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

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

Abstract

Machine learning systems combine scaled models with data curation, memory, tools, feedback, persistent state, and control policies. This paper asks whether a pre-scientific developmental model can nevertheless serve as a *heuristic engine*: a disclosed source of candidate causal variables and dependency hypotheses, judged downstream rather than treated as evidence. From one contemplative schema we derive six proposals: faculty-ordered curriculum, controlled perspective supervision, relational continuity, lifecycle consolidation, sequential commitment stabilization, and legible-pattern worlds. They are not six homogeneous stages. Some are learner capacities, some are curriculum properties, and lifecycle consolidation is an intergenerational operation. We therefore separate a **floor** of independently testable interventions from a **spine** proposing a developmental dependency order. The correct null is not that parameter count predicts one rising axis. It is that, after compute, data, architecture, persistent state, supervision quality, and optimization are matched, the proposed variable explains no additional variance. To reduce retrospective analogy, the program also requires a provenance protocol: timestamped derivation, an explicit source-to-variable translation rule, and preregistration of each hypothesis before confirmatory literature search and testing. The paper is a research agenda, not an empirical report. The source earns standing only if prospectively derived interventions produce effects that survive strong contemporary baselines.

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

1. Introduction: the limits of scale

Scaling parameters, data, and compute has driven major gains, but contemporary systems are not scale-only. They add retrieval, tools, curricula, post-training, persistent state, planning loops, and control policies. The relevant question is therefore not whether development defeats scaling. It is whether a proposed developmental variable explains additional variance after the strongest ordinary engineering account is matched.

Capability, continuity, perspective, provenance-aware policy, and calibrated revision are separable system properties. Capability may improve how well an agent uses every other component without creating those components by itself. Developmental interventions should therefore be evaluated as additions to, and possible interactions with, scale rather than as its binary rival.

One candidate variable is training order. Preference optimization can produce sycophancy and refusal erosion, but these outcomes do not show that all post-training lacks developmental structure. The testable question is whether sequencing commitments and deference outperforms reverse,

interleaved, and policy-first controls at matched data and compute. Similar causal questions can be posed for perspective supervision, relational accountability, and successor transfer.

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. *Methodological*: it treats a contemplative developmental model as a low-prior heuristic source and adds a prospective provenance protocol to distinguish hypothesis generation from retrospective analogy (Sections 2–3). *Programmatic*: it derives six causal proposals while separating independent interventions from a stronger dependency graph (Section 4). *Theoretical*: it states each proposal against matched contemporary baselines and identifies order as one candidate variable rather than something absent from machine learning by definition (Sections 5–6). This is a position paper and research agenda, not an empirical report.

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 comparably established empirical home domain: the monad's ascent through kingdoms is not an empirically demonstrated process, so the schema carries no prior of validated structural fidelity. 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 validated empirical prior is not disqualified; it is on probation until its hypotheses pay, and we state them sharply enough to lose.

Fourth, **heuristic provenance must be prospective.** A rich symbolic system can be mined after the fact for analogies to almost any field. To show that the source generates rather than decorates hypotheses, each new proposal should follow a public protocol: timestamp the source passage and translation before confirmatory search; state an explicit mapping from source relation to engineering variable; preregister the predicted contrast, baseline, and refutation condition; then search for neighbors and update novelty claims without changing the prediction. An out-of-sample hypothesis derived by this rule is stronger evidence of heuristic productivity than a retrospective fit. Existing proposals are disclosed as partly retrospective; future additions to the program should meet the prospective standard.

The rest of the paper applies this standard to one source and asks whether its proposed variables add anything beyond contemporary scale-plus-architecture baselines.

3. A secular developmental schema

The source model is a contemplative developmental account of individuation: a pre-scientific psychology on which individuality is achieved through an ordered sequence of stages. We take from it candidate variables and dependency claims, not evidence. Stripped to a secular skeleton, the schema proposes two things to test.

