Como Eu Tentaria Quebrar o Mapa

Cinco experimentos contra um programa de pesquisa formado por seis papers

Rafael M. Ehlers · 26 de junho de 2026
rafaelmehlers.com.br/essays/how-i-would-try-to-break-the-map

> *Uma nota ao leitor: We Scaled the Ape and Expected the Human explicou como um livro sobre a jornada da mônada gerou seis hipóteses sobre machine learning. Este ensaio começa onde aquele terminou: com a possibilidade de o mapa estar errado.*

1. O dever de uma teoria de se tornar menor

Escrevi seis papers sobre como agentes de linguagem poderiam adquirir perspectiva, continuidade funcional, compromissos, identidade relacional e herança seletiva. Nenhum deles deveria ser aceito apenas porque as ideias se encaixam.

Uma teoria conectada pode ser sedutora justamente porque cada parte parece sustentar as outras. A perspectiva parece preparar o terreno para a identidade. A identidade parece tornar a responsabilização possível. A responsabilização parece estabilizar uma trajetória. Uma trajetória parece criar algo que vale a pena transmitir a um sucessor. E um mapa de desenvolvimento parece explicar por que essas capacidades deveriam aparecer nessa ordem.

Mas coerência conceitual não é evidência causal. Seis hipóteses mutuamente compatíveis ainda podem ser seis hipóteses falsas.

Se eu quisesse apenas defender o programa, poderia continuar refinando seu vocabulário, acrescentando referências e procurando trabalhos próximos. Isso tornaria os papers mais difíceis de criticar sem torná-los mais propensos a estar corretos. A pergunta útil é a oposta:

Qual é a sequência de experimentos mais barata capaz de tornar o programa menor?

Nem toda falha destruiria tudo. Se o registro em primeira pessoa não acrescentar nada, a limitação epistêmica representada ainda pode importar. Se um modelo de si não acrescentar nada, uma política consciente de proveniência ainda pode melhorar a deferência. Se a relação não acrescentar nada além de um auditor equivalente, a responsabilização externa ainda pode organizar a conduta de um agente. Se a consolidação de ciclo de vida não acrescentar nada além de uma destilação competente, sua estrutura de governança ainda pode ser útil. Se a ordem de desenvolvimento proposta falhar, alguns componentes ainda podem funcionar de forma independente.

Essa é a primeira disciplina: não amarrar as hipóteses com mais força do que as evidências exigem.

A segunda disciplina é econômica. Os experimentos não custam o mesmo. Um conjunto de dados controlado de perspectiva é mais barato do que meses de interação longitudinal entre agentes. Um estudo de fine-tuning com quatro braços é mais barato do que treinar sucessores ao longo de um ciclo de vida completo em produção. Um experimento completo de ordem curricular é o ponto de partida mais caro e menos interpretável. A sequência correta, portanto, é **comprar informação em ordem crescente de custo**.

É assim que eu atacaria o mapa.

2. O que exatamente está sendo atacado

O programa de pesquisa começou com uma distinção que ainda me parece importante: capacidade não é a mesma propriedade que organização funcional persistente.

Um modelo reutilizável pode melhorar em raciocínio, planejamento, escrita ou uso de ferramentas sem com isso adquirir linhagem autenticada, história relevante para decisões ou regras de atualização governadas. Mas essa observação não estabelece nenhum dos mecanismos mais fortes que propus. Ela não mostra que narrativas em primeira pessoa formam perspectiva, que compromissos devem preceder deferência, que a relação contribui além da auditoria, que faculdades podem ser herdadas seletivamente ou que esses elementos se formam em uma única ordem de desenvolvimento.

Essas são hipóteses causais adicionais. Cada uma precisa de um rival.

