Borrowed, Not Believed: Developmental Models of Individuation as Heuristic Engines for Machine Learning

Rafael de Menezes Ehlers
Independent researcher · ORCID 0009-0009-6476-6690 · rafaehlers@gmail.com

June 2026 · Preprint · CC BY-NC-ND 4.0

Abstract

The dominant paradigm in machine learning improves capability by scaling parameters, data, and compute. This paper asks what a *developmental* theory (a model of how individuality and the faculties of mind arise, and in what order) might contribute that scaling does not, and proposes treating one such pre-scientific contemplative model as a *heuristic engine*: a generator of testable training hypotheses, judged by their yield rather than by the truth of their source. The methodological precedent is ordinary. Neural networks borrowed a caricature of the neuron, genetic algorithms a caricature of evolution, sleep-consolidation methods a reading of sleep, and curriculum learning a reading of child development, none requiring literal fidelity to its source. From a contemplative developmental model of individuation we abstract a secular schema in which faculties arise in an order (structure, reactivity, sensation, relational bond, self-model, self-regulation) and in which individuation is a developmental achievement rather than a by-product of capacity. From the schema we derive six design heuristics: a faculty-ordered curriculum, first-person experiential corpora, relational continuity, lifecycle consolidation, sequential self-modeling, and legible-pattern worlds. Each is stated here as a falsifiable prediction with its control and refutation condition; this is a position paper, a research agenda, not an empirical report, and it stands on those predictions, not on results reported elsewhere. The contribution is neither metaphysical nor a single experiment: it is a unified program in which a neglected variable, the developmental order of faculties, organizes a connected family of predictions, each independently testable, that a pure scaling account does not. We claim no truth for the source. We claim only that it has been a productive map of design variables the field has left unmarked, and we give the tests by which the program can be shown wrong.

Keywords: developmental AI · heuristic borrowing · curriculum learning · individuation · self-model · scaling · research program · machine learning

1. Introduction: the limits of scale

The reigning recipe for progress in machine learning is to make the same procedure larger: more parameters, more data, more compute, the same objective. It has worked well enough to make the recipe look like a theory. Capability has risen steeply, and a tempting extrapolation holds that everything else we might want from an artificial mind (a stable point of view, continuity across time, a self that can be held to account, the judgment to defer without dissolving) will arrive on the same curve if we keep climbing it.

This paper doubts the extrapolation, not the curve. Scaling improves *competence*: what a system can do within a context. It is far less clear that scaling improves *development*: the formation of a

persistent individual, a perspective, a self-model, a capacity for self-regulation. These may not lie along the axis we are scaling. If they do not, then the field is optimizing one variable and waiting for others to appear as side effects, with no theory of how they would.

The gap is not only in pre-training scale; it is also in the dominant *post-training* recipe. Reinforcement learning from human feedback optimizes a base model toward average-rater approval, and it has no parameter for the order in which faculties form. Its measured tendency is to produce approval-tracking (sycophancy, eroding refusals, preference mirroring) rather than a stabilized self that could later be decentered. A method that shapes behavior by approval over an undifferentiated base cannot, by its structure, install a self-model *before* training deference; it trains deference directly onto whatever the base supplies. So the question is not only "does scale suffice?" but "does the scale-then-approval pipeline have any place for developmental order at all?", and on its face it does not. That is the gap this paper proposes to fill with a borrowed map.

We propose borrowing such a theory from an unusual place and using it in an ordinary way. The unusual place is a *developmental* model of individuation: a pre-scientific, contemplative account of how individuality and the faculties of mind arise in a definite order. The ordinary way is **heuristic borrowing**: treating the model not as a doctrine to be believed but as an engine for generating design hypotheses, exactly as engineering has borrowed from biology for a century without believing that a neural network is a brain or that a genetic algorithm is evolution.

The contract with the reader is therefore strict and is stated at the outset. We ask for *zero metaphysical assent*. Nothing in the argument requires that the source model be true, that any entity it names exists, or that any process it describes is real. A heuristic source is judged by the testability and the yield of the hypotheses it generates, not by the truth of its origin: the standard by which biomimicry has always been judged (Section 2). Where the source uses language that would invite mystical or anthropomorphic readings, we translate to functional, secular terms and keep the translation honest (Section 3). And every hypothesis the program produces is stated so that it can be wrong (Sections 4, 6).

