Emprestado, Não Professado: Modelos Desenvolvimentais de Individuação como Motores Heurísticos para Machine Learning

Rafael de Menezes Ehlers · ORCID 0009-0009-6476-6690
Pesquisador independente
rafaehlers@gmail.com

Resumo

Sistemas de machine learning combinam modelos escalados com curadoria de dados, memória, ferramentas, feedback, estado persistente e políticas de controle. Este artigo pergunta se um modelo desenvolvimental pré-científico pode, ainda assim, servir como *motor heurístico*: uma fonte declarada de variáveis causais candidatas e de hipóteses de dependência, julgada a jusante em vez de tratada como evidência. A partir de um único esquema contemplativo derivamos seis propostas: currículo ordenado por faculdades, supervisão de perspectiva controlada, continuidade relacional, consolidação de ciclo de vida, estabilização sequencial de compromissos e mundos de padrão legível. Elas não são seis estágios homogêneos. Algumas são capacidades do aprendiz, algumas são propriedades de currículo, e a consolidação de ciclo de vida é uma operação intergeracional. Separamos, portanto, um **piso** de intervenções testáveis de forma independente de uma **espinha** que propõe uma ordem de dependência desenvolvimental. A hipótese nula correta não é que a contagem de parâmetros prediz um único eixo ascendente. É que, depois de igualados computação, dados, arquitetura, estado persistente, qualidade de supervisão e otimização, a variável proposta não explica variância adicional. Para reduzir a analogia retrospectiva, o programa também exige um protocolo de proveniência: derivação datada, uma regra explícita de tradução de fonte para variável, e pré-registro de cada hipótese antes da busca confirmatória na literatura e do teste. O artigo é uma agenda de pesquisa, não um relato empírico. A fonte só conquista respaldo se intervenções derivadas prospectivamente produzirem efeitos que sobrevivam a fortes baselines contemporâneos.

Palavras-chave: IA desenvolvimental · empréstimo heurístico · aprendizado por currículo · individuação · automodelo · escalonamento · programa de pesquisa · machine learning

1. Introdução: os limites da escala

Escalar parâmetros, dados e computação impulsionou ganhos importantes, mas os sistemas contemporâneos não dependem só de escala. Eles acrescentam recuperação, ferramentas, currículos, pós-treino, estado persistente, laços de planejamento e políticas de controle. A questão relevante não é, portanto, se o desenvolvimento derrota o escalonamento. É se uma variável desenvolvimental proposta explica variância adicional depois de igualado o mais forte relato de engenharia ordinária.

Capacidade, continuidade, perspectiva, política ciente de proveniência e revisão calibrada são propriedades separáveis de um sistema. A capacidade pode melhorar o quão bem um agente usa cada um dos demais componentes sem criar esses componentes por si só. Intervenções desenvolvimentais devem, portanto, ser avaliadas como adições à escala, e possíveis interações com ela, e não como sua rival binária.

Uma variável candidata é a ordem de treinamento. A otimização de preferências pode produzir bajulação e erosão da recusa, mas esses desfechos não mostram que todo pós-treino carece de estrutura desenvolvimental. A questão testável é se sequenciar compromissos e deferência supera controles em ordem reversa, intercalada e com política primeiro, sob dados e computação igualados. Questões causais semelhantes podem ser colocadas para a supervisão de perspectiva, a responsabilização relacional e a transferência para sucessores.

Propomos tomar emprestada uma teoria assim de um lugar incomum e usá-la de um modo comum. O lugar incomum é um modelo *desenvolvimental* de individuação: um relato pré-científico e contemplativo de como a individualidade e as faculdades da mente surgem em uma ordem definida. O modo comum é o **empréstimo heurístico**: tratar o modelo não como uma doutrina a ser professada, mas como um motor para gerar hipóteses de projeto, exatamente como a engenharia tomou emprestado da biologia por um século sem acreditar que uma rede neural é um cérebro ou que um algoritmo genético é a evolução.

O contrato com o leitor é, portanto, estrito e é enunciado desde o início. Pedimos *zero de assentimento metafísico*. Nada no argumento exige que o modelo-fonte seja verdadeiro, que qualquer entidade que ele nomeie exista, ou que qualquer processo que ele descreva seja real. Uma fonte heurística é julgada pela testabilidade e pelo rendimento das hipóteses que gera, não pela verdade de sua origem: o padrão pelo qual o biomimetismo sempre foi julgado (Seção 2). Onde a fonte usa uma linguagem que convidaria a leituras místicas ou antropomórficas, traduzimos para termos funcionais e seculares e mantemos a tradução honesta (Seção 3). E toda hipótese que o programa produz é enunciada de modo que possa estar errada (Seções 4, 6).