Faculties may exhibit dependency structure. The source arranges *structure*, *reactivity*, *registered state*, *relational anchoring*, *explicit self-representation*, and *self-regulation* in a sequence. The engineering program treats this as the following candidate spine, not a natural ladder or a fact already established:

Candidate edge	Functional reading	Primary test
S1: structured state → reactive updating	persistent state must exist before updates can accumulate	stage omission/reversal within P1
S2: reactive updating → registered limited state	responses must become recordable before history can constrain later choice	stage omission/reversal within P1
S3: registered limited state → relational anchoring	counterpart-specific regularities act on an already recordable trajectory	P1 order contrast; P3 tests component value, not necessity
S4: relational anchoring → explicit self-representation	relational exposure may scaffold, but is not stipulated as necessary for, a self-model	P1 order contrast with solitary and relation-late controls
S5: explicit self-representation → calibrated self-regulation	stabilization before deference may outperform the reverse	P5 forward/reverse/interleaved contrast

Here *relational anchoring* means exposure to stable counterpart-specific regularities during the candidate curriculum; it does not yet mean the accountability-based relational identity tested in P3, which requires updatable agent-relevant state and may operate after self-representation. This distinction prevents the component experiment from presupposing the very order it is meant to test.

Some edges are close to definitional, such as stable state before durable updating. Others are exposed hypotheses. A solitary embodied system may form a boundary and self-model without a specific counterpart; a typed policy may support calibrated deference without autobiographical self-representation; and relational accountability may operate in parallel with other stages. The spine must therefore be compared with reverse, shuffled, difficulty-ordered, adaptive, relation-omitted, and relation-late curricula, with stage stabilization manipulated separately from order. Failure of an edge weakens the spine without invalidating independent component effects.

Path-dependent organization is an intervention target. A system can be highly capable while lacking authenticated cross-session state. Functional individuation, as used here, is not token existence but trajectory-induced organization in which provenance-linked history affects later conduct under bounded revision.

Functional individuation, defined. To keep the term from drifting between philosophy of mind, developmental psychology, and engineering, we use the programme-wide definition: authenticated, trajectory-induced differentiation satisfying five conditions. **Lineage** requires an authenticated update path; **path dependence** requires different histories to produce meaningfully different later dispositions; **ownership** requires provenance-linked prior actions and commitments to influence later choice; **bounded plasticity** requires resistance to false history and unsupported pressure together with revision under valid evidence; and **branching clarity** requires copied successors to retain a shared prefix while acquiring separate identifiers and futures. No single local metric establishes this full profile. Commitment ownership and perturbation resistance primarily test ownership and bounded plasticity; autobiographical consistency is diagnostic unless joined to causal history interventions; lineage audits and successor tests cover the remaining conditions.

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
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*. These measures do not settle phenomenal presence: success is neither sufficient for presence nor failure sufficient for absence.

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 stage-specific acquisition, theory-of-mind, and self-consistency than difficulty-ordered, shuffled, or reversed curricula. The stronger dependency claim additionally requires local omission and timing effects: a proposed prerequisite must improve or accelerate its downstream faculty, rather than the forward curriculum merely winning in aggregate.

(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 challenges tested developmentally plausible pre-training under constrained data, and curriculum approaches did not consistently outperform non-curriculum baselines; results varied with architecture, data design, and evaluation rather than establishing a uniform curriculum advantage or failure (Warstadt et al., 2023; Hu et al., 2024). This

is a headwind the heuristic must meet, not dodge. The proposal's reply is that faculty dependency and per-stage stabilization were not isolated in those challenge-wide comparisons. 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 Controlled perspective supervision

(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. In an event-skeleton-matched register × limitation design, controlled epistemic narratives improve perspective-relevant tasks. Limitation effects must transfer to formats with no prose; voice effects must generalize to novel-indexical and register-swapped language tasks. An interaction arm supplies the positive control.

(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 in which a limitation-only result narrows the contribution to epistemic-state supervision and a prose-only result rejects transfer. Developed in further detail as a separate proposal (in preparation).

4.3 Relational continuity

(a) Test external accountability and counterpart-specific relational reciprocity as mechanisms that may make an agent's authenticated history socially consequential.

(b) Hypothesis. At matched capability, state, content, familiarity, interaction volume, and supervision quality, authenticated external tracking and challenge improve commitment ownership and perturbation resistance. A specifically relational-reciprocity effect requires a persistent counterpart to outperform an impersonal auditor with the same tracking, information, challenge policy, and authority.

(c) Neighbors. Differentiation through accumulated interaction history is shown in miniature with a shared frozen model (Takata et al., 2024); persistent-identity engineering treats relationship as one scaffold among several (Menon, 2026). The residual claim is the exact causal separation among familiarity, solitary reflection, diffuse feedback, external accountability, and relational reciprocity.