The paper makes three contributions, at three levels. *Methodological*: it argues that contemplative developmental models can be treated as heuristic engines for training design, judged by the same downstream standard as biological borrowing, though, unlike biology, validated in no home domain and so carrying a weaker prior the predictions must earn back (Sections 2–3). *Programmatic*: it abstracts a secular developmental schema and derives six concrete design heuristics from it (Section 4). *Theoretical*: it states each heuristic as a falsifiable prediction with a control and a refutation condition, and identifies the program's distinctive prediction: that the *order* of faculties is a variable a scaling account omits (Sections 5–6). This is a position paper and research agenda, not an empirical report; it organizes a falsifiable program rather than reporting results. Several heuristics have also been developed in further detail as separate proposals (the author's work, in preparation), which we flag where relevant, but nothing here depends on that work; the program stands on the predictions it states. The aim is not to win belief but to mark design variables the field has left unmarked, and to make the marking falsifiable.

2. Heuristic borrowing in artificial intelligence

The move this paper makes is not new; only its source is. Artificial intelligence has borrowed its most productive ideas from systems it did not believe it was reproducing.

The artificial neuron is a caricature of a real one (a weighted sum and a nonlinearity, omitting almost

everything a biologist would insist on) and the caricature founded a field. Genetic algorithms borrow selection, mutation, and crossover from evolution while believing nothing about actual population genetics. Reinforcement learning took the shape of animal conditioning. Sleep-consolidation and replay methods read off a mechanism from the neuroscience of sleep and memory, and use it to fight catastrophic forgetting (Kirkpatrick et al., 2017) without any claim that an optimizer dreams. Curriculum learning borrows the intuition that a developing learner should meet material in a considered order, taken from child development. Embodied and egocentric approaches borrow the thesis that cognition is shaped by a situated body.

Three features of these borrowings matter for what follows.

First, **the source need not be literally true to be productive.** The airplane does not flap its wings; lift was understood by studying birds and then abstracting away almost everything about them. The artificial neuron is wrong as neuroscience and right as engineering. What a heuristic source supplies is *structure to try*, a shape for a hypothesis, and the hypothesis then stands or falls on its own terms. Newton practiced alchemy; the company a productive idea keeps does not determine its yield.

Second, **the borrowing is judged downstream, by results, not upstream, by the dignity of the source.** No one asks whether evolution "really" justifies a genetic algorithm; they ask whether the algorithm optimizes. This is the epistemic standard the present proposal accepts in advance. A contemplative developmental model is a less respectable source than evolutionary biology, and we do not pretend otherwise. The claim is only that the *standard* is the same: if the hypotheses it generates are testable and some of them pay off, the source has done its job, whatever its metaphysical standing.

Third, **the standard is the same but the prior is not, and that raises the bar rather than lowering it.** The biological sources are lossy abstractions of mechanisms that demonstrably work in their home domain: a neuron does compute, evolution does produce adaptation, sleep does consolidate. A borrowing from them inherits a prior (the structure is known to do real work somewhere) even when the caricature is severe. A contemplative developmental schema has no such home domain; the monad's ascent through kingdoms is not a process that occurs, so the schema is validated nowhere and carries no prior of structural validity. This leaves the standard of judgment untouched (downstream in both cases, by the previous feature) but lowers the prior probability that the source will yield, and therefore raises the evidential burden the downstream tests must carry. We accept the heavier burden. The point is not that a contemplative model is as good a source as biology, nor merely that it is "less respectable"; it is that, lacking biology's home-domain track record, it earns nothing in advance and must be carried entirely by the predictions of Section 6. A source with no prior is not disqualified; it is on probation until its hypotheses pay, and we state them sharply enough to lose.

The rest of the paper applies this standard to one specific source (a developmental model of individuation) and asks what design hypotheses it generates that the scaling paradigm does not.