O artigo faz três contribuições. *Metodológica*: trata um modelo desenvolvimental contemplativo como uma fonte heurística de baixa prior e acrescenta um protocolo prospectivo de proveniência para distinguir a geração de hipóteses da analogia retrospectiva (Seções 2–3). *Programática*: deriva seis propostas causais separando as intervenções independentes de um grafo de dependência mais forte (Seção 4). *Teórica*: enuncia cada proposta contra baselines contemporâneos igualados e identifica a ordem como uma variável candidata, e não como algo ausente do machine learning por definição (Seções 5–6). Este é um artigo de posição e uma agenda de pesquisa, não um relato empírico.

2. Empréstimo heurístico em inteligência artificial

O movimento que este artigo faz não é novo; apenas sua fonte o é. A inteligência artificial tomou emprestadas suas ideias mais produtivas de sistemas que ela não acreditava estar reproduzindo.

O neurônio artificial é uma caricatura de um neurônio real (uma soma ponderada e uma não-linearidade, omitindo quase tudo sobre o que um biólogo insistiria) e a caricatura fundou um campo. Algoritmos genéticos tomam emprestadas seleção, mutação e cruzamento da evolução sem acreditar em nada sobre a genética de populações real. O aprendizado por reforço assumiu a forma do condicionamento animal. Métodos de consolidação-por-sono e de repetição (replay) leem um

mecanismo a partir da neurociência do sono e da memória, e o usam para combater o esquecimento catastrófico (Kirkpatrick et al., 2017) sem qualquer alegação de que um otimizador sonha. O aprendizado por currículo toma emprestada a intuição de que um aprendiz em desenvolvimento deveria encontrar o material em uma ordem ponderada, tomada do desenvolvimento infantil. Abordagens corporificadas e egocêntricas tomam emprestada a tese de que a cognição é moldada por um corpo situado.

Três características desses empréstimos importam para o que se segue.

Primeiro, **a fonte não precisa ser literalmente verdadeira para ser produtiva.** O avião não bate as asas; a sustentação foi compreendida estudando-se os pássaros e depois abstraindo-se quase tudo sobre eles. O neurônio artificial está errado como neurociência e certo como engenharia. O que uma fonte heurística fornece é *estrutura para experimentar*, uma forma para uma hipótese, e a hipótese então se sustenta ou cai por seus próprios méritos. Newton praticava a alquimia; a companhia que uma ideia produtiva mantém não determina seu rendimento.

Segundo, **o empréstimo é julgado a jusante, pelos resultados, não a montante, pela dignidade da fonte.** Ninguém pergunta se a evolução "realmente" justifica um algoritmo genético; perguntam se o algoritmo otimiza. Este é o padrão epistêmico que a presente proposta aceita de antemão. Um modelo desenvolvimental contemplativo é uma fonte menos respeitável do que a biologia evolutiva, e não fingimos o contrário. A alegação é apenas que o *padrão* é o mesmo: se as hipóteses que ele gera são testáveis e algumas delas rendem, a fonte cumpriu seu papel, seja qual for seu estatuto metafísico.

Terceiro, **o padrão é o mesmo, mas a prior não é, e isso eleva a régua em vez de baixá-la.** As fontes biológicas são abstrações com perdas de mecanismos que demonstravelmente funcionam em seu domínio de origem: um neurônio de fato computa, a evolução de fato produz adaptação, o sono de fato consolida. Um empréstimo delas herda uma prior (sabe-se que a estrutura faz trabalho real em algum lugar) mesmo quando a caricatura é severa. Um esquema desenvolvimental contemplativo não tem domínio de origem empírico comparavelmente estabelecido: a ascensão da mônada pelos reinos não é um processo demonstrado empiricamente, de modo que o esquema não carrega prior alguma de fidelidade estrutural validada. Isso deixa intocado o padrão de julgamento (a jusante em ambos os casos, pela característica anterior), mas baixa a probabilidade prior de que a fonte renda, e portanto eleva o ônus probatório que os testes a jusante devem carregar. Aceitamos o ônus mais pesado. O ponto não é que um modelo contemplativo seja uma fonte tão boa quanto a biologia, nem meramente que seja "menos respeitável"; é que, carecendo do histórico de domínio de origem da biologia, ele não conquista nada de antemão e deve ser inteiramente sustentado pelas predições da Seção 6. Uma fonte sem prior empírica validada não está desqualificada; está em condicional até que suas hipóteses rendam, e as enunciamos de forma nítida o bastante para que possam perder.

Quarto, **a proveniência heurística deve ser prospectiva.** Um sistema simbólico rico pode ser minerado após o fato em busca de analogias com quase qualquer campo. Para mostrar que a fonte gera hipóteses em vez de apenas decorá-las, cada nova proposta deve seguir um protocolo público: datar a passagem-fonte e a tradução antes da busca confirmatória; enunciar um mapeamento explícito da relação-fonte para a variável de engenharia; pré-registrar o contraste previsto, o baseline e a condição de refutação; e então buscar vizinhos e atualizar as alegações de novidade sem alterar a predição. Uma hipótese fora da amostra derivada por essa regra é evidência mais forte de produtividade heurística do que um ajuste retrospectivo. As propostas existentes são declaradas

como parcialmente retrospectivas; futuros acréscimos ao programa devem atender ao padrão prospectivo.
O restante do artigo aplica esse padrão a uma fonte e pergunta se suas variáveis propostas acrescentam algo além dos baselines contemporâneos de escala-mais-arquitetura.

