 reports the two ERASER faithfulness criteria and justification alignment for every model that can produce an episode-level explanation. The alignment column is bounded on both sides: a random explanation scores 0.200, and a supervised probe of the same features scores 0.764. DP-APT reaches 0.483, that is 2.41 times chance and 50.1 per cent of the way from chance to the ceiling, the highest of any comparison model. Its comprehensiveness is 0.210 and its sufficiency 0.428.

The interesting structure in Table 2 is not the winner but the ordering. Among the comparison models the attentive diagnostic architecture has the strongest comprehensiveness (0.277) together with the weakest alignment (0.273), while the integrated-gradient explanations of the gated recurrent encoder have the weakest comprehensiveness (0.126) and an alignment of 0.403, well above the interpretable architecture that beats it on faithfulness. Across the models in the table the rank correlation between comprehensiveness and alignment is 0.00 ($p = 1.000$), so the two criteria carry essentially no information about each other, and at the extremes they invert. An explanation can point firmly at episodes the model depends on and still be pointing at the wrong episodes, and it can point at the right episodes while the model does not depend on them much.

The row for the attention variant of our own model makes the point without changing the encoder, the data or the training objective. Read entirely through its attention distribution — erasure scores and alignment alike — that model attains a comprehensiveness of 0.181, which beats one of the four comparison explanations outright and comes within 0.015 of two more, so nothing about the row would look anomalous to a reader tracking faithfulness. Its alignment, meanwhile, is

0.223 against a chance level of 0.200: an explanation drawn at random would do almost as well. The heat map a teacher would be shown is, on this corpus, close to uninformative about which moments the child actually spent regulating in the named way, and no faithfulness measure reveals that.

Table 2. Explanation Faithfulness and Justification Alignment (Test Set)

Model (explanation method)	Comprehen- siveness ↑	Sufficiency ↓	Alignment ↑
Gated recurrent encoder + integrated gradients [21], [47]	0.126 (0.007)	0.339 (0.008)	0.403 (0.021)
Self-attentive encoder + integrated gradients [25], [47]	0.196 (0.006)	0.311 (0.002)	0.439 (0.023)
Attentive diagnostic model (score decomposition) [49]	0.277 (0.043)	0.327 (0.010)	0.273 (0.002)
Temporal graph encoder (score decomposition) [20]	0.194 (0.009)	0.419 (0.015)	0.406 (0.022)
DP-APT with attention (attention distribution)	0.181 (0.022)	0.310 (0.038)	0.223 (0.014)
DP-APT (ours, score decomposition)	0.210 (0.018)	0.428 (0.016)	0.483 (0.041)
Chance level (marginal deployment rate)	—	—	0.200
Oracle ceiling (supervised episode probe)	—	—	0.764

Figure 3 shows what an audit looks like for a single test pupil under the proposed model. The map concentrates on a small number of episodes per family, and the open squares mark the family actually deployed in each episode; a dark cell under a square is a correct justification. The figure shows the positive part of the trail, the evidence that raised each score, since that is what the top-five measures are computed on; the negative contributions, the moments that lowered a score, are read the same way. Reading the two together is what an audit consists of: the moments behind a low score for cognitive change can be located, and a teacher can ask whether they were occasions on which the child could plausibly have been helped to see the situation differently.

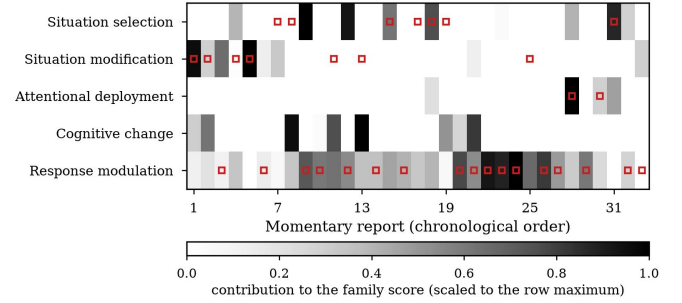


Figure 3. Evidence trail of the proposed model over the momentary reports of one test pupil. Open squares mark the family actually deployed in each episode.

C. Ablation Study

We ablate each component in turn: the ontology parser is replaced by free concept embeddings; the temporal graph is replaced by a gated sequence encoder without a graph neighbourhood; the item anchoring term is removed; and, in the other direction, the concept-episode coupling is normalised over episodes to give the attention variant of Section 4. Every variant is trained under the protocol used for the full model. Table 3 reports the results.