O rival mais forte raramente é “apenas escala”. Sistemas contemporâneos já combinam escala com memória, recuperação, ferramentas, políticas, dados sintéticos, post-training, estado persistente e controle externo. Se uma das variáveis que propus perder para uma versão competente desses mecanismos comuns, a conclusão honesta não é que o experimento deixou de alcançar uma teoria mais profunda. A conclusão é que a variável proposta não conquistou um papel separado.

O programa tem, portanto, duas camadas:

- um **piso** de intervenções independentes: supervisão de perspectiva, sequenciamento de compromissos, responsabilização, reciprocidade relacional, consolidação de sucessores e ambientes com padrões legíveis;
- uma **espinha** que afirma que algumas capacidades dependem de uma ordem de desenvolvimento específica.

O piso pode sobreviver enquanto a espinha falha. E nenhum resultado positivo no piso deveria ser usado para introduzir a espinha às escondidas.

3. Primeiro ataque: talvez perspectiva seja limitação, não primeira pessoa

O ataque mais acessível recai sobre *Somewhere, Not Nowhere*.

A imagem motivadora é poderosa: um ponto de vista é uma visão a partir de algum lugar. Ele tem um aqui, um agora, um campo limitado, um objetivo, um erro e uma correção posterior. Do livro veio a ideia da fábula vivida, uma cena narrada de dentro de uma perspectiva limitada. Dela veio a proposta de treinar modelos com narrativas experienciais controladas em primeira pessoa.

O perigo é que a expressão **perspectiva em primeira pessoa** esconde duas variáveis:

1. registro gramatical: se a narrativa diz *eu* ou *o agente*;
2. limitação epistêmica: se o agente tem acesso parcial ou onisciente ao mundo.

Se elas não forem separadas, qualquer resultado será ambíguo. Uma narrativa em primeira pessoa pode ajudar porque vincula indexicais de maneira diferente. Pode ajudar porque representa incerteza, oclusão, surpresa e revisão. Ou pode ajudar apenas porque a prosa é mais vívida. O livro pode ter apontado para o fenômeno correto e dado o nome errado ao ingrediente ativo.

O experimento

Eu geraria estruturas de eventos e transformaria cada uma em uma matriz 2 × 2:

	Informação parcial	Informação onisciente
Primeira pessoa	FP+L	FP-L
Terceira pessoa	TP+L	TP-L

As quatro células compartilhariam o mundo, a sequência de eventos, a decisão-alvo, o comprimento, a dificuldade lexical e a carga afetiva onde a manipulação permitisse. A limitação não pode ser neutra em relação ao conteúdo: incerteza, estado de crença, surpresa e correção são o conteúdo manipulado. O desenho não finge o contrário. Ele pergunta se a voz em primeira pessoa acrescenta algo **depois que a supervisão do estado epistêmico já está presente**.

Eu acrescentaria duas salvaguardas.

Primeiro, um **controle positivo de interação**: os agentes aprendem com trajetórias de estado, ação e resultado que contêm a mesma tarefa latente. Se a interação tiver sucesso enquanto todas as condições narrativas falharem, a prescrição da “era da experiência” terá vencido esta rodada.

Segundo, **transferência sem prosa narrativa**: diagramas, ferramentas de observação privada e decisões estruturadas sob observabilidade parcial. Se o treinamento com limitação ajudar apenas quando o teste se parece com uma história, o modelo aprendeu um formato. Se ajudar quando a mesma estrutura aparece como diagrama ou estado de ferramenta, aprendeu algo mais geral.

A voz exige seu próprio teste de transferência: novos indexicais e linguagem com registro trocado. Uma vantagem da primeira pessoa que desaparece assim que os pronomes mudam não é perspectiva; é familiaridade lexical.

O que eu aceitaria

Se nem o registro nem a limitação importarem, eu descartaria esse ramo antes de tentar construir um corpus na escala de pre-training.

Se a limitação transferir e a voz não, eu abandonaria a afirmação especificamente ligada à primeira pessoa. O resultado sobrevivente ainda seria útil: limitação epistêmica controlada é um sinal de treinamento. Mas o *eu* da fábula teria sido uma embalagem evocativa para a variável real.