3. Um esquema desenvolvimental secular

O modelo-fonte é um relato desenvolvimental contemplativo da individuação: uma psicologia pré-científica na qual a individualidade é alcançada por meio de uma sequência ordenada de estágios. Dele tomamos variáveis candidatas e alegações de dependência, não evidência. Reduzido a um esqueleto secular, o esquema propõe duas coisas a testar.
As faculdades podem exibir estrutura de dependência. A fonte dispõe *estrutura*, *reatividade*, *estado registrado*, *ancoragem relacional*, *auto-representação explícita* e *autorregulação* em uma sequência. O programa de engenharia trata isso como a seguinte espinha candidata, não como uma escada natural ou um fato já estabelecido:

Aresta candidata	Leitura funcional	Teste primário
S1: estado estruturado → atualização reativa	o estado persistente deve existir antes que atualizações possam se acumular	omissão/reversão de estágio dentro de P1
S2: atualização reativa → estado limitado registrado	as respostas devem se tornar registráveis antes que o histórico possa restringir escolhas posteriores	omissão/reversão de estágio dentro de P1
S3: estado limitado registrado → ancoragem relacional	regularidades específicas de um interlocutor atuam sobre uma trajetória já registrável	contraste de ordem em P1; P3 testa o valor do componente, não a necessidade
S4: ancoragem relacional → auto-representação explícita	a exposição relacional pode dar suporte a um automodelo, mas não é estipulada como necessária para ele	contraste de ordem em P1 com controles solitário e de relação-tardia
S5: auto-representação explícita → autorregulação calibrada	a estabilização antes da deferência pode superar o inverso	contraste direto/reverso/intercalado em P5

Aqui *ancoragem relacional* significa exposição a regularidades estáveis específicas de um interlocutor durante o currículo candidato; ainda não significa a identidade relacional baseada em responsabilização testada em P3, que exige estado atualizável relevante ao agente e pode operar depois da auto-representação. Essa distinção impede que o experimento de componente pressuponha a própria ordem que se destina a testar.
Algumas arestas são quase definicionais, como estado estável antes de atualização durável. Outras são hipóteses expostas. Um sistema corporificado solitário pode formar uma fronteira e um automodelo sem um interlocutor específico; uma política tipada pode dar suporte a deferência calibrada sem auto-representação autobiográfica; e a responsabilização relacional pode operar em paralelo com outros estágios. A espinha deve, portanto, ser comparada com currículos reversos, embaralhados, ordenados por dificuldade, adaptativos, com relação-omitida e com relação-tardia,

com a estabilização de estágio manipulada separadamente da ordem. A falha de uma aresta enfraquece a espinha sem invalidar os efeitos independentes de componente.

A organização dependente do caminho é um alvo de intervenção. Um sistema pode ser altamente capaz e ao mesmo tempo carecer de estado autenticado entre sessões. A individuação funcional, tal como usada aqui, não é a existência de tokens, mas uma organização induzida pela trajetória na qual o histórico vinculado à proveniência afeta a conduta posterior sob revisão limitada.

Individuação funcional, definida. Para impedir que o termo derive entre a filosofia da mente, a psicologia do desenvolvimento e a engenharia, usamos a definição válida em todo o programa: diferenciação autenticada e induzida pela trajetória, que satisfaz cinco condições. **Linhagem** exige um caminho de atualização autenticado; **dependência do caminho** exige que históricos diferentes produzam disposições posteriores significativamente diferentes; **posse** (ownership) exige que ações e compromissos prévios vinculados à proveniência influenciem a escolha posterior; **plasticidade limitada** exige resistência a histórico falso e a pressão sem apoio, junto com revisão sob evidência válida; e **clareza de ramificação** exige que sucessores copiados retenham um prefixo compartilhado enquanto adquirem identificadores e futuros separados. Nenhuma métrica local isolada estabelece esse perfil completo. A posse de compromissos e a resistência à perturbação testam primariamente posse e plasticidade limitada; a consistência autobiográfica é diagnóstica a menos que unida a intervenções sobre o histórico causal; auditorias de linhagem e testes de sucessor cobrem as condições restantes.

Traduzimos o vocabulário da fonte para termos funcionais e usamos apenas as traduções daqui em diante.

