

Herdado, Não Lembrado: Consolidação de Ciclo de Vida para Agentes de Linguagem Sucessores

Rafael de Menezes Ehlers · ORCID [0009-0009-6476-6690](https://orcid.org/0009-0009-6476-6690)

Pesquisador independente

rafaehlers@gmail.com

Preprint. © 2026 Rafael de Menezes Ehlers. Licenciado sob [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/).

Resumo

Agentes de linguagem de longa duração acumulam cada vez mais experiência durante a implantação, e no entanto os ciclos de vida de modelos carecem de uma explicação amplamente adotada e experimentalmente distinguida de como evidências de implantação seletivamente auditadas deveriam moldar um sucessor. Este artigo propõe a **consolidação de ciclo de vida**: a evidência do predecessor é auditada, abstraída em faculdades candidatas, transferida e independentemente validada, enquanto traços específicos de episódios são excluídos da memória ativa do sucessor. O problema central de identificação é a proveniência causal. Se os projetistas ou um abstrator mais forte já conhecem a lição-alvo, a melhoria pode vir de um pós-treino sintético comum, e não de algo que o predecessor descobriu. Propomos, portanto, um ambiente sintético privado contendo uma regra causal inédita e chamarizes sensíveis a intervenção gerados após o treino do modelo base. Apenas o predecessor encontra a regra por meio de resultados de implantação. Os três braços de transferência ao sucessor — transferência direta de episódios, destilação curada e consolidação de ciclo de vida — recebem evidência do predecessor, computação, acesso ao gerador, tempo de revisor, orçamentos de seleção e portões de validação equiparados. O sucessor do zero é um controle negativo sem histórico; a adaptação contínua atualiza o predecessor e é, portanto, uma alternativa operacional em vez de um braço sucessor equiparado em informação e identidade. O resultado distintivo não é apenas capacidade sem recordação, o que a destilação também pode produzir, mas a transferência seletiva de uma regra derivada da implantação com linhagem auditável, ao mesmo tempo que se rejeitam chamarizes e detalhes dos episódios de origem. A substituição de modelos torna-se, assim, um problema testável de herança derivada de evidência, em vez de simples substituição.

Palavras-chave: agentes de linguagem · consolidação de ciclo de vida · aprendizado contínuo · sucessão de modelos · memória · destilação · inteligência artificial

1. Introdução

Modelos de linguagem costumam ser discutidos como artefatos versionados, com cada geração nomeada suplantando a anterior. Um modelo é treinado, implantado, atualizado, descontinuado e, por fim, aposentado ou preservado. Esse vocabulário é operacionalmente conveniente, mas obscurece um problema conceitual crescente. À medida que os modelos de linguagem se tornam agentes com memória de longo prazo, históricos específicos de usuários, uso de ferramentas, preferências,

compromissos e trajetórias de implantação estendidas, a substituição deixa de ser uma mera questão de substituir um artefato estático por outro mais capaz. Torna-se um problema de sucessão.

Os arcabouços de governança de IA já tratam a gestão de risco como uma atividade de ciclo de vida, incluindo o monitoramento após a implantação (Tabassi, 2023). Eles não especificam, contudo, como a experiência acumulada por um agente implantado deveria moldar seu substituto.

O que, se é que algo, um modelo sucessor deveria herdar de seu predecessor?

As respostas usuais são parciais. Uma resposta é o **ajuste fino contínuo**: continuar atualizando o mesmo modelo à medida que novos dados ou feedback chegam, aceitando riscos como o esquecimento catastrófico e a deriva comportamental (Kirkpatrick et al., 2017; Rolnick et al., 2019).

Outra é a **memória episódica**: preservar registros, resumos, repositórios vetoriais ou registros acessíveis por recuperação, como já fazem as arquiteturas de agentes aumentadas por recuperação e memória (Lewis et al., 2020; Park et al., 2023; Packer et al., 2023; Shinn et al., 2023). Benchmarks de diálogo de longo prazo demonstram, adicionalmente, tanto a relevância prática quanto as limitações atuais da memória ao longo de muitas sessões (Maharana et al., 2024). A memória episódica preserva o passado, mas por si só não o transforma em competência. Uma terceira resposta é a **destilação**, que transfere comportamento ou desempenho do professor para o aluno (Hinton et al., 2015) sem necessariamente distinguir o que um predecessor fez, lembrou ou aprendeu como uma faculdade generalizável.

Este artigo argumenta que a sucessão de modelos requer um quarto conceito: a **consolidação de ciclo de vida**.