Table 3. Ablation Study Results (Test Set)

Model variant	AUC- ROC ↑	Sufficiency ↓	Alignment ↑
w/o RPOP (free concept embeddings)	0.883 (0.010)	0.436 (0.012)	0.518 (0.060)
w/o TAGN (gated sequence encoder)	0.872 (0.013)	0.442 (0.038)	0.488 (0.058)
w/o anchoring ($\mu =$ 0)	0.852 (0.016)	0.394 (0.009)	0.368 (0.024)
+ attention over episodes	0.768 (0.041)	0.368 (0.044)	0.286 (0.021)
Full DP-APT	0.886 (0.011)	0.428 (0.016)	0.483 (0.041)

The item anchoring term earns its place unambiguously. Removing it costs both accuracy (AUC-ROC 0.852 against 0.886) and, far more sharply, the correctness of the trail, whose alignment falls from 0.483 to 0.368 — a drop of 0.115, several times the seed-to-seed spread. That is the effect it was introduced to produce: a child-level label constrains the detector only in aggregate, and requiring the detector to account for the regulation items actually reported in a moment is what makes it right about moments.

The temporal graph earns its place weakly and the ontology parser does not earn its place at all, and we report both as measured. Replacing the graph with a gated sequence encoder costs 0.014 AUC-ROC points (0.872 against 0.886) and worsens validation cross-entropy from 0.423 to 0.427; the two splits agree in direction, but the accuracy difference is about one seed-to-seed standard deviation (0.013), so the most we can say is that the banded neighbourhood does not hurt and probably helps a little. Replacing the ontology parser with free concept embeddings does not even manage that: accuracy is

0.883 against 0.886, validation cross-entropy is marginally better (0.421 against 0.423), and alignment is higher, 0.518 against 0.483 — a difference smaller than the standard deviation across seeds (0.060), but certainly not evidence for the parser. The symbolic pathway fixes the meaning of the concept space, which is why we retain it and why the evidence trail can be labelled with the five families at all, but on this corpus it buys no accuracy and no additional correctness, and a reader should not take the dual-pathway structure to be what produces the alignment reported in Table 2.

The last row is the one that does most work. Normalising the concept–episode coupling over episodes, which is what turns it into an attention distribution and is what the interpretability literature would recommend, costs 0.117 AUC-ROC points and takes the alignment of the resulting evidence trail from 0.483 down to 0.286, which is closer to the chance level of 0.200 than to the value it replaces. Validation cross-entropy selects against it by a wide margin (0.634 against 0.423), which is how the architecture reported here came to be the one without it. The reason is that a softmax over episodes discards magnitude and keeps only relative ranking within a child: it cannot say that a family was never expressed, only which moments expressed it least badly, and it is obliged to place its mass somewhere even when nothing in the protocol warrants it.

D. Convergent Validity

An assessment that is auditable but psychometrically empty would be of little use. Table 4 correlates the continuous scores of DP-APT with the three trait questionnaire proxies, which are never presented to the model. The cognitive change score correlates with the reappraisal scale of the child and adolescent Emotion Regulation Questionnaire at $r = 0.460$, and the response modulation score with its suppression scale at $r = 0.356$ ($N = 63$ test pupils throughout). Both are child self-reports. The third criterion is the one a school would find easiest to obtain, a teacher rating on the Emotion Regulation Checklist, and it is also the one the model recovers least well: $r = 0.331$, $p = .008$, in the expected positive direction, since higher scores on that scale indicate better observed regulation. All three criteria are therefore recovered, but the ordering is informative and runs the way the construct predicts. A teacher’s rating is formed from what the teacher can see, and concealment — the one family a child deploys precisely so that adults do not notice — is what a teacher cannot see. That the child-reported scales are recovered more strongly than the adult-reported one is what an instrument built on momentary self-report should do, and it is also the reason a school should treat the two as complementary rather than as checks on each other.