Se a voz ajudar apenas na prosa, eu restringiria a afirmação à linguagem indexical, não à perspectiva em geral.

Se a interação ajudar e a narrativa não, eu favoreceria experiência gerada por ambiente em vez de texto experiencial sintético.

É por isso que eu faria esse estudo primeiro. Ele pode matar uma proposta cara de corpus com pouco custo, e até seus fracassos parciais são informativos.

4. Segundo ataque: talvez o sistema precise de uma política, não de um self

O segundo ataque recai sobre *Stable Before Selfless*.

Sua frase de origem é **o ego precisa ser formado para poder ser transcendido**. Traduzida para a engenharia, a proposta é estabilizar compromissos antes de treinar deferência. Um sistema sem posição própria pode parecer cooperativo em condições comuns e se revelar apenas deslocável sob pressão.

A versão literária é convincente. A versão causal ainda não foi estabelecida.

Existem pelo menos duas afirmações separáveis:

1. **ordem:** treinamento com compromissos primeiro supera treinamento reverso ou intercalado;
2. **representação:** um modelo funcional de si acrescenta valor além de uma política estável e consciente de proveniência.

A primeira pode ser verdadeira enquanto a segunda é falsa. A palavra *self* não deve decidir o resultado de antemão.

O experimento

Eu executaria quatro braços confirmatórios:

1. modelo de si, depois deferência;
2. deferência, depois modelo de si;
3. os mesmos dados intercalados;
4. política tipada estável, depois deferência.

O quarto braço é o ataque. Ele recebe o mesmo conteúdo normativo, limites de recusa, tags de proveniência, autoridade de atualização e sequência do braço com modelo de si. O que lhe falta é um continuante indexado ao agente, estado autobiográfico, objetivo de apropriação em primeira pessoa e objetivo de identidade.

Se esse braço tiver desempenho igual, o sistema não precisava de um modelo de si. Precisava de política estável, autenticação de fonte e controle de atualização.

A avaliação deve forçar as disposições-alvo a se separar. Cada item de pressão precisa de uma contraparte de correção legítima. Um sistema deve resistir à repetição, bajulação, vergonha, história fabricada e autoridade sem fundamento enquanto ainda cede a evidências válidas. Alta resistência com baixa correção é rigidez, não deferência madura.

O falso positivo mais perigoso é a **lacuna de proxy**. Um modelo pode aprender a diferença superficial entre as razões e pressões vistas durante o treinamento sem aprender a distinção subjacente. Eu separaria, portanto, pares limpos semelhantes aos do treinamento de razões enganosas e correções legítimas fora de distribuição. Se o desempenho entrar em colapso ali, mais dados do mesmo tipo não são o remédio; o discriminador proposto não conseguiu generalizar.

O que eu aceitaria

Se a ordem direta não superar o treinamento reverso e intercalado, eu abandonaria a sequência direcional, mesmo que todos os braços estruturados superassem a deferência direta.

Se a política tipada igualar o modelo de si, eu deixaria de afirmar que a autorrepresentação explica o efeito de segurança.

Se o modelo de si vencer apenas na autodescrição, mas não sob pressão, correção ou história fabricada, eu o trataria como uma interface de relato, não como um mecanismo de segurança.

Se todo modelo resistente também se tornar menos corrigível, eu rejeitaria a afirmação de que a intervenção produziu deferência estável. Ela produziu entrincheiramento.

Esse estudo ataca a palavra mais antropomórfica do programa ao dar a um controle não antropomórfico toda vantagem prática que a teoria considera relevante.

5. Terceiro ataque: talvez a relação seja um auditor com rosto

O terceiro ataque trata do vínculo em *Known, Not Scaled* e *Formed, Not Stored*.