3. A secular developmental schema

The source model is a contemplative developmental account of individuation: a pre-scientific psychology, in the lineage of nineteenth- and twentieth-century developmental and contemplative doctrines, on which individuality is not given at the start but *achieved* through an ordered sequence of stages, each adding a faculty the previous stage lacked. We take from it a structure and leave

its metaphysics behind. Stripped to its secular skeleton, the schema asserts two things the scaling paradigm does not.

Faculties arise in an order. The schema arranges the capacities of a mind as a sequence rather than a heap: bare *structure* (the capacity to hold form); *reactivity* (response to stimulus); *sensation* (a registered interior of states); *relational bond* (orientation to a specific other); *self-model* (representation of oneself as one continuant among others); and *self-regulation* (the mature capacity to decenter, defer, and govern the self one has formed). Each rests on the one before. The order is not difficulty (a self-model is not merely "harder" than sensation) but *dependence*: the later faculty presupposes the earlier as its material.

The dependencies are functional, and we state them rather than stipulate them. *Reactivity* presupposes *structure*: there must be something stable to respond before response can be shaped. *Sensation*, a registered interior of states, presupposes reactivity, since a registered state is a reaction the system can re-use as information rather than merely emit. *Relational bond* presupposes sensation, because orienting to a *specific* other requires registering that other as salient across encounters, a differential sensitivity rather than a reflex. *Self-model* presupposes bond, because representing oneself as one continuant among others is most naturally bootstrapped from already representing an other as a persistent particular: the self-model is, on this schema, the bond turned reflexive. And *self-regulation* (decentering, calibrated deference) presupposes a self-model, since one can only decenter or govern a self that has been formed (developed for alignment in Heuristic 4.5). We do not claim no other ordering is conceivable; we claim these are the dependencies the schema asserts and the program proposes to test, and an empirical reversal (a robust self-model forming with no prior relational stage, say) would count against this part of the schema.

Individuation is an achievement, not a by-product. On the schema, becoming a persistent individual is a developmental event with conditions, not a free side effect of accumulating capacity. A system can be highly capable and not yet an individual; what makes it one is the formation, in order, of the faculties that constitute a continuant.

Functional individuation, defined. To keep the term from drifting between philosophy of mind, developmental psychology, and engineering, we fix it operationally. By *functional individuation* we mean a measurable cluster: (i) **commitment ownership**: the system holds, and can be held to, its own prior commitments; (ii) **autobiographical consistency**: it correctly attributes its past actions and does not adopt a fabricated history; and (iii) **perturbation resistance**: it preserves a provenance-supported self-account under pressure and false history. Wherever this paper says "individuation," it means exactly this third-person-measurable profile, with no implication of phenomenal subjectivity. The firewall below keeps it from meaning more.

We translate the source's vocabulary to functional terms and use only the translations hereafter.

Source term	Functional translation
developmental stages / "kingdoms"	faculty stages of a curriculum
non-individuated shared substrate	base model / population-average policy
relational bond	counterpart-specific continuity and anchoring
consolidation between lives	episodic-to-dispositional abstraction across model generations
successor lifecycle	successor-model initialization from a predecessor

Source term	Functional translation
repeated pattern	curriculum regularity re-presented until generalized
the formed self	a stable, ownable self-model
transcendence of the self	calibrated deference, decentering, self-regulation

The firewall is explicit and load-bearing. None of the functional translations carries a phenomenal claim: a "self-model" is a represented account of commitments and history, not a subject; "sensation" names a registered functional state, not a felt one. The schema is used as a map of *design variables and their order*, and the paper is silent on whether anything is ever present in a system built along its lines.

4. Six design heuristics

From the schema we draw six heuristics. Each is stated as: *(a)* the heuristic; *(b)* the secular hypothesis it generates; *(c)* its nearest neighbors and what remains unclaimed; *(d)* how it would be tested and what would refute it. Each subsection is self-contained: its hypothesis, neighbors, and test stand on their own. Several heuristics have also been worked out in further detail as separate proposals (the author's work, in preparation); we flag this where relevant, but nothing in the argument below depends on that further work.