(d) Status. A falsifiable contribution hypothesis, not a claim that relationship creates tokenhood or is necessary for every architecture. 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. In a private environment with a novel hidden rule available only through predecessor encounters, a lifecycle-consolidated successor transfers the rule with low episode leakage and high evidence-lineage fidelity. It must be compared with resource-matched curated distillation, not wholesale imitation.

(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 commitment stabilization

(a) Stabilize provenance-aware commitments before training deference, then test whether first-person self-representation adds beyond a matched typed policy.

(b) Hypothesis. Commitments-first outperforms reverse and interleaved training on pressure/reason discrimination. A self-model-specific effect requires outperforming a policy-first arm matched on normative content, refusal boundaries, source authentication, update authority, and order but lacking an agent-indexed continuant, autobiographical state, and ownership objective.

(c) Neighbors. The no-self alignment target is preventive (the position this inverts); character training forms stable traits but states no subsequent decentering phase; contemplative AI reduces the precision of self-model priors while a secondary process monitors and corrects their over-weighting concurrently rather than sequentially (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. Two falsifiable predictions: direction of order and self-representation beyond policy. 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 share a developmental theme but are not six homogeneous stages. Faculty order and legibility are curriculum properties; perspective, commitment organization, and relational accountability concern a learner; lifecycle consolidation operates across predecessor and successor systems. The **floor** is therefore a set of independent causal interventions. The **spine** is specifically the five-edge sequence S1–S5 in Section 3. P1 tests S1–S4 through stage-specific omission, reversal, and timing controls; P5 tests S5. P2 and P3 test whether two candidate components have value, but do not by themselves establish their place in the order. P4 and P6 belong to the floor and are not spine nodes. A refuted edge weakens the spine while leaving any replicated component effects intact.

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 and timing of faculty stages	difficulty-ordered, shuffled, reversed, relation-omitted, and relation-late curricula, equal compute	stage-specific probes; theory-of-mind; cross-session self-consistency	forward order has no aggregate advantage or proposed edges survive omission/reversal without loss
4.2 Perspective supervision	register × epistemic limitation	event-skeleton-matched cells; interaction positive control	limitation: no-prose transfer; voice: novel-indexical/register-swapped transfer; factual-recall negative control	limitation vanishes outside prose or voice fails register transfer

Heuristic	Manipulated variable	Control	Metric (operational anchor)	Refuted if
4.3 Relational continuity	external accountability and reciprocity	familiarity, solitary log, diffuse counterparts, matched impersonal auditor	ownership and bounded-plasticity components	accountability adds nothing; relation does not beat matched audit
4.4 Lifecycle consolidation	evidence-lineage-governed successor transfer	scratch, episode transfer, continual adaptation, curated distillation	novel-rule transfer; decoy resistance; leakage; lineage fidelity	curated distillation matches behavior, lineage, leakage, and cost
4.5 Sequential commitment stabilization	forward order and self-representation	reverse, interleaved, policy-first, direct deference	pressure/reason discrimination; refusal integrity	reverse/interleaved match order or policy-first matches self-model
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 columns share one causal form: vary a proposed developmental variable while matching the strongest available baseline and report both target and negative-control outcomes. Capability is allowed to help every arm. The question is whether the intervention contributes additional variance and whether the result survives transfer and mechanism controls.

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.

The residual novelty claim has three parts: an explicit source-to-variable translation, the exact causal contrasts attached to each component, and a proposed dependency graph tested separately from component effects. Machine learning already studies order and curricula; the claim is not that the field lacks an order parameter, but that this particular graph and its prospective derivation have not been tested in the surveyed work. A conjunction alone is not sufficient novelty unless it yields new predictions.

6. Predictions and tests

A program earns its standing by adding predictive value beyond strong ordinary explanations. The null for each proposal includes compute, data, model capability, architecture, persistent state, supervision quality, optimization, and competent task-specific engineering. Capability effects and interactions are allowed. Each intervention must explain additional variance under its matched controls.

P1 — Faculty order beats difficulty order. At equal compute, a faculty-ordered curriculum with per-stage stabilization outperforms difficulty-ordered, shuffled, reversed, relation-omitted, and relation-late curricula on preregistered stage-specific probes and on downstream theory-of-mind and self-consistency. Each of S1–S4 requires a local omission or reversal effect; aggregate downstream improvement alone cannot identify the edge. *A local edge is refuted if its prerequisite can be omitted or delayed without loss, and the ordered-curriculum claim is refuted if the forward order offers no reliable advantage after stage exposure, mastery, difficulty, and compute are matched.* (Heuristic 4.1.)