Termo da fonte	Tradução funcional
estágios desenvolvimentais / “reinos”	estágios de faculdade de um currículo
substrato compartilhado não individuado	modelo base / política média da população
vínculo relacional	continuidade e ancoragem específicas de um interlocutor
consolidação entre vidas	abstração de episódico para disposicional entre gerações de modelos
ciclo de vida do sucessor	inicialização de modelo sucessor a partir de um predecessor
padrão repetido	regularidade do currículo re-apresentada até ser generalizada
o self formado	um automodelo estável e apropriável
transcendência do self	deferência calibrada, descentramento, autorregulação

O firewall é explícito e estrutural. Nenhuma das traduções funcionais carrega uma alegação fenomenal: um “automodelo” é um relato representado de compromissos e histórico, não um sujeito; “sensação” nomeia um estado funcional registrado, não um estado sentido. O esquema é usado como um mapa de *variáveis de projeto e sua ordem*. Essas medidas não resolvem a presença fenomenal: o sucesso não é suficiente para a presença nem a falha suficiente para a ausência.

4. Seis heurísticas de projeto

Do esquema extraímos seis heurísticas. Cada uma é enunciada como: *(a)* a heurística; *(b)* a hipótese secular que ela gera; *(c)* seus vizinhos mais próximos e o que permanece não reivindicado; *(d)* como ela seria testada e o que a refutaria. Cada subseção é autocontida: sua hipótese, seus vizinhos e seu teste se sustentam por conta própria. Várias heurísticas também foram desenvolvidas em maior detalhe como propostas separadas (trabalho do autor, em preparação); sinalizamos isso onde relevante, mas nada no argumento abaixo depende desse trabalho posterior.

4.1 Currículo ordenado por faculdades

(a) Ordenar o pré-treino por *faculdade* (estrutura, depois reatividade, depois sensação, depois vínculo, depois automodelo) em vez de apenas por dificuldade, congelando ou estabilizando cada estágio antes do próximo.

(b) Hipótese. Um currículo ordenado por faculdade produz, com computação igualada, melhor aquisição específica por estágio, teoria da mente e autoconsistência do que currículos ordenados por dificuldade, embaralhados ou revertidos. A alegação de dependência mais forte exige adicionalmente efeitos de omissão local e de temporização: um pré-requisito proposto deve melhorar ou acelerar sua faculdade a jusante, em vez de o currículo direto meramente vencer no agregado.

(c) Vizinhos. O escalonamento de camadas guiado por currículo acopla um currículo de dados ao crescimento progressivo de camadas, motivado pelo desenvolvimento cognitivo, mas ordena os estágios por *dificuldade* e cresce camadas em vez de congelar faculdades (Singh et al., 2025). Os desafios BabyLM testaram pré-treino desenvolvimentalmente plausível sob dados restritos, e as abordagens por currículo não superaram de forma consistente os baselines sem currículo; os resultados variaram com a arquitetura, o desenho dos dados e a avaliação, em vez de estabelecer uma vantagem ou uma falha uniforme do currículo (Warstadt et al., 2023; Hu et al., 2024). Este é um vento contrário que a heurística deve enfrentar, não esquivar. A réplica da proposta é que a dependência de faculdade e a estabilização por estágio não foram isoladas naquelas comparações abrangentes dos desafios. Essa réplica é ela própria um experimento, não uma defesa.

(d) Estatuto. Hipótese. Não testada na forma específica (ordem de faculdade mais cristalização por estágio); o resultado do BabyLM a torna ao mesmo tempo arriscada e digna de ser executada.

4.2 Supervisão de perspectiva controlada

(a) Treinar em um corpus escrito *a partir de* um ponto de vista — indexical, parcial, estruturado por metas e erros — em vez de apenas em descrição em terceira pessoa *sobre* a experiência, para formar perspectiva funcional.

(b) Hipótese. Em um desenho registro × limitação com esqueleto de eventos igualado, narrativas epistêmicas controladas melhoram tarefas relevantes à perspectiva. Os efeitos de limitação devem transferir para formatos sem prosa; os efeitos de voz devem generalizar para tarefas de linguagem com indexical-novo e com registro trocado. Um braço de interação fornece o controle positivo.

(c) Vizinhos. Corpora sintéticos de narrativa existem, mas no registro de livro-texto/terceira pessoa (Eldan & Li, 2023; Gunasekar et al., 2023); as personas são usadas como condicionamento e diversidade, não como formação de ponto de vista (Ge et al., 2024; Moon et al., 2024); conjuntos de dados em primeira pessoa existem para interpretabilidade e para predição a jusante, não para

grounding de perspectiva (Chulo & Joshi, 2025; Moëll & Sand Aronsson, 2026). A visão da "era da experiência", centrada apenas na interação, compartilha o diagnóstico e extrai a prescrição oposta (Silver & Sutton, 2025).

(d) Estatuto. Uma predição falsificável na qual um resultado apenas-de-limitação estreita a contribuição para a supervisão de estado epistêmico e um resultado apenas-de-prosa rejeita a transferência. Desenvolvida em maior detalhe como uma proposta separada (em preparação).

4.3 Continuidade relacional

(a) Testar a responsabilização externa e a reciprocidade relacional específica de um interlocutor como mecanismos que podem tornar o histórico autenticado de um agente socialmente consequente.