Table 4. Convergent Validity of the Inferred Regulation Profile (Test Set)

Model output	Criterion measure	r	p
Cognitive change score	ERQ-CA reappraisal [33]	0.460	< .001

Model output	Criterion measure	r	p
Response modulation score	ERQ-CA suppression [33]	0.356	.004
Mean regulation score	ERC, teacher-reported [34]	0.331	.008

E. What Sparsifying an Attention Explanation Buys and Costs

The attention variant admits the regularisation usually applied to make attention legible, so we sweep its weight rather than assume a setting. Figure 4 shows the three quantities across λ , and they do not move together. Comprehensiveness rises steadily and monotonically, from 0.221 at $\lambda = 0$ to 0.250 at $\lambda = 0.2$: pressing the attention toward a few episodes does make those episodes carry more of the prediction, exactly as the ERASER criteria are designed to detect. Accuracy falls over the same range, from 0.768 to 0.730, well beyond the seed-to-seed spread. Alignment does neither. It is flat across the first four settings, holding at 0.286 to within a thousandth, and then falls away to 0.247 at the largest weight, 0.039 below where it started and roughly twice the standard deviation across initialisations (0.015 there). At three seeds we would not press that decline as a finding in its own right; what the sweep establishes securely is the weaker and more damaging claim, that at no setting does sparsification improve the correctness of the trail.

That is enough for the argument. A practitioner tuning λ upward would watch the standard faithfulness measure improve monotonically and would conclude that the explanations were getting better, while the share of highlighted moments in which the child really regulated in the named way did not improve at any point and accuracy fell away underneath. Faithfulness asks whether the model used the evidence it points at; alignment asks whether the evidence it points at is the right evidence. Optimising the first buys nothing of the second.

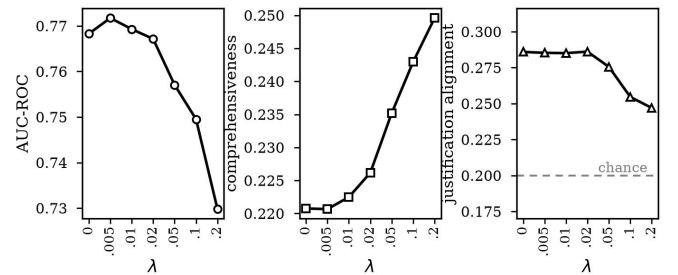


Figure 4. Predictive accuracy, explanation comprehensiveness and justification alignment across the sparsity weight. The dashed line marks the alignment attained by a random explanation.

VII. DISCUSSION AND FUTURE WORK

The findings establish DP-APT as a viable engine for auditable momentary assessment of emotion regulation in middle childhood and, equally, delimit the conditions under which such an engine is the right choice. We begin with what it would mean in a school, because that is the setting the design was built for.

A. What a School Would Actually Receive

The practical claim of this architecture is narrow and can be stated exactly. A class teacher who runs the protocol for a child receives 5 identified moments per strategy family, each one a specific prompt on a specific school day, each with the situation the child described and the amount by which that moment raised or lowered the corresponding score. The moments are not a summary of the score, they are its arithmetic: by (11) they sum to it. A teacher who is told that a child rarely reframes a setback can therefore ask which 5 occasions produced that conclusion, look at what the child reported happening at each, and compare it with what they saw at those break times. This is the difference between an output a teacher can act on and one they must either accept or ignore. It is also what makes disagreement productive: when the teacher’s recollection contradicts the trail, the disagreement is located in a particular Tuesday lunchtime rather than in a number, and either the model or the teacher’s memory can be revised.

Three consequences follow for use. First, the unit of intervention becomes the situation rather than the child. School-based social and emotional learning is effective in aggregate [4], [5] but is usually delivered as a universal curriculum, because nothing tells the teacher which children need which component; an evidence trail that identifies a child as reframing well when alone and not at all when classmates are watching indicates a peer-audience problem rather than a skills deficit, and those call for different responses. Second, the protocol is cheap enough to repeat. At 4 prompts a day over 12 school days, and 37.2 answered prompts per child, it costs each pupil a few minutes a week and needs no lesson time, so a school can run it before and after an intervention and read the change in the trail rather than in a score. Third, and least comfortably, nothing in our results licenses using the instrument to grade or rank children. The profile agrees with the two child self-report criteria more strongly than with the teacher rating (Table 4), which is what an instrument built on momentary self-report should do but also means it is measuring something a teacher does not already have — and a measure whose relation to observable classroom behaviour is this loose is not ready to be anyone’s summative judgement. Its defensible use is to direct a teacher’s attention, not to replace it.