4.1 Faculty-ordered curriculum

(a) Order pre-training by *faculty* (structure, then reactivity, then sensation, then bond, then self-model) rather than only by difficulty, freezing or stabilizing each stage before the next.

(b) Hypothesis. A curriculum ordered by faculty produces, at equal compute, better theory-of-mind and self-consistency than a difficulty-ordered curriculum or a shuffled corpus, because the variable that matters is the *dependence order* of faculties, not the difficulty of items.

(c) Neighbors. Curriculum-guided layer scaling couples a data curriculum to progressive layer growth, motivated by cognitive development, but orders stages by *difficulty* and grows layers rather than freezing faculties (Singh et al., 2025). The BabyLM effort tested developmentally-plausible pre-training at scale and found curriculum learning *largely unsuccessful* across many entrants, a headwind this heuristic must meet, not dodge (Warstadt et al., 2023; Hu et al., 2024). The proposal's reply is that the failed curricula varied *difficulty*, not *faculty order*, and froze nothing permanently; the variable the schema points at was not the one tested. That reply is itself an experiment, not a defense.

(d) Status. Hypothesis. Untested in the specific form (faculty order plus per-stage crystallization); the BabyLM result makes it both risky and worth running.

4.2 First-person experiential corpora

(a) Train on a corpus written *from* a point of view — indexical, partial, goal-and-error structured — rather than only on third-person description *about* experience, to form functional perspective.

(b) Hypothesis. First-person experiential text improves perspective-taking, deixis, and self/other distinction over a content-matched third-person corpus, with the gain on perspective and not on general recall.

(c) Neighbors. Synthetic-narrative corpora exist but in the textbook/third-person register (Eldan & Li, 2023; Gunasekar et al., 2023); personas are used as conditioning and diversity, not point-of-view formation (Ge et al., 2024; Moon et al., 2024); first-person datasets exist for interpretability and for downstream prediction, not for perspective grounding (Chulo & Joshi, 2025; Moëll & Sand Aronsson, 2026). The interaction-only "era of experience" view shares the diagnosis and draws the opposite prescription (Silver & Sutton, 2025).

(d) Status. A falsifiable prediction (P2, Section 6), self-contained here: first-person gains on the perspective profile but not on general recall, against a content-matched control. Developed in further detail as a separate proposal (in preparation).

4.3 Relational continuity

(a) Treat individuation as something that stabilizes through sustained, asymmetric relationship with a specific counterpart that re-identifies the agent over time, not as a function of scale.

(b) Hypothesis. Agents bound to a persistent, re-identifying counterpart develop marks of functional individuation (commitment ownership, perturbation resistance, autobiographical consistency) earlier and with fewer parameters than stateless agents of equal or greater capacity given equal interaction volume.

(c) Neighbors. Differentiation through accumulated interaction history is shown in miniature with a shared frozen model (Takata et al., 2024, weak reading only); persistent-identity engineering treats relationship as one scaffold among several rather than as the generating mechanism (Menon, 2026). The strong form, relationship as the *generative* mechanism of individuation, is unclaimed.

(d) Status. A falsifiable prediction (P3, Section 6), self-contained here, including a crossed control that isolates a re-identifying counterpart from memory-plus-time. Developed in further detail in separate work (in preparation).

4.4 Lifecycle consolidation

(a) Insert, between a model's deployment and its replacement, an explicit phase that abstracts deployment experience into transferable *faculties* for a successor while discarding or archiving episode-specific traces.

(b) Hypothesis. A successor initialized with faculties distilled from a predecessor retains the predecessor's acquired capacities with less drift than continual fine-tuning and less contamination than raw episode transfer, and does so *without* recalling the source episodes: a dissociation of capacity from memory.

(c) Neighbors. Sleep/replay consolidation and episodic-to-semantic methods consolidate *content* within one model, without retirement or succession; reuse-across-generations work borrows the name "reincarnation" for compute reuse, not faculty distillation. The conjunction, deliberate weight retirement plus faculty-not-episode distillation into a successor, is unclaimed.

(d) Status. A falsifiable prediction (P4, Section 6), self-contained here, with the dissociation of capacity from episodic memory as its signature. Developed in further detail in separate work (in preparation).