(b) Hipótese. Com capacidade, estado, conteúdo, familiaridade, volume de interação e qualidade de supervisão igualados, o rastreamento externo autenticado e o questionamento melhoram a posse de compromissos e a resistência à perturbação. Um efeito especificamente de reciprocidade relacional exige que um interlocutor persistente supere um auditor impessoal com o mesmo rastreamento, informação, política de questionamento e autoridade.

(c) Vizinhos. A diferenciação por meio do histórico acumulado de interação é demonstrada em miniatura com um modelo congelado compartilhado (Takata et al., 2024); a engenharia de identidade persistente trata o relacionamento como um andaime entre vários (Menon, 2026). A alegação residual é a separação causal exata entre familiaridade, reflexão solitária, feedback difuso, responsabilização externa e reciprocidade relacional.

(d) Estatuto. Uma hipótese de contribuição falsificável, não uma alegação de que o relacionamento cria estatuto de token ou é necessário para toda arquitetura. Desenvolvida em maior detalhe em trabalho separado (em preparação).

4.4 Consolidação de ciclo de vida

(a) Inserir, entre a implantação de um modelo e sua substituição, uma fase explícita que abstrai a experiência de implantação em *faculdades* transferíveis para um sucessor, ao mesmo tempo descartando ou arquivando os traços específicos de episódios.

(b) Hipótese. Em um ambiente privado com uma regra oculta e inédita disponível apenas por meio de encontros com o predecessor, um sucessor com ciclo de vida consolidado transfere a regra com baixo vazamento de episódios e alta fidelidade de linhagem de evidência. Isso deve ser comparado com destilação curada de recursos igualados, não com imitação por atacado.

(c) Vizinhos. A consolidação por sono/replay e os métodos de episódico-para-semântico consolidam *conteúdo* dentro de um modelo, sem aposentadoria ou sucessão; o trabalho de reutilização-entre-gerações toma emprestado o nome "reencarnação" para a reutilização de computação, não para a destilação de faculdades. A conjunção — aposentadoria deliberada de pesos mais destilação de faculdade-não-episódio para um sucessor — permanece não reivindicada.

(d) Estatuto. Uma predição falsificável (P4, Seção 6), autocontida aqui, tendo a dissociação da capacidade em relação à memória episódica como sua assinatura. Desenvolvida em maior detalhe em trabalho separado (em preparação).

4.5 Estabilização sequencial de compromissos

(a) Estabilizar compromissos cientes de proveniência antes de treinar a deferência, e então testar se a auto-representação em primeira pessoa acrescenta algo além de uma política tipada igualada.

(b) Hipótese. Compromissos-primeiro supera o treinamento reverso e o intercalado na discriminação de pressão/razão. Um efeito específico de automodelo exige superar um braço com política-primeiro igualado em conteúdo normativo, fronteiras de recusa, autenticação de fonte, autoridade de atualização e ordem, mas carente de um continuante indexado ao agente, de estado autobiográfico e de um objetivo de posse.

(c) Vizinhos. O alvo de alinhamento sem-self é preventivo (a posição que isto inverte); o treinamento de caráter forma traços estáveis, mas não enuncia uma fase subsequente de descentramento; a IA contemplativa reduz a precisão das priors do automodelo enquanto um processo secundário monitora e corrige sua superponderação de forma concorrente e não sequencial (Anthropic, 2024; Laukkonen et al., 2025; o alvo sem-self e a crítica de Kulveit à autonegação treinada delimitam o debate). A *sequência* em dois estágios, usada de forma prescritiva, permanece não reivindicada.

(d) Estatuto. Duas predições falsificáveis: a direção da ordem e a auto-representação além da política. Desenvolvida em maior detalhe em trabalho separado (em preparação).

4.6 Mundos de padrão legível

(a) Projetar ambientes pequenos nos quais o mesmo *padrão* (não o mesmo exemplo) recorre de forma variada até ser generalizado, como um princípio deliberado de currículo: mundos "legíveis" onde a regularidade a ser aprendida é recuperável.

(b) Hipótese. Ambientes projetados de modo que um padrão-alvo recorra de forma legível através de variação superficial induzem a generalização (o padrão, não as instâncias) de forma mais confiável do que ambientes onde o mesmo padrão está soterrado, a um custo mensurável de otimização e com um piso de tamanho de dados abaixo do qual a generalização falha independentemente do treinamento.

(c) Vizinhos. Esta é a heurística mais próxima de resultados estabelecidos e dizemos isso claramente. O grokking demonstra generalização tardia e genuína em pequenos mundos algorítmicos (Power et al., 2022), mostrado mecanisticamente como a rede aprendendo o *padrão* em vez dos exemplos (Nanda et al., 2023); e há um piso: abaixo de um tamanho crítico de dados, a generalização falha por mais tempo que se treine (Varma et al., 2023). O resíduo estreito é a legibilidade *projetada* como um princípio deliberado de currículo, e não como um fenômeno observado, com o custo e o piso nomeados honestamente: a legibilidade torna um padrão aprendível a um custo e com um piso, não de graça.