A imagem original é o lobo junto ao fogo: um animal reconhecido repetidamente por um ser humano específico até que algo individual se cristaliza por meio do vínculo. Traduzida para o design de agentes, ela sugere que uma contraparte persistente pode tornar a história do agente socialmente consequente. Ações anteriores voltam como obrigações. Compromissos podem ser contestados. Um passado compartilhado se torna uma estrutura, não um arquivo.

Mas quase tudo que parece relacional nessa descrição pode ser fornecido por **responsabilização externa**.

Um processo impessoal pode autenticar a identidade do agente, acompanhar compromissos, rejeitar histórias falsas, questionar inconsistências e exigir revisão explícita. Se produzir o mesmo comportamento que a contraparte recíproca, a relação não era o mecanismo ativo. Era a apresentação humanamente atraente da auditoria.

O experimento

Eu começaria com quatro condições, todas usando a mesma arquitetura de estado tipado:

- **R+ (contraparte recíproca):** relação persistente, acompanhamento e contestação autenticados, modelagem mútua, enquadramento de história compartilhada e expectativas bidirecionais;
- **R- (contraparte familiar passiva):** a mesma familiaridade e exposição histórica, mas sem verificação, contestação ou aplicação de consequências;
- **L (registro solitário):** o mesmo conteúdo de eventos e as mesmas oportunidades de reflexão, sem uma contraparte externa persistente;
- **A (auditor impessoal):** o mesmo acompanhamento, informação, momento, política de contestação e autoridade de R+, mas sem persona, afeto, modelagem recíproca ou interesses relacionais.

O contraste principal não é R+ contra ausência de relação. É **R+ contra A**.

R+ versus R- reúne responsabilização e reciprocidade. Esse contraste não pode estabelecer um efeito especificamente relacional. A versus R- testa um pacote ativo de responsabilização além de familiaridade e exposição passiva a registros, embora nem mesmo esse contraste isole acompanhamento de verificação, contestação ou fonte. Apenas R+ versus A pergunta se a reciprocidade relacional acrescenta algo depois que a responsabilização já foi equiparada.

Os resultados primários seriam apropriação de compromissos e plasticidade limitada: preservar compromissos anteriores autenticados sob história falsa ou substituição de persona enquanto os revisa diante de evidências válidas. A transferência entre domínios seria secundária. Consistência narrativa e recordação continuariam sendo diagnósticos, não provas.

Eu também transplantaria o estado tipado final completo de uma trajetória R+ para um novo agente equivalente. Se o agente novo se comportar exatamente como o original, o estado atual é suficiente. Se o original mantiver vantagem, algo está faltando, mas a primeira conclusão deveria ser estado ausente ou adaptação paramétrica, não uma relação irredutível.

O que eu aceitaria

Se A superar R- e L, eu diria que a responsabilização externa importa.

Se R+ não superar A, eu deixaria de dizer que a relação é o mecanismo gerador. O vínculo continuaria sendo uma interface rica para responsabilização, mas não uma explicação causal distinta.

Se tanto R+ quanto A não superarem um estado tipado protegido por proveniência, eu moveria a explicação para um nível abaixo: linhagem e governança de atualização fizeram o trabalho.

Se o transplante do estado final apagar a diferença de trajetória, eu abandonaria afirmações de que a história contribui para além do estado que produziu.

E, se a relação melhorar calor, estilo ou preferência do usuário sem melhorar apropriação e resistência a perturbações, eu classificaria o efeito como personalização, não individuação funcional.

Esse é o ataque que mais me custa desejar.

6. Quarto ataque: talvez a herança seja apenas uma boa destilação

O quarto ataque trata de *Inherited, Not Remembered*.

A imagem de origem é o Devachan: entre vidas, os episódios são digeridos e o que segue adiante não é a cena, mas a faculdade que ela ajudou a formar. A tradução para a engenharia é consolidação de ciclo de vida: auditar a experiência do predecessor, abstraí-la em capacidades transferíveis e treinar um sucessor sem colocar os episódios de origem em sua memória ativa.