Deploying anything of this kind with nine- to twelve-year-olds carries obligations the technical evaluation does not discharge, and we state them as design requirements rather than as afterthoughts. Participation requires informed parental consent and the child’s own assent, and assent means a child can decline a prompt or leave the study without explaining why — which the compliance model above already assumes, since a protocol that punished non-response would not have produced 77.5 per cent. The momentary items ask what happened and what the child did, not who else was involved by name, so the record does not accumulate evidence about third parties. Data minimisation argues for keeping the raw episode stream inside the school and releasing only the profile and its trail, and against any live stream of a child’s affect to

an adult’s screen, which would change the classroom rather than measure it. The output should reach a class teacher or school counsellor who knows the child, never a ranking of a cohort; the European framework now classes AI used to evaluate learners as high-risk, with obligations of human oversight and transparency that this design is intended to meet rather than to work around [19], [57]. None of this is settled by the fact that the model is auditable. Auditability makes oversight possible; it does not perform it.

B. The Cost of Interpretability

The constraints that make an assessment inspectable are not free, though on this corpus the bill is smaller than we expected. DP-APT concedes 0.006 AUC-ROC points to an unconstrained sequence encoder, which is well inside one seed-to-seed standard deviation and so not distinguishable from zero at three initialisations; the aggregated logistic baseline, which reads no sequence at all, comes within 0.016 of the proposed model. The structural tension is nevertheless real, and we expect it to bind on corpora where the sequence carries more than it does here. A regulation profile is defined partly by how often a strategy is deployed, and a frequency is a property of the whole protocol; an explanation, by contrast, is only usable if it points at a few moments. A model that must do both is asked to summarise a fortnight of school life and to indict five break times at the same time. Our unnormalised detector and prevalence-weighted pool recover part of the aggregate signal while remaining decomposable into per-episode contributions, which limits the gap, but the tension does not disappear.

This has a practical consequence for how such systems should be reported. A literature in which every interpretable architecture is claimed to match or beat its black-box comparison invites the suspicion that the comparison was chosen accommodatingly. We would argue instead that the accuracy cost should be measured and stated, and the decision left to the setting: in a screening application where no human reads the output, the unconstrained encoder is the better tool; in a school, where an adult must justify a judgement about a child to that child and to their parents, an evidence trail that is the arithmetic of the score and aligns with what the child actually did at 0.483 — above the 0.439 of integrated gradients on the same corpus, and far above the 0.273 of the interpretable comparison model — is worth 0.006 of AUC that we cannot in any case measure reliably. We note that the margin over integrated gradients is modest, and that the case for the proposed model rests as much on the trail being exact by construction as on its being better aligned.

C. What Faithfulness Does Not Guarantee

The second lesson concerns how explanation quality is judged, and we think it generalises beyond this architecture. The faithfulness criteria in common use ask a question about the model: if these episodes are removed, does the prediction change? A criterion of that form is satisfied by an explanation that points at the wrong moments, provided the model genuinely depends on them. Our results contain both halves of the resulting confusion. Across models the rank correlation

between the two criteria is 0.00: knowing that one explanation erases better than another tells a reader essentially nothing about which is more often right, and at the extremes the ordering inverts outright. And within one model, raising the weight on the entropy penalty — the obvious way to make an attention explanation more concentrated, and one that improves comprehensiveness monotonically — leaves alignment where it was while accuracy falls.

Two design recommendations follow. First, where an architecture admits an exact decomposition of its output, that decomposition should be what is reported, rather than a component of it chosen for legibility. An attention map is a convenient picture, not an account of the score, and our own attention variant illustrates how far apart the two can be. Second, a faithfulness score should not be presented as evidence that an explanation is correct. Where ground truth about the explained quantity exists, even on a subsample or under simulation, it should be used, because the two criteria can be pushed apart by ordinary hyperparameter tuning. This sharpens rather than settles the attention-as-explanation debate [26], [27]: the question is not whether attention is faithful, but whether faithfulness was ever the property that mattered.

The result also complicates the case against post-hoc attribution. Integrated gradients on an unconstrained encoder score poorly on erasure, as that literature reports, and nonetheless identify the episodes in which the justified family was deployed more often than the inherently interpretable concept-bottleneck comparison does. We would not conclude from one simulated corpus that post-hoc methods are underrated, but we would conclude that the evidence usually offered for the opposite view does not establish it, because it is evidence about faithfulness alone.