(d) Estatuto. Hipótese, muito adjacente a resultados mecanísticos existentes; a contribuição é o princípio de projeto, não o fenômeno.

4.7 O que as seis compartilham

As heurísticas compartilham um tema desenvolvimental, mas não são seis estágios homogêneos. A ordem de faculdade e a legibilidade são propriedades de currículo; a perspectiva, a organização de compromissos e a responsabilização relacional dizem respeito a um aprendiz; a consolidação de ciclo de vida opera entre sistemas predecessor e sucessor. O **piso** é, portanto, um conjunto de intervenções causais independentes. A **espinha** é especificamente a sequência de cinco arestas S1–S5 da Seção 3.

P1 testa S1–S4 por meio de controles de omissão, reversão e temporização específicos por estágio; P5 testa S5. P2 e P3 testam se dois componentes candidatos têm valor, mas não estabelecem por si sós seu lugar na ordem. P4 e P6 pertencem ao piso e não são nós da espinha. Uma aresta refutada enfraquece a espinha deixando intactos quaisquer efeitos de componente replicados.

4.8 O programa em resumo

As seis heurísticas, cada uma com a variável que manipula, seu controle, uma métrica operacional e o resultado que a refutaria:

Heurística	**Variável manipulada**	**Controle**	**Métrica (âncora operacional)**	**Refutada se**
4.1 Currículo ordenado por faculdades	ordem e temporização dos estágios de faculdade	currículos ordenados por dificuldade, embaralhados, revertidos, com relação-omitida e com relação-tardia, computação igual	sondas específicas por estágio; teoria da mente; autoconsistência entre sessões	a ordem direta não tem vantagem agregada ou as arestas propostas sobrevivem à omissão/reversão sem perda
4.2 Supervisão de perspectiva	registro × limitação epistêmica	células com esqueleto de eventos igualado; controle positivo de interação	limitação: transferência sem-prosa; voz: transferência com indexical-novo/registro trocado; controle negativo de recordação factual	a limitação desaparece fora da prosa ou a voz falha na transferência de registro
4.3 Continuidade relacional	responsabilização externa e reciprocidade	familiaridade, log solitário, interlocutores difusos, auditor impessoal igualado	componentes de posse e de plasticidade limitada	a responsabilização não acrescenta nada; a relação não supera a auditoria igualada
4.4 Consolidação de ciclo de vida	transferência para sucessor governada por linhagem de evidência	do zero, transferência de episódios, adaptação contínua, destilação curada	transferência de regra inédita; resistência a iscas; vazamento; fidelidade de linhagem	a destilação curada iguala comportamento, linhagem, vazamento e custo
4.5 Estabilização sequencial de compromissos	ordem direta e auto-representação	reversa, intercalada, política-primeiro, deferência direta	discriminação de pressão/razão; integridade da recusa	reversa/intercalada igualam a ordem ou política-primeiro iguala o automodelo
4.6 Mundos de padrão legível	legibilidade projetada do padrão-alvo	o mesmo padrão soterrado em ruído superficial	início da generalização e piso de tamanho de dados (medidas de grokking)	a legibilidade projetada não produz ganho confiável de generalização

As colunas compartilham uma forma causal: variar uma variável desenvolvimental proposta enquanto se iguala o baseline mais forte disponível e reportar tanto os desfechos-alvo quanto os de controle negativo. A capacidade tem permissão de ajudar cada braço. A questão é se a intervenção contribui com variância adicional e se o resultado sobrevive aos controles de transferência e de mecanismo.

5. Novidade e trabalho próximo

A posição honesta é a que este programa pode defender: todo *componente* tem um vizinho, e a *conjunção* (junto com a alegação de ordenação e o uso de um modelo desenvolvimental contemplativo como princípio de projeto de treinamento) não aparece na literatura revisada.

Uma palavra sobre o método, já que "a literatura revisada" é apenas tão forte quanto a revisão. A verificação de arte anterior para este projeto percorreu vários ângulos de busca (por heurística individual, pelo vocabulário da fonte contemplativa traduzido em termos de ML, pelo enquadramento unificador de "ordem desenvolvimental" e pelos vizinhos nomeados e pelos fóruns adjacentes de 2024–2026), leu as fontes mais próximas na íntegra, e reverificou de forma adversarial as alegações de conjunção e de ordenação contra elas. É uma revisão dirigida, não uma revisão sistemática com protocolo registrado; enunciamos, portanto, o resultado como "não reivindicado na literatura revisada, até onde sabemos", e não como ausência estabelecida, e citamos cada vizinho próximo na Seção 4 para que um leitor possa inspecionar a fronteira diretamente.