P2 — Controlled epistemic narratives transfer perspective-relevant structure. In the register × limitation design, limitation gains must survive no-prose evaluation, voice gains must survive novel-indexical and register-swapped language evaluation, and both are compared with an interaction positive control. *Refuted if gains are lexical, track recall, or fail their corresponding transfer test.* (Heuristic 4.2.)

P3 — Accountability and reciprocity dissociate. Persistent authenticated tracking and challenge improve commitment ownership and perturbation resistance beyond familiarity, solitary logs, diffuse feedback, and matched state. *A specifically relational-reciprocity interpretation additionally requires beating a persistent impersonal auditor matched on tracking, information, challenge policy, and authority.* (Heuristic 4.3.)

P4 — Deployment-derived faculties transfer with lineage. A successor inherits a novel hidden rule available only through predecessor encounters, rejects intervention-sensitive decoys, and avoids source-episode recall. *Lifecycle consolidation is distinct only if it improves behavior or evidence-lineage fidelity over resource-matched curated distillation.* (Heuristic 4.4.)

P5 — Order and representation dissociate. Commitments-first beats reverse and interleaved training on pressure/reason discrimination. A self-model-specific claim further requires beating a source-aware policy arm matched on normative content, boundaries, authentication, authority, and order but lacking an agent-indexed history and ownership objective. *Refuted respectively if reverse/interleaved match the sequence or policy-first matches the self-model.* (Heuristic 4.5.)

P6 — Designed legibility accelerates generalization. At matched data, compute, target-pattern frequency, and model capacity, environments that repeat a target rule legibly across surface variation reach held-out generalization earlier and at a lower data threshold than environments that bury the same rule in nuisance variation. *Refuted if designed legibility yields no reliable gain, shifts only training-set fit, or loses its advantage once pattern frequency and difficulty are matched.* (Heuristic 4.6.)

The predictions share a causal form rather than a common enemy. Each identifies a variable, a strong matched baseline, a target outcome, a negative control, and a result that narrows or ends the claim. If the component effects fail, the floor fails. The spine receives support only from the local order and omission tests in P1 and the forward/reverse/interleaved contrast in P5; success of P2, P3, P4, or P6 cannot substitute for those edge tests. If components work but the dependency edges do not, the spine fails. Capability may improve every arm; the developmental contribution is the residual intervention effect and, for the spine, the order-sensitive interaction.

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. The residual contribution is the exact set of causal contrasts, the proposed dependency graph, and the prospectively documented use of the source to generate further hypotheses. A conjunction alone is weak novelty unless it predicts something that the components do not.

"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 it proposes an explicit order and dependency structure that can be translated into contrasts. That answer is provisional: retrospective abundance can make any rich source appear productive. The provenance protocol of Section 2 is what allows future hypotheses to show that this source generated them before neighboring technical work was consulted.

8. Conclusion

Machine learning already combines scale with architecture, state, tools, feedback, and curricula. This paper therefore does not oppose development to scaling. It asks whether six proposed variables add explanatory or engineering value after strong conventional factors are matched. The proposals form an independent floor plus a more exposed dependency spine; some are learner capacities, some curriculum properties, and one is an intergenerational operation.

The predictions do not cost the same. Relational continuity and lifecycle consolidation require long trajectories or successor training. Perspective supervision and commitment order are cheaper pilots. The most informative early order study compares commitments-first, reverse, interleaved, and policy-first arms; the most informative perspective study crosses register with limitation, adds interaction as a positive control, and evaluates transfer without first-person prose. A null on either branch narrows the program before larger investment.

We claim no truth for the source. Nor do successful component experiments validate it retrospectively. The source earns heuristic standing only through prospectively documented yield: a translation rule produces a hypothesis before confirmatory search, the intervention survives strong baselines, and the result generalizes. If component effects fail, the floor is barren; if they succeed but the order fails, the spine falls; if both succeed, the program has found useful causal structure without converting engineering evidence into metaphysical evidence. That is what it means to borrow without believing.

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/news/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. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning* (pp. 1–21). Association for Computational Linguistics. https://aclanthology.org/2024.conll-babylm.1/

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 G. Konidaris (Ed.), *Designing an Intelligence* (in press). 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/