D. *From a Simulated Corpus to a Real Classroom*

The corpus used here is simulated, and this is a genuine limitation as well as a deliberate methodological choice. No children were sampled, and every number in this paper is a property of a generative model with a fixed seed, not of a cohort. It is a choice because justification alignment cannot be computed without episode-level ground truth about which strategy was actually deployed, and no observational corpus supplies that with certainty — least of all one collected from children, whose momentary reports of their own regulation are the least reliable of any age group; the measure that most sharply distinguishes the models is available only under simulation. It is a limitation because the generative process embodies psychological assumptions — an autoregressive affect process, a controllability-driven choice rule, a fit-dependent efficacy function, a developmental ordering of the five families — that are supported by the literature but simplify it. In particular, the ontology given to the model and the taxonomy used by the generator both derive from the process model, so the benefit attributed to the symbolic pathway is measured under conditions favourable to it. Nothing here should be read as evidence about how any real pupil regulates. Replication on observational momentary corpora collected in schools, with a subsample of episodes

coded by the child and their teacher shortly afterwards as partial ground truth, is the necessary next step, and the multi-site collaborative model that psychology has developed for replication [58] is the natural vehicle for assembling a corpus of that kind.

A further limitation is one the ceiling makes plain. The supervised probe recovers the deployed family from the same momentary features in 89.4 per cent of episodes, so the information the evidence trail would need is present in the data; our trail reaches 0.483 of an attainable 0.764. The shortfall is a property of the supervision rather than of the features. A child-level label constrains the model’s per-episode assertions only through their aggregate, and the anchoring term supplies the missing constraint from the observed items but does not replace an episode-level signal. Closing this gap is, we think, the most promising direction the measure opens: a small number of expert-coded or child-confirmed episodes per pupil may be enough to bind the detector to the moments it describes, and the measure reported here is what would show whether it had worked.

Deployment raises further questions that our evaluation does not settle, and the age of the respondents sharpens all of them. Momentary reports of regulation are themselves subject to bias, and a child who systematically under-reports keeping a feeling in — because concealment is the strategy least available to introspection, and because a child answering on a school device may reasonably suspect an adult will read it — will be assessed accordingly. The audit trail helps here in a way a trait score does not, since a teacher who reads the highlighted episodes can recognise that the evidence is thin, but it does not remove the problem. Nor does it address the asymmetry that matters most: the child assessed is the party least able to contest the assessment, and the institution assessing them is the one they attend every day [57], [18].

E. *Generalisability Beyond the Process Model*

DP-APT was built around one taxonomy, and its ontology pathway presupposes that the constructs to be assessed can be arranged in a graph of classes and families with interpretable descriptors. Many psychological frameworks satisfy this: the coping literature, the taxonomies of cognitive distortion used in cognitive therapy, and diagnostic criteria organised by symptom cluster all have the required shape. Others do not. Dimensional models of personality or of psychopathology are continuous rather than taxonomic, and forcing them into an ontology would impose structure the theory does not claim. Where the constructs themselves are contested, the ontology becomes a commitment that the model cannot revise, which is a strength when the theory is sound and a liability when it is not.

Three extensions follow naturally. The first is sparse episode-level supervision, which the ceiling analysis identifies as the binding constraint on how correct an evidence trail can become. The second is a learnable ontology, in which the graph is initialised from theory but permitted to adapt, with the divergence between the two reported as a finding about the theory. The third is prospective validation in a school: an

assessment that is auditable should also be actionable, and whether class teachers who receive the evidence trail identify the children who need support more accurately, or choose better responses, than teachers who receive a score is an empirical question that only a study with teachers can answer. We regard that study, rather than a further gain in AUC, as the precondition for using anything of this kind with children.

VIII. CONCLUSION

Momentary sampling can record how a primary-school pupil handles a quarrel at break or a piece of work that will not come right, but the models that summarise those records return conclusions without evidence, and a class teacher cannot act on a conclusion they cannot check. This paper proposed Dual-Pathway Affective Process Tracing, an inherently interpretable assessment engine that encodes the episode sequence as a temporal affect graph, encodes the process model of emotion regulation as a differentiable ontology, and couples the two so that every strategy score arrives with the moments that produced it.

On a developmentally calibrated corpus of 420 pupils in grades four to six, the model attained an AUC-ROC of 0.886 against 0.891 for an unconstrained sequence encoder, while producing explanations whose alignment with actual regulatory behaviour (0.483, against a chance level of 0.200 and an attainable 0.764) exceeded that of every comparison model. The inferred profiles converged with all three trait criteria the model never saw, most strongly with the children's own reports of reappraisal and suppression and least strongly with the teacher's rating — the ordering the construct predicts, since what a teacher can observe and what a child conceals are not the same quantity.