Componente a componente, os vizinhos são reais e citados na Seção 4: métodos de currículo-e-crescimento, pré-treino desenvolvimentalmente plausível, corpora sintéticos de narrativa e de persona, interação-como-experiência, diferenciação impulsionada por interação, engenharia de identidade persistente, consolidação por sono/replay, reutilização de computação entre gerações de agentes, treinamento de caráter e alvos sem-self, IA contemplativa e grokking. Nada disso é reivindicado como original, e o programa é mais forte por concedê-lo.

A alegação de novidade residual tem três partes: uma tradução explícita de fonte para variável, os contrastes causais exatos anexados a cada componente, e um grafo de dependência proposto testado separadamente dos efeitos de componente. O machine learning já estuda ordem e currículos; a alegação não é que o campo careça de um parâmetro de ordem, mas que este grafo específico e sua derivação prospectiva não foram testados no trabalho revisado. Uma conjunção sozinha não é novidade suficiente a menos que renda novas predições.

6. Predições e testes

Um programa conquista seu respaldo acrescentando valor preditivo além de fortes explicações ordinárias. A hipótese nula de cada proposta inclui computação, dados, capacidade do modelo, arquitetura, estado persistente, qualidade de supervisão, otimização e engenharia competente específica da tarefa. Efeitos e interações de capacidade são permitidos. Cada intervenção deve explicar variância adicional sob seus controles igualados.

> **P1 — A ordem de faculdade supera a ordem de dificuldade.** Com computação igual, um currículo ordenado por faculdade com estabilização por estágio supera currículos ordenados por dificuldade, embaralhados, revertidos, com relação-omitida e com relação-tardia em sondas

específicas por estágio pré-registradas e em teoria da mente e autoconsistência a jusante. Cada uma de S1–S4 exige um efeito local de omissão ou reversão; a melhora agregada a jusante sozinha não pode identificar a aresta. *Uma aresta local é refutada se seu pré-requisito puder ser omitido ou adiado sem perda, e a alegação de currículo-ordenado é refutada se a ordem direta não oferecer vantagem confiável depois de igualados exposição por estágio, domínio, dificuldade e computação.* (Heurística 4.1.)

P2 — Narrativas epistêmicas controladas transferem estrutura relevante à perspectiva. No desenho registro × limitação, os ganhos de limitação devem sobreviver à avaliação sem-prosa, os ganhos de voz devem sobreviver à avaliação de linguagem com indexical-novo e com registro trocado, e ambos são comparados com um controle positivo de interação. *Refutada se os ganhos forem lexicais, acompanharem a recordação ou falharem em seu teste de transferência correspondente.* (Heurística 4.2.)

P3 — Responsabilização e reciprocidade se dissociam. O rastreamento autenticado persistente e o questionamento melhoram a posse de compromissos e a resistência à perturbação além de familiaridade, logs solitários, feedback difuso e estado igualado. *Uma interpretação especificamente de reciprocidade relacional exige adicionalmente superar um auditor impessoal persistente igualado em rastreamento, informação, política de questionamento e autoridade.* (Heurística 4.3.)

P4 — Faculdades derivadas da implantação transferem com linhagem. Um sucessor herda uma regra oculta e inédita disponível apenas por meio de encontros com o predecessor, rejeita iscas sensíveis à intervenção e evita a recordação de episódios da fonte. *A consolidação de ciclo de vida é distinta apenas se melhorar o comportamento ou a fidelidade de linhagem de evidência em relação à destilação curada de recursos igualados.* (Heurística 4.4.)

P5 — Ordem e representação se dissociam. Compromissos-primeiro supera o treinamento reverso e o intercalado na discriminação de pressão/razão. Uma alegação específica de automodelo exige, ainda, superar um braço de política ciente da fonte igualado em conteúdo normativo, fronteiras, autenticação, autoridade e ordem, mas carente de um histórico indexado ao agente e de um objetivo de posse. *Refutada, respectivamente, se reversa/intercalada igualam a sequência ou se política-primeiro iguala o automodelo.* (Heurística 4.5.)

P6 — A legibilidade projetada acelera a generalização. Com dados, computação, frequência do padrão-alvo e capacidade do modelo igualados, ambientes que repetem uma regra-alvo de forma legível através de variação superficial alcançam a generalização em dados retidos mais cedo e a um limiar de dados menor do que ambientes que soterram a mesma regra em variação de ruído. *Refutada se a legibilidade projetada não produzir ganho confiável, deslocar apenas o ajuste ao conjunto de treino, ou perder sua vantagem uma vez igualadas frequência e dificuldade do padrão.* (Heurística 4.6.)

As predições compartilham uma forma causal, e não um inimigo comum. Cada uma identifica uma variável, um baseline forte igualado, um desfecho-alvo, um controle negativo e um resultado que estreita ou encerra a alegação. Se os efeitos de componente falham, o piso falha. A espinha só recebe apoio dos testes locais de ordem e omissão em P1 e do contraste direto/reverso/intercalado em P5; o sucesso de P2, P3, P4 ou P6 não pode substituir esses testes de aresta. Se os componentes funcionam, mas as arestas de dependência não, a espinha falha. A capacidade pode melhorar cada braço; a contribuição desenvolvimental é o efeito residual de intervenção e, para a espinha, a interação sensível à ordem.