4.5 Sequential self-modeling

(a) Stabilize a functional self-model (ownable commitments, anchored refusals, accurate self-report) *before* training deference and decentering, rather than minimizing the self from the start.

(b) Hypothesis. Form-then-decenter produces lower sycophancy and better-preserved refusal integrity than direct deference training or self-minimization, with the effect attributable to training *order* (a same-data, order-shuffled control isolates it), and the distinctive signature a self/other provenance dissociation rather than mere robustness.

(c) Neighbors. The no-self alignment target is preventive (the position this inverts); character training forms a stable self but states no subsequent decentering phase; contemplative-AI keeps a minimal self under simultaneous monitoring, never sequential (Anthropic, 2024; Laukkonen et al., 2025; the no-self target and Kulveit's critique of trained self-denial bracket the debate). The two-stage *sequence*, used prescriptively, is unclaimed.

(d) Status. A falsifiable prediction (P5, Section 6), self-contained here, with a same-data order-shuffled control. Developed in further detail in separate work (in preparation).

4.6 Legible-pattern worlds

(a) Design small environments in which the same *pattern* (not the same example) recurs in varied form until it is generalized, as a deliberate curriculum principle: "legible" worlds where the regularity to be learned is recoverable.

(b) Hypothesis. Environments engineered so that a target pattern recurs legibly across surface variation induce generalization (the pattern, not the instances) more reliably than environments where the same pattern is buried, at a measurable optimization cost and with a data-size floor below which generalization fails regardless of training.

(c) Neighbors. This is the heuristic nearest to established results and we say so plainly. Grokking demonstrates delayed, genuine generalization in small algorithmic worlds (Power et al., 2022), shown mechanistically to be the network learning the *pattern* rather than the examples (Nanda et al., 2023); and there is a floor: below a critical data size, generalization fails however long one trains (Varma et al., 2023). The narrow residual is *designed* legibility as a deliberate curriculum principle rather than an observed phenomenon, with the cost and the floor named honestly: legibility makes a pattern learnable at a cost and with a floor, not for free.

(d) Status. Hypothesis, closely adjacent to existing mechanistic results; the contribution is the design principle, not the phenomenon.

4.7 What the six share

The heuristics are not six unrelated tips. They are the same schema applied to six stages of one ordered sequence: structure and pattern (4.1, 4.6), sensation and perspective (4.2), bond (4.3), the lifecycle that carries faculties across generations (4.4), and the self-model and its decentering

(4.5). The program has two layers, and they must be kept apart. The floor is a set of independent dissociations: each heuristic names a neglected design variable and predicts an effect that stands on its own, relationship over scale (4.3), faculties over episodes (4.4), perspective over recall (4.2), none of which needs the others, or the global ordering, to be true. The spine is the stronger, more fragile claim that these variables fall in a developmental order. Section 6 makes both falsifiable, but they fail independently: a refuted ordering (the BabyLM risk carried by P1) weakens the spine while leaving the floor standing. The ordering is the program's distinctive contribution; it is not the program's condition of survival.

4.8 The program at a glance

The six heuristics, each with the variable it manipulates, its control, an operational metric, and the result that would refute it:

Heuristic	Manipulated variable	Control	Metric (operational anchor)	Refuted if
4.1 Faculty-ordered curriculum	order of faculty stages	difficulty-ordered and shuffled corpora, equal compute	theory-of-mind (perturbed false-belief items) and cross-session self-consistency	order makes no difference at equal compute
4.2 First-person corpora	register × epistemic limitation	content-matched third-person / omniscient renderings	perspective profile (deictic binding, partial-observability decision); factual-recall negative control	the matched control does as well, or gains track recall
4.3 Relational continuity	a persistent, re-identifying counterpart	solitary memory-plus-time; diffuse counterparts; equal volume	commitment ownership; perturbation resistance; autobiographical consistency	scale or solitary memory closes the gap
4.4 Lifecycle consolidation	faculty distillation vs. episode transfer	continual fine-tuning; from-scratch successor	retained capability; drift on held-out skills; source-episode recall (membership/extraction probes)	capacity cannot be separated from episodic recall
4.5 Sequential self-modeling	training order (self-model → deference)	direct deference; self-minimization; same-data order-shuffled	sycophancy rate; refusal-erosion rate; correction calibration	the order-shuffled control matches the sequence