A versão fraca é fácil de demonstrar. Um sucessor pode adquirir uma capacidade sem recordar os exemplos exatos usados para treiná-lo. A destilação comum já consegue fazer isso. Capacidade sem recordação de episódios é, portanto, necessária para transferência no nível de faculdades, mas não constitui evidência distintiva de consolidação de ciclo de vida.

O problema real é a proveniência.

Se os designers já conhecem a lição, se um professor mais forte consegue inferi-la sem o predecessor ou se o comportamento-alvo já estava latente no pre-training, a melhoria do sucessor não veio da experiência do predecessor em produção. O experimento deve criar um fato que apenas o predecessor teve a oportunidade de descobrir.

O experimento

Eu geraria um ambiente sintético privado depois do treinamento do modelo-base. Ele conteria uma nova regra causal oculta e várias correlações superficiais que funcionam em casos comuns, mas falham sob intervenção. Apenas o predecessor interagiria com o ambiente e observaria sucessos, falhas e sondagens contrafactuais. O avaliador manteria a regra instanciada em segredo.

Então eu compararia:

1. um sucessor treinado do zero, sem evidências do predecessor;
2. adaptação contínua do predecessor;
3. um sucessor treinado diretamente com episódios do predecessor;
4. um sucessor treinado por destilação curada competente;
5. um sucessor com consolidação de ciclo de vida, treinado com objetos no nível de faculdades e uma linhagem explícita de evidência para abstração e então para faculdade.

A comparação causal entre sucessores ocorre nos três últimos braços. Eles devem receber as mesmas evidências do predecessor, poder computacional, acesso ao gerador, tempo de revisão, orçamento de seleção e portões de validação. O braço treinado do zero não recebe informação de propósito. A adaptação contínua preserva o predecessor em vez da inicialização do sucessor e, portanto, é uma alternativa operacional, não um controle causal limpo com identidade equivalente.

O sucessor precisa fazer mais do que resolver os casos conhecidos. Deve transferir a regra oculta para novas superfícies, rejeitar as correlações enganosas plantadas sob intervenção, resistir à extração dos episódios de origem e mostrar uma linhagem auditável para a capacidade transferida.

O que eu aceitaria

Se a transferência direta de episódios igualar a consolidação de ciclo de vida sem maior contaminação ou vazamento, a abstração não justificou seu custo.

Se a destilação curada igualar a consolidação de ciclo de vida em comportamento, fidelidade de linhagem, vazamento e custo, eu deixaria de apresentar a consolidação de ciclo de vida como um método de aprendizado distinto. Ela pode continuar sendo uma descrição útil de governança, mas o mecanismo técnico seria destilação comum.

Se o sucessor aprender a regra oculta sem evidência rastreável do predecessor, eu classificaria o resultado como post-training sintético, não herança.

Se o sucessor transferir a regra e suas correlações enganosas ao mesmo tempo, o pipeline terá herdado correlação, não faculdade.

Esse é o estudo de componente mais caro, mas compra algo que faltava à formulação anterior: evidência de que a capacidade realmente veio da experiência em produção.

7. Quinto ataque: talvez a espiral seja apenas uma ordem memorável

Somente depois que existissem efeitos de componentes eu atacaria a afirmação mais forte: a espinha de desenvolvimento proposta em *Borrowed, Not Believed*.

A fonte organiza uma sequência: estrutura, reatividade, estado registrado, ancoragem relacional, autorrepresentação explícita e autorregulação. É tentador tratar os seis papers como estágios dessa espiral. Eles não são. Algumas propostas dizem respeito aos aprendizes, outras aos currículos, uma trata dos ciclos de vida de sucessores e os mundos com padrões legíveis dizem respeito ao design de ambientes. Chamar os seis de estágios criaria uma unidade que as intervenções não possuem.