Por consolidação de ciclo de vida entendemos uma fase explícita entre implantação e sucessão na qual a experiência acumulada de um modelo implantado é auditada, abstraída e transformada em faculdades transferíveis para um sucessor, enquanto traços específicos de episódios são descartados, arquivados ou mantidos fora da memória ativa do sucessor. O objetivo não é preservar tudo o que o modelo encontrou nem meramente comprimir o predecessor no sucessor. É distinguir **episódios** de **faculdades**: interações, casos e traços particulares das capacidades gerais que eles podem ter ajudado a formar.

A distinção importa porque agentes de longa duração acumularão mais do que dados. Eles acumularão correções, falhas, encontros adversariais, adaptações específicas de usuários, estratégias de domínio, lições de segurança, hábitos de uso de ferramentas e regularidades comportamentais. Alguns desses não deveriam ser herdados de forma alguma. Alguns deveriam permanecer auditáveis, porém inativos. Alguns deveriam ser convertidos em capacidades gerais. Tratar tudo isso como “dados para ajuste fino” colapsa distinções que uma teoria de ciclo de vida adequada deve manter separadas.

“Ciclo de vida” é usado aqui funcionalmente para nomear o arco operacional que vai do treino, passando pela implantação e adaptação, até a aposentadoria e a substituição. A proposta não faz nenhuma afirmação sobre consciência, personalidade ou continuidade pessoal. O sucessor é um novo token de agente, não o mesmo agente continuando por outra via. **Linhagem de evidência**, neste artigo, significa uma derivação auditável dos episódios do predecessor a abstrações e objetos de treino de faculdade; não é a condição autenticada de continuação de estado usada na definição, adotada em todo o programa, de individuação funcional. A consolidação de ciclo de vida pode transferir uma capacidade sem satisfazer nenhuma das outras condições de individuação — dependência de caminho de um agente contínuo, propriedade, plasticidade limitada ou clareza de ramificação. Ela trata a sucessão como um problema de projeto: se um passo de transformação

fundamentado em princípios deveria extrair faculdades generalizáveis do histórico de implantação antes que um modelo substitua outro.

O firewall fenomenal é igualmente rigoroso. Essas medidas não decidem a presença fenomenal: o sucesso não é suficiente para presença nem o fracasso suficiente para ausência. Transferência de capacidade, vazamento episódico e linhagem de evidência permanecem resultados de engenharia, independentemente do status fenomenal de qualquer um dos sistemas.

O artigo faz quatro contribuições. Distingue a consolidação de ciclo de vida do ajuste fino contínuo, da transferência episódica e da destilação padrão; define episódios e faculdades como objetos distintos de herança; propõe uma arquitetura de referência de seis estágios; e deriva previsões experimentais. O resultado característico seria a dissociação: um sucessor preserva uma habilidade adquirida ao mesmo tempo que falha, por projeto arquitetônico, em recordar seus episódios de origem.

As Seções 2 a 4 definem o problema da sucessão, a distinção episódio/faculdade e métodos vizinhos. As Seções 5 a 7 apresentam a arquitetura, as previsões experimentais e as implicações de governança; a Seção 8 conclui.

2. O problema da sucessão

A substituição de modelos costuma ser tratada como um problema de capacidade e compatibilidade. Um novo modelo é lançado porque é mais capaz, mais barato, mais seguro, mais rápido ou mais bem alinhado com requisitos de produto. O modelo mais antigo é descontinuado, contornado por roteamento, preservado para reprodutibilidade ou mantido disponível para usuários que ainda dependem dele. Dessa perspectiva, as questões centrais são práticas: como migrar usuários, por quanto tempo manter o acesso, como preservar a retrocompatibilidade e como documentar diferenças comportamentais.

Esse enquadramento é adequado quando os modelos são tratados como ferramentas sem estado. Torna-se incompleto quando os modelos se tornam agentes de horizonte longo.

Uma ferramenta sem estado (stateless) pode ser substituída por uma melhor com pouco resíduo conceitual. Se o novo sistema executa as mesmas tarefas melhor, a substituição foi bem-sucedida. Um agente de horizonte longo, contudo, não é meramente um mapeamento de prompts para saídas. Ele pode acumular memórias, interagir com usuários ao longo de sessões, adaptar-se a domínios específicos, adquirir hábitos de uso de ferramentas, receber correções, encontrar situações adversariais e desenvolver padrões comportamentais estáveis. Em tais sistemas, a implantação não é mera distribuição; ela gera experiência. A substituição levanta então uma questão que o versionamento comum não responde: o que deveria acontecer com o que o predecessor aprendeu durante a implantação?