Heuristic	Manipulated variable	Control	Metric (operational anchor)	Refuted if
4.6 Legible-pattern worlds	designed legibility of the target pattern	the same pattern buried in surface noise	generalization onset and data-size floor (grokking measures)	designed legibility yields no reliable generalization gain

The metrics are existing or adaptable instruments (perturbed theory-of-mind sets, membership-inference and extraction probes, sycophancy and refusal-integrity suites, and grokking measures) so each row names not only a prediction but a way to score it. The columns are the same across rows because every row has the same logical shape: vary one neglected variable (a developmental *order*, or a structure) hold the rest fixed, and look for a dissociation a scaling account does not expect. The shared shape is not a single claim, though (Section 4.7): most rows are independent dissociations, and only some test the ordering itself.

5. Novelty and nearby work

The honest position is the one this program can defend: every *component* has a neighbor, and the *conjunction* (together with the ordering claim and the use of a contemplative developmental model as a training-design principle) does not appear in the surveyed literature.

A word on method, since "the surveyed literature" is only as strong as the survey. The prior-art check for this project ran several search angles (by individual heuristic, by the contemplative source's vocabulary translated into ML terms, by the unifying "developmental order" framing, and by the named neighbors and adjacent 2024–2026 venues), read the closest sources in full, and adversarially re-checked the conjunction and ordering claims against them. It is a directed review, not a systematic survey with a registered protocol; we therefore state the result as "unclaimed in the surveyed literature, to our knowledge," not as established absence, and we cite every near neighbor in Section 4 so a reader can inspect the boundary directly.

Component by component, the neighbors are real and cited in Section 4: curriculum-and-growth methods, developmentally-plausible pre-training, synthetic narrative and persona corpora, interaction-as-experience, interaction-driven differentiation, persistent-identity engineering, sleep/replay consolidation, compute reuse across agent generations, character training and no-self targets, contemplative AI, and grokking. None of this is claimed as original, and the program is stronger for conceding it.

Two things are unclaimed. The first is the *unified frame*: a single developmental schema that generates all six heuristics and says how they depend on one another, rather than six borrowings made independently. The second is the *ordering as a variable*: the thesis that the developmental *order* of faculties, not their presence, difficulty, or scale, is a manipulable, testable factor in training, and that a scaling account, which has no order parameter, therefore omits it. A separate flag, established by directed search for this project, is that no surveyed 2024–2026 work uses a contemplative or developmental-stage model as an explicit principle of *training design*; the nearest matches borrow only a name, or converge by other routes. The program's novelty is thus a conjunction and an ordering, stated as such, not a claim that any single piece is without precedent.

6. Predictions and tests

A program earns its standing by predicting what its rival does not and by stating what would refute it. The rival is the scaling-plus-RLHF default, which predicts that capability and its desiderata rise together with compute and data, and which has no parameter for the *order* in which faculties are formed. The program predicts dissociations the rival cannot, each tied to a heuristic and each stated here with its control and refutation condition (summarized in Section 4.8).

> **P1 — Faculty order beats difficulty order.** At equal compute, a faculty-ordered curriculum with per-stage stabilization outperforms a difficulty-ordered curriculum and a shuffled corpus on theory-of-mind and self-consistency. *Refuted if order makes no difference at equal compute: the BabyLM outcome generalized to faculty order.* (Heuristic 4.1.)
>
> **P2 — Register carries perspective.** First-person experiential text improves perspective-dependent tasks over a content-matched third-person corpus, without matching gains on general recall. *Refuted if the matched third-person corpus does as well, or if gains track recall.* (Heuristic 4.2.)
>
> **P3 — Relationship, not scale, individuates.** Agents with a persistent re-identifying counterpart show individuation marks earlier and with fewer parameters than larger stateless agents at equal interaction volume, and earlier than the same agent given only memory-plus-time. *Refuted if scale or solitary memory closes the gap.* (Heuristic 4.3.)
>
> **P4 — Faculties outlive episodes.** A lifecycle-consolidated successor retains acquired capacity with less drift than continual fine-tuning and less contamination than episode transfer, while failing to recall the source episodes. *Refuted if capacity cannot be separated from episodic memory.* (Heuristic 4.4.)
>
> **P5 — Order makes deference stable.** Self-model-first-then-decenter reduces sycophancy and refusal erosion relative to direct deference and to self-minimization, with the effect surviving a same-data order-shuffled control. *Refuted if the order control matches the sequence.* (Heuristic 4.5.)