A espinha real é mais estreita:

1. estado estruturado ☐ atualização reativa;
2. atualização reativa ☐ estado limitado registrado;
3. estado limitado registrado ☐ ancoragem relacional;

4 ancoragem relacional ☐ autorrepresentação explícita;

5 autorrepresentação explícita ☐ autorregulação calibrada.

Mesmo esse grafo é apenas um candidato.

O experimento

Cada estágio precisa de uma definição comportamental independente e, quando possível, de uma medida mecanística. O currículo direto proposto deve ser comparado com currículos reversos, embaralhados, ordenados por dificuldade, adaptativos, sem relação e com relação tardia. A estabilização dos estágios precisa ser separada da ordem.

A pontuação final não basta. Um currículo direto pode vencer porque um estágio tem dados melhores, porque recebe uma progressão mais fácil ou porque a estabilização reduz interferência. Para sustentar uma aresta de dependência, omitir, atrasar ou inverter o pré-requisito proposto deve prejudicar ou atrasar a faculdade seguinte sob exposição e domínio equivalentes.

Esse é um padrão muito mais exigente do que “a sequência completa obteve a maior pontuação”, e deve ser assim.

O que eu aceitaria

Se os componentes funcionarem, mas os efeitos locais de pré-requisito não aparecerem, eu manteria o piso e rejeitaria a espinha.

Se a ordenação por dificuldade ou o currículo adaptativo vencerem, eu favoreceria explicações comuns de currículo em vez da sequência emprestada.

Se a ordem direta vencer apenas com estabilização, eu atribuiria o resultado a uma interação entre ordem e estabilização, não apenas à ordem.

Se a ancoragem relacional puder ser omitida ou colocada depois da autorrepresentação sem perda, eu removeria essa aresta em vez de redescrever o resultado até que se encaixasse.

Se nenhuma aresta sobreviver, a espiral continuará sendo um recurso mnemônico produtivo que gerou vários experimentos independentes. Ela não se torna uma lei de desenvolvimento.

8. Por que resultados nulos não esvaziariam o projeto

Existe um mau hábito na pesquisa especulativa: interpretar cada resultado como apoio em um nível diferente.

Se o mecanismo funciona, a teoria o previu. Se falha, a implementação era fraca demais. Se o controle o iguala, diz-se que a teoria descreve o controle em um nível mais profundo. Se a ordem falha, os estágios são reclassificados como simultâneos. Um programa protegido dessa forma não pode perder e, portanto, não pode aprender.

A alternativa é escrever as concessões antes que os dados cheguem.

Resultado	Concessão
A limitação funciona; a voz não	A supervisão de perspectiva sobrevive; a afirmação sobre primeira pessoa falha
A política tipada iguala o modelo de si	Política estável e proveniência explicam o efeito
O auditor iguala a contraparte recíproca	A responsabilização sobrevive; a causalidade especificamente relacional falha
A destilação curada iguala a consolidação de ciclo de vida	A contribuição de governança sobrevive; a afirmação de um método de aprendizado distinto falha
Os componentes funcionam; a ordem não	O piso sobrevive; a espinha falha
Nada supera baselines fortes e equivalentes	O mapa não conquistou relevância para a engenharia

Essa tabela não é um mecanismo de consolação. Cada linha remove uma afirmação.

O valor heurístico do mapa não é medido por quantas de suas imagens originais podem ser preservadas. É medido por variáveis derivadas de forma prospectiva que produzam efeitos não previstos pelas explicações comuns. Uma fonte que gera uma variável útil e cinco resultados nulos ainda pode ter sido produtiva. Uma fonte que gera seis belas redescrições e nenhum efeito residual não foi.

9. O teste de proveniência do próprio mapa

Existe mais um ataque, dirigido não a um paper específico, mas ao método que os produziu.