Um agente de longa duração acumula experiência que é valiosa demais para descartar, ruidosa demais para herdar por completo e consequente demais para incorporar cegamente em seus pesos. A atualização contínua arrisca deriva e contaminação; a memória externa preserva traços sem necessariamente produzir competência; e a destilação indiscriminada transfere comportamento sem decidir quais capacidades adquiridas na implantação merecem preservação. O problema da sucessão não é resolvido, portanto, escolhendo um desses mecanismos, mas especificando uma operação intermediária que governe a herança.

Este artigo chama essa operação de **consolidação de ciclo de vida**. O termo pretende enfatizar três características.

Primeiro, a consolidação é seletiva. Nem tudo o que aconteceu a um modelo implantado deveria ser passado adiante. Algumas interações são anômalas, adversariais, privadas, de baixa qualidade, específicas de usuários ou eticamente inapropriadas para generalizar. Uma fase de consolidação deve, portanto, incluir filtragem e auditoria, não mera acumulação.

Segundo, a consolidação é abstrativa. O alvo da transferência não é o episódio em si, mas o padrão extraído dele. Um modelo pode encontrar milhares de correções de usuários, mas o sucessor não deveria herdá-las como uma coleção de casos lembrados. Ele deveria herdar melhor reconhecimento de erros, melhor calibração de incerteza ou maior integridade de recusa, se a evidência sustentar essas capacidades gerais.

Terceiro, a consolidação é orientada ao sucessor. O objetivo não é preservar o predecessor tal como era nem reproduzi-lo exatamente. É preparar um sucessor que se beneficie do que o predecessor aprendeu sem depender desnecessariamente do histórico particular do predecessor. Nesse sentido, a consolidação de ciclo de vida difere tanto do arquivamento quanto da imitação. Ela transforma a experiência selecionada em capacidades futuras, em vez de armazenar um modelo passado.

A distinção pode ser posta da seguinte forma. O histórico de implantação de um modelo contém **episódios**: interações, falhas, correções, decisões, encontros adversariais e casos específicos de domínio particulares. Alguns episódios sustentam **abstrações**: padrões ou lições que se repetem em vários casos. Algumas abstrações tornam-se **faculdades**: capacidades gerais que podem operar em novos contextos sem recordar os casos originais que as formaram. A consolidação de ciclo de vida é o processo pelo qual episódios selecionados são convertidos em abstrações e, em seguida, em faculdades transferíveis para um sucessor.

Isso fornece uma resposta fundamentada em princípios à questão da herança. Um sucessor não deveria herdar a vida inteira do predecessor. Ele deveria herdar aquelas faculdades que a vida do predecessor tornou disponíveis para consolidação.

Entre um modelo implantado e seu sucessor pode existir, assim, uma fase estruturada: auditar o que aconteceu, identificar o que foi aprendido, abstrair capacidades transferíveis e testá-las independentemente de seus episódios de origem. O predecessor pode permanecer arquivado para reprodutibilidade ou análise de segurança, mas o sucessor leva adiante aquilo para que essas memórias servissem, em vez das memórias em si.

A próxima seção formaliza a distinção episódio/faculdade da qual esta proposta depende.

3. Episódios, abstrações e faculdades

A consolidação de ciclo de vida depende de uma distinção que o vocabulário atual de desenvolvimento de modelos frequentemente embaralha: a distinção entre um episódio, uma abstração e uma faculdade.

Um **episódio** é um traço particular de experiência. Em um agente de linguagem implantado, os episódios incluem conversas, chamadas de ferramentas, correções, falhas, recusas, encontros adversariais, feedback de usuários, casos específicos de domínio e decisões tomadas sob restrições particulares. Os episódios são indexados a tempo, contexto, participantes e circunstâncias. Eles registram o que aconteceu.

Uma **abstração** é um padrão extraído de um ou mais episódios. Ela remove alguns detalhes específicos do episódio enquanto preserva uma relação, regularidade, modo de falha ou estratégia que pode recorrer em outro lugar. Correções repetidas de usuários, por exemplo, podem indicar que o agente superestima a confiança quando a recuperação está incompleta. Conversas adversariais podem mostrar que elogio, urgência ou vergonha corroem os limites de recusa. Falhas repetidas de uso de ferramentas podem indicar que uma ação requer verificação antes da execução.

Uma **faculdade**, tal como o termo é usado aqui, é uma capacidade generalizável moldada pela experiência, mas não mais dependente de recordar os episódios a partir dos quais foi abstraída. Uma faculdade não é meramente a proposição armazenada de que