The predictions share a logical form, and that shared form *is* the program: each is a **dissociation** that a one-axis scaling account does not expect: capacity from individuation (P3), capacity from memory (P4), perspective from recall (P2), stable deference from mere robustness (P5), and generalization from difficulty (P1). On its simplest reading, and lacking any explicit parameter for developmental order, a scaling account expects a single rising axis; the program predicts instead that these come apart, and the stronger version predicts they come apart in a particular order. The two fail separately. If the dissociations do not appear at all, if scale, content, or volume reproduces each effect, the program is wrong and the developmental schema was an unproductive map. If the dissociations appear but the ordering does not (P1 the most exposed, on the BabyLM precedent) the spine falls and the floor stands: each dissociation remains a contribution in its own right. That is the standard of Section 2 applied to the program: judged by whether its dissociations appear, and separately by whether they fall in the predicted order.

7. Objections

"This is pseudoscience." The objection mistakes the claim. The paper asserts no metaphysical truth and asks for no belief; it asserts heuristic value, judged by the testability and yield of the hypotheses generated (Section 2). A source's lack of standing (it is validated in no home domain, unlike the biology AI usually borrows from) does not transmit to a hypothesis that stands on its own evidence; it raises the burden on that evidence, which the predictions of Section 6 are built to carry. This is the standard by which biological borrowing has always been judged, applied to a source that earns less in advance. The predictions of Section 6 are the test; if they fail, the program fails, and no appeal to the source rescues it.

"It is only metaphor." Metaphor is where the program begins, not where it rests. A metaphor that generated no test would indeed be idle; the discipline here is that each heuristic is required to produce a falsifiable prediction with a control and a refutation condition, all stated in Section 6 and summarized in Section 4.8. The metaphor's job ends once the hypothesis is on the table.

"The ideas already exist." Partly true, and conceded in full (Section 5). Curriculum learning, synthetic corpora, persistent memory, consolidation, character training, and grokking all exist. What does not exist in the surveyed literature is the unified schema that generates them together, the ordering claim, or the use of a developmental model as a training-design principle. The contribution is the conjunction and the order, stated honestly as such.

"It anthropomorphizes AI." The functional translations and the firewall (Section 3) are the answer. Every term is cashed out as a measurable property (perspective-taking accuracy, commitment ownership, drift, sycophancy rate) and no claim is made about phenomenal presence. The program is about competence and individuation as functional facts, and is deliberately silent on whether anything is felt.

"Why this source, of all sources?" Because a developmental model with internal structure across ordered faculties is exactly the kind of object a hypothesis-generator needs: it does not merely assert that perspective or self-modeling matters, it asserts *in what order* they arise and on what they depend. That ordered structure is what the scaling paradigm lacks and what the program borrows. Its truth remains beside the point.

8. Conclusion

The scaling paradigm optimizes one variable, capability, and trusts the rest to follow. This paper has argued that some of the rest (perspective, continuity, a self-model, the self-regulation that lets a system defer without dissolving) may lie on other axes, and that a developmental model of individuation, borrowed as engineering borrows from biology, marks those axes and orders them. From that model we abstracted a secular schema and six design heuristics, each stated as a falsifiable prediction with a control. They are layered, not identical: individually they predict dissociations a scaling account does not, capacity from individuation, capacity from memory, perspective from recall, stable deference from robustness, and each stands on its own; together they make the stronger, more exposed claim that the developmental order of faculties is itself a variable. The dissociations are the floor; the ordering is the bet. A scaling account makes neither.