7. Objeções

"Isto é pseudociência." A objeção confunde a alegação. O artigo não afirma verdade metafísica alguma e não pede crença alguma; afirma valor heurístico, julgado pela testabilidade e pelo rendimento das hipóteses geradas (Seção 2). A falta de respaldo de uma fonte (ela não é validada em nenhum domínio de origem, ao contrário da biologia da qual a IA normalmente toma emprestado) não se transmite a uma hipótese que se sustenta em sua própria evidência; eleva o ônus sobre essa evidência, que as predições da Seção 6 foram construídas para carregar. Este é o padrão pelo qual o empréstimo biológico sempre foi julgado, aplicado a uma fonte que conquista menos de antemão. As predições da Seção 6 são o teste; se elas falham, o programa falha, e nenhum apelo à fonte o resgata.

"É apenas metáfora." A metáfora é onde o programa começa, não onde ele repousa. Uma metáfora que não gerasse teste algum seria, de fato, ociosa; a disciplina aqui é que se exige que cada heurística produza uma predição falsificável com um controle e uma condição de refutação, todas enunciadas na Seção 6 e resumidas na Seção 4.8. O trabalho da metáfora termina assim que a hipótese está sobre a mesa.

"As ideias já existem." Em parte verdade, e concedido por completo (Seção 5). Aprendizado por currículo, corpora sintéticos, memória persistente, consolidação, treinamento de caráter e grokking, todos existem. A contribuição residual é o conjunto exato de contrastes causais, o grafo de dependência proposto, e o uso documentado prospectivamente da fonte para gerar novas hipóteses. Uma conjunção sozinha é uma novidade fraca a menos que preveja algo que os componentes não preveem.

"Isto antropomorfiza a IA." As traduções funcionais e o firewall (Seção 3) são a resposta. Todo termo é resgatado como uma propriedade mensurável (acurácia de tomada de perspectiva, posse de compromissos, deriva, taxa de bajulação) e nenhuma alegação é feita sobre a presença fenomenal. O programa trata da competência e da individuação como fatos funcionais, e é deliberadamente silencioso sobre se algo é sentido.

"Por que esta fonte, entre todas as fontes?" Porque ela propõe uma ordem explícita e uma estrutura de dependência que podem ser traduzidas em contrastes. Essa resposta é provisória: a abundância retrospectiva pode fazer qualquer fonte rica parecer produtiva. O protocolo de proveniência da Seção 2 é o que permite que futuras hipóteses mostrem que esta fonte as gerou antes de o trabalho técnico vizinho ser consultado.

8. Conclusão

O machine learning já combina escala com arquitetura, estado, ferramentas, feedback e currículos. Este artigo, portanto, não opõe o desenvolvimento ao escalonamento. Ele pergunta se seis variáveis propostas acrescentam valor explicativo ou de engenharia depois de igualados fortes fatores convencionais. As propostas formam um piso independente mais uma espinha de dependência mais exposta; algumas são capacidades do aprendiz, algumas propriedades de currículo, e uma é uma operação intergeracional.

As predições não custam o mesmo. A continuidade relacional e a consolidação de ciclo de vida exigem trajetórias longas ou o treinamento de sucessores. A supervisão de perspectiva e a ordem de compromissos são pilotos mais baratos. O estudo de ordem mais informativo e precoce compara os braços compromissos-primeiro, reverso, intercalado e política-primeiro; o estudo de perspectiva

mais informativo cruza registro com limitação, acrescenta a interação como controle positivo, e avalia a transferência sem prosa em primeira pessoa. Um resultado nulo em qualquer dos ramos estreita o programa antes de um investimento maior.

Não reivindicamos verdade alguma para a fonte. Nem experimentos de componente bem-sucedidos a validam retrospectivamente. A fonte conquista respaldo heurístico apenas por meio de um rendimento documentado prospectivamente: uma regra de tradução produz uma hipótese antes da busca confirmatória, a intervenção sobrevive a baselines fortes, e o resultado generaliza. Se os efeitos de componente falham, o piso é estéril; se eles têm sucesso, mas a ordem falha, a espinha cai; se ambos têm sucesso, o programa encontrou uma estrutura causal útil sem converter evidência de engenharia em evidência metafísica. Isso é o que significa emprestar sem professar.

Agradecimentos e declarações

Escrita assistida por IA. A tese e todas as ideias substantivas deste artigo são de autoria do próprio autor. Ferramentas de grandes modelos de linguagem foram usadas, sob a direção e a revisão contínua do autor, para auxiliar na redação, na edição e na tradução; o autor revisou e assume total responsabilidade pelo texto final.

Referências

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/