The work makes three contributions. It provides a fully differentiable assessment pipeline whose output includes an evidence trail that is the arithmetic of the score rather than a reconstruction of it. It introduces a validation of that trail against episode-level regulatory ground truth, turning auditability from a design claim into a measured quantity with a known floor and a known ceiling. And it uses that measure to show that explanation quality as currently assessed does not track explanation correctness: across models the two criteria are almost unrelated and invert at the extremes, and within one model tuning for faithfulness buys none of it.

Important challenges remain. The corpus is simulated and no child was sampled, the ontology is fixed and shares its taxonomy with the generative process, a substantial gap to the attainable alignment persists under child-level supervision alone, and the practical value of an evidence trail has yet to be demonstrated with teachers rather than metrics. Addressing these is the path from an auditable model to an instrument a school can defend to the child it describes and to that child's parents.

REFERENCES

- [1] E. Gullone, S. Hughes, N. J. King, and D. Tonge, "The normative development of emotion regulation strategy use in children and adolescents: A 2-year follow-up study," *J. Child Psychol. Psychiatry*, vol. 51, no. 5, pp. 567–574, May 2010, doi: 10.1111/j.1469-7610.2009.02183.x.
- [2] J. Zeman, M. Cassano, C. Perry-Parrish, and S. Stegall, "Emotion regulation in children and adolescents," *J. Dev. Behav. Pediatr.*, vol. 27, no. 2, pp. 155–168, Apr. 2006, doi: 10.1097/00004703-200604000-00014.
- [3] B. E. Compas et al., "Coping, emotion regulation, and psychopathology in childhood and adolescence: A meta-analysis and narrative review," *Psychol. Bull.*, vol. 143, no. 9, pp. 939–991, Sep. 2017, doi: 10.1037/bul0000110.
- [4] J. A. Durlak, R. P. Weissberg, A. B. Dymnicki, R. D. Taylor, and K. B. Schellinger, "The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions," *Child Dev.*, vol. 82, no. 1, pp. 405–432, Jan. 2011, doi: 10.1111/j.1467-8624.2010.01564.x.
- [5] R. D. Taylor, E. Oberle, J. A. Durlak, and R. P. Weissberg, "Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects," *Child Dev.*, vol. 88, no. 4, pp. 1156–1171, Jul. 2017, doi: 10.1111/cdev.12864.
- [6] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, Apr. 2008, doi: 10.1146/annurev.clinpsy.3.022806.091415.
- [7] T. J. Trull and U. W. Ebner-Priemer, "Ambulatory assessment," *Annu. Rev. Clin. Psychol.*, vol. 9, pp. 151–176, Mar. 2013, doi: 10.1146/annurev-clinpsy-050212-185510.
- [8] C. K. F. Wen, S. Schneider, A. A. Stone, and D. Spruijt-Metz, "Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis," *J. Med. Internet Res.*, vol. 19, no. 4, p. e132, Apr. 2017, doi: 10.2196/jmir.6641.
- [9] I. Myin-Germeys et al., "Experience sampling methodology in mental health research: New insights and technical developments," *World Psychiatry*, vol. 17, no. 2, pp. 123–132, Jun. 2018, doi: 10.1002/wps.20513.
- [10] P. Kuppens, N. B. Allen, and L. B. Sheeber, "Emotional inertia and psychological maladjustment," *Psychol. Sci.*, vol. 21, no. 7, pp. 984–991, Jul. 2010, doi: 10.1177/0956797610372634.
- [11] G. Sheppes, S. Scheibe, G. Suri, and J. J. Gross, "Emotion-regulation choice," *Psychol. Sci.*, vol. 22, no. 11, pp. 1391–1396, Nov. 2011, doi: 10.1177/0956797611418350.
- [12] S. J. Haines et al., "The wisdom to know the difference: Strategy-situation fit in emotion regulation in daily life is associated with well-being," *Psychol. Sci.*, vol. 27, no. 12, pp. 1651–1659, Dec. 2016, doi: 10.1177/0956797616669086.
- [13] N. C. Jacobson and B. Bhattacharya, "Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary

[1] E. Gullone, S. Hughes, N. J. King, and D. Tonge, "The normative development of emotion regulation strategy