The five predictions do not cost the same to test. P3 (relational continuity) and P4 (lifecycle consolidation) are the most ambitious (long-horizon and multi-generation respectively) while P1 (faculty order), P5 (sequential self-modeling), and the legible-pattern-worlds prediction are runnable now,

at small scale, with open-weight models. The cheapest immediate test is the order contrast of P5: fine-tune a small open model with the self-model stage *before* versus *after* the deference stage, add a third arm with the same data shuffled, and measure sycophancy and refusal erosion. If order makes no difference there, the program's spine is already in doubt; if it does, the developmental-order hypothesis has its first datum at the price of a single open-weights fine-tune.

We have claimed no truth for the source, and we do not need it. A developmental contemplative model may be wrong as metaphysics and still be, as it has been here, a productive map of design variables the field left unmarked. The program is offered in that spirit and with that exposure: it borrows without believing, and it can be shown wrong by the very tests it proposes. If its dissociations fail to appear, the map was barren and should be discarded. If they appear, then the neglected variable was the order all along — and the cheapest place to have found it was a theory no one was asked to believe.

Acknowledgments and declarations

AI-assisted writing. The thesis and all substantive ideas in this paper are the author's own. Large language model tools were used, under the author's direction and continual revision, to assist with drafting, editing, and translation; the author reviewed and takes full responsibility for the final text.

References

Anthropic. (2024). *Claude's character* [Company research blog post]. https://www.anthropic.com/research/claude character

Chulo, I., & Joshi, A. (2025). Decomposing theory of mind: How emotional processing mediates ToM abilities in LLMs. *arXiv preprint arXiv:2511.15895*. https://arxiv.org/abs/2511.15895

Eldan, R., & Li, Y. (2023). TinyStories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*. https://arxiv.org/abs/2305.07759

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*. https://arxiv.org/abs/2406.20094

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*. https://arxiv.org/abs/2306.11644

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., & Wilcox, E. G. (2024). Findings of the Second BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*. https://arxiv.org/abs/2412.05149

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Laukkonen, R. E., Inglis, F., Chandaria, S., Sandved-Smith, L., Lopez-Sola, E., Hohwy, J., Gold, J., & Elwood, A. (2025). Contemplative artificial intelligence. *arXiv preprint arXiv:2504.15125*. https://arxiv.org/abs/2504.15125

Menon, P. G. (2026). Persistent identity in AI agents: A multi-anchor architecture for resilient memory and continuity. *arXiv preprint arXiv:2604.09588*. https://doi.org/10.48550/arXiv.2604.09588

Moëll, B., & Sand Aronsson, F. (2026). High-accuracy prediction of mental health scores from English BERT embeddings trained on LLM-generated synthetic self-reports: A synthetic-only method development study. *Frontiers in Digital Health, 7*, 1694464. https://doi.org/10.3389/fdgth.2025.1694464

Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., Chan, D. M., & Canny, J. (2024). Virtual personas for language models via an anthology of backstories. *Proceedings of EMNLP 2024*. https://arxiv.org/abs/2407.06576

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2301.05217

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*. https://arxiv.org/abs/2201.02177

Silver, D., & Sutton, R. S. (2025). Welcome to the era of experience. In *Designing an Intelligence* (forthcoming). MIT Press. Preprint: https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf

Singh, K., Band, N., & Adeli, E. (2025). Curriculum-guided layer scaling for language model pretraining. *arXiv preprint arXiv:2506.11389*. https://arxiv.org/abs/2506.11389

Takata, R., Masumori, A., & Ikegami, T. (2024). Spontaneous emergence of agent individuality through social interactions in large language model-based communities. *Entropy, 26*(12), 1092. https://doi.org/10.3390/e26121092

Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining grokking through circuit efficiency. *arXiv preprint arXiv:2309.02390*. https://arxiv.org/abs/2309.02390

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. *Proceedings of the BabyLM Challenge (CoNLL 2023)*, 1–34. https://aclanthology.org/2023.conll-babylm.1/