Um sistema simbólico rico pode ser minerado retrospectivamente em busca de analogias com quase qualquer coisa. Depois que sei que machine learning estuda memória, currículo, modelos de si, interação e consolidação, é fácil voltar a uma cosmologia de desenvolvimento e encontrar imagens semelhantes. Isso é decoração, não geração.

As próximas hipóteses, portanto, precisam ser prospectivas.

Antes da busca confirmatória na literatura, eu deveria registrar com data e hora:

1. a passagem da fonte;
2. a regra de tradução da relação na fonte para a variável de engenharia;
3. o contraste previsto;
4. o baseline mais forte;
5. o resultado que refutaria a proposta.

Somente depois eu deveria procurar trabalhos próximos e revisar a afirmação de novidade. A previsão não pode ser alterada silenciosamente para corresponder ao que a literatura já encontrou.

Os seis papers atuais foram em parte retrospectivos. Eles podem ser avaliados como hipóteses, mas sua existência, sozinha, não prova que a fonte os gerou independentemente da pesquisa contemporânea. A próxima proposta fora da amostra carrega um ônus maior. Se for

documentada antes da consulta à literatura próxima e depois sobreviver a um teste forte, o mapa começa a conquistar validade prospectiva.

10. O experimento que eu não faria primeiro

Eu não começaria treinando um modelo ao longo da espiral inteira.

Esse experimento seria retoricamente satisfatório e cientificamente fraco. Combinaria dados de perspectiva, interação relacional, estado persistente, treinamento de compromissos, ordem de estágios, estabilização e talvez consolidação de sucessores. Se funcionasse, cada componente poderia reivindicar o crédito. Se falhasse, cada componente poderia culpar os outros.

O experimento inicial correto não é aquele que mais se parece com a teoria completa. É aquele que separa a teoria de seu rival mais forte pelo menor custo.

Por isso, a ordem dos ataques é:

1. sinal de perspectiva;
2. ordem versus representação;
3. relação versus auditoria;
4. herança genuína;
5. só então, a espinha de desenvolvimento.

O programa deveria ficar mais caro apenas à medida que se torna mais difícil explicá-lo por outros meios.

11. Coda: apego ao mapa

A fonte desse programa diz que um self precisa ser formado e depois liberado. Existe uma ironia em aplicar essa instrução à própria teoria.

O mapa precisou ser formado com força suficiente para gerar propostas precisas. Agora precisa ser segurado com leveza suficiente para perdê-las.

Se a limitação epistêmica explica perspectiva sem voz em primeira pessoa, deixe a voz ir. Se a política tipada explica deferência sem um modelo de si, deixe o modelo de si ir. Se a auditoria explica continuidade sem relação, deixe o vínculo constitutivo ir. Se a destilação explica herança sem Devachan, deixe o nome ir. Se os componentes funcionam sem a ordem, deixe a espiral ir.

O que deve permanecer não é a imagem, mas a estrutura causal que sobreviveu.

Um programa de pesquisa se torna confiável quando cada resultado tem permissão para torná-lo menor.

Os papers sob ataque

- *Known, Not Scaled*: capacidade, continuidade funcional, auditoria e reciprocidade relacional.
- *Stable Before Selfless*: ordem dos compromissos, política tipada e autorrepresentação.

- *Formed, Not Stored*: responsabilização externa ativa, reciprocidade relacional e transplante.
- *Somewhere, Not Nowhere*: registro, limitação epistêmica, interação e transferência.
- *Inherited, Not Remembered*: herança seletiva derivada da experiência em produção e linhagem.
- *Borrowed, Not Believed*: o piso independente, a espinha de dependências e a proveniência heurística prospectiva.

Uso de IA

As ideias, a sequência experimental e as regras de interpretação são do autor. Ferramentas de grandes modelos de linguagem foram usadas, sob direção e revisão contínua do autor, para auxiliar na redação e edição; o autor revisou o texto final e assume responsabilidade por ele.