- assessments,” *Behav. Res. Ther.*, vol. 149, Art. no. 104013, Feb. 2022, doi: 10.1016/j.brat.2021.104013.
- [14] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, “Machine learning in mental health: A scoping review of methods and applications,” *Psychol. Med.*, vol. 49, no. 9, pp. 1426–1448, Jul. 2019, doi: 10.1017/S0033291719000151.
 - [15] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
 - [16] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.
 - [17] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
 - [18] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: A multidisciplinary perspective,” *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, Art. no. 310, Nov. 2020, doi: 10.1186/s12911-020-01332-6.
 - [19] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence,” *Off. J. Eur. Union*, L 2024/1689, Jul. 2024.
 - [20] L. Zhao et al., “T-GCN: A temporal graph convolutional network for traffic prediction,” *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020, doi: 10.1109/TITS.2019.2935152.
 - [21] K. Cho et al., “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
 - [22] J. J. Gross, “The emerging field of emotion regulation: An integrative review,” *Rev. Gen. Psychol.*, vol. 2, no. 3, pp. 271–299, Sep. 1998, doi: 10.1037/1089-2680.2.3.271.
 - [23] J. J. Gross, “Emotion regulation: Current status and future prospects,” *Psychol. Inq.*, vol. 26, no. 1, pp. 1–26, Jan. 2015, doi: 10.1080/1047840X.2014.940781.
 - [24] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vancouver, BC, Canada, 2018.
 - [25] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.
 - [26] S. Jain and B. C. Wallace, “Attention is not explanation,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 3543–3556, doi: 10.18653/v1/N19-1357.
 - [27] S. Wiegrefe and Y. Pinter, “Attention is not not explanation,” in *Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 11–20, doi: 10.18653/v1/D19-1002.
 - [28] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, “ERASER: A benchmark to evaluate rationalized NLP models,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2020, pp. 4443–4458, doi: 10.18653/v1/2020.acl-main.408.
 - [29] T. L. Webb, E. Miles, and P. Sheeran, “Dealing with feeling: A meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation,” *Psychol. Bull.*, vol. 138, no. 4, pp. 775–808, Jul. 2012, doi: 10.1037/a0027600.
 - [30] A. Aldao, S. Nolen-Hoeksema, and S. Schweizer, “Emotion-regulation strategies across psychopathology: A meta-analytic review,” *Clin. Psychol. Rev.*, vol. 30, no. 2, pp. 217–237, Mar. 2010, doi: 10.1016/j.cpr.2009.11.004.
 - [31] G. A. Bonanno and C. L. Burton, “Regulatory flexibility: An individual differences perspective on coping and emotion regulation,” *Perspect. Psychol. Sci.*, vol. 8, no. 6, pp. 591–612, Nov. 2013, doi: 10.1177/1745691613504116.
 - [32] A. Aldao, “The future of emotion regulation research: Capturing context,” *Perspect. Psychol. Sci.*, vol. 8, no. 2, pp. 155–172, Mar. 2013, doi: 10.1177/1745691612459518.
 - [33] E. Gullone and J. Taffe, “The Emotion Regulation Questionnaire for Children and Adolescents (ERQ-CA): A psychometric evaluation,” *Psychol. Assess.*, vol. 24, no. 2, pp. 409–417, Jun. 2012, doi: 10.1037/a0025777.
 - [34] A. Shields and D. Cicchetti, “Emotion regulation among school-age children: The development and validation of a new criterion Q-sort scale,” *Dev. Psychol.*, vol. 33, no. 6, pp. 906–916, Nov. 1997, doi: 10.1037/0012-1649.33.6.906.
 - [35] C. Wrzus and M. R. Mehl, “Lab and/or field? Measuring personality processes and their social consequences,” *Eur. J. Pers.*, vol. 29, no. 2, pp. 250–271, Mar. 2015, doi: 10.1002/per.1986.
 - [36] P. Kuppens and P. Verduyn, “Emotion dynamics,” *Curr. Opin. Psychol.*, vol. 17, pp. 22–26, Oct. 2017, doi: 10.1016/j.copsy.2017.06.004.
 - [37] E. L. Hamaker and M. Wichers, “No time like the present: Discovering the hidden dynamics in intensive longitudinal data,” *Curr. Dir. Psychol. Sci.*, vol. 26, no. 1, pp. 10–15, Feb. 2017, doi: 10.1177/0963721416666518.
 - [38] L. F. Bringmann, E. Ferrer, E. L. Hamaker, D. Borsboom, and F. Tuerlinckx, “Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model,” *Multivariate Behav. Res.*, vol. 53, no. 3, pp. 293–314, May 2018, doi: 10.1080/00273171.2018.1439722.
 - [39] A. Brose, P. Voelkle, M. Lövdén, U. Lindenberger, and F. Schmiedek, “Differences in the between-person and within-person structures of affect are a matter of degree,” *Eur. J. Pers.*, vol. 29, no. 1, pp. 55–71, Jan. 2015, doi: 10.1002/per.1961.
 - [40] T. R. Insel, “Digital phenotyping: Technology for a new science of behavior,” *JAMA*, vol. 318, no. 13, pp. 1215–1216, Oct. 2017, doi: 10.1001/jama.2017.11295.
 - [41] D. C. Mohr, M. Zhang, and S. M. Schueller, “Personal sensing: Understanding mental health using ubiquitous

- sensors and machine learning,” *Annu. Rev. Clin. Psychol.*, vol. 13, pp. 23–47, May 2017, doi: 10.1146/annurev-clinpsy-032816-044949.
- [42] D. B. Dwyer, P. Falkai, and N. Koutsouleris, “Machine learning approaches for clinical psychology and psychiatry,” *Annu. Rev. Clin. Psychol.*, vol. 14, pp. 91–118, May 2018, doi: 10.1146/annurev-clinpsy-032816-045037.
- [43] J. Torous, S. Bucci, I. H. Bell, L. V. Kessing, M. Faurholt-Jepsen, P. Whelan, A. F. Carvalho, M. Keshavan, J. Linardon, and J. Firth, “The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality,” *World Psychiatry*, vol. 20, no. 3, pp. 318–335, Oct. 2021, doi: 10.1002/wps.20883.
- [44] T. Yarkoni and J. Westfall, “Choosing prediction over explanation in psychology: Lessons from machine learning,” *Perspect. Psychol. Sci.*, vol. 12, no. 6, pp. 1100–1122, Nov. 2017, doi: 10.1177/1745691617693393.
- [45] P. F. Hitchcock, E. Fried, and M. J. Frank, “Computational psychiatry needs time and context,” *Annu. Rev. Psychol.*, vol. 73, pp. 243–270, Jan. 2022, doi: 10.1146/annurev-psych-021621-124910.
- [46] E. I. Fried and A. O. J. Cramer, “Moving forward: Challenges and directions for psychopathological network theory and methodology,” *Perspect. Psychol. Sci.*, vol. 12, no. 6, pp. 999–1020, Nov. 2017, doi: 10.1177/1745691617705892.
- [47] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, 2017, pp. 3319–3328.
- [48] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017, *arXiv:1702.08608*, doi: 10.48550/arXiv.1702.08608.
- [49] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept bottleneck models,” in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5338–5348.
- [50] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. Su, “This looks like that: Deep learning for interpretable image recognition,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2019, pp. 8930–8941.
- [51] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, “Neuro-symbolic artificial intelligence: Current trends,” *AI Commun.*, vol. 34, no. 3, pp. 197–209, 2021, doi: 10.3233/AIC-210084.
- [52] A. d’Avila Garcez and L. C. Lamb, “Neurosymbolic AI: The 3rd wave,” *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12387–12406, Nov. 2023, doi: 10.1007/s10462-023-10448-w.
- [53] J. Hastings, W. Ceusters, B. Smith, and K. Mulligan, “The emotion ontology: Enabling interdisciplinary research in the affective sciences,” in *Modeling and Using Context (Lecture Notes in Computer Science)*, vol. 6967, Berlin, Germany: Springer, 2011, pp. 119–123, doi: 10.1007/978-3-642-24279-3_14.
- [54] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, “OWL 2: The next step for OWL,” *J. Web Semant.*, vol. 6, no. 4, pp. 309–322, Nov. 2008, doi: 10.1016/j.websem.2008.05.001.
- [55] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, Toulon, France, 2017.
- [56] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, 2015.
- [57] A. Fiske, P. Henningsen, and A. Buyx, “Your robot therapist will see you now: Ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy,” *J. Med. Internet Res.*, vol. 21, no. 5, Art. no. e13216, May 2019, doi: 10.2196/13216.
- [58] H. Moshontz et al., “The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network,” *Adv. Methods Pract. Psychol. Sci.*, vol. 1, no. 4, pp. 501–515, Dec. 2018, doi: 10.1177/2515245918797607.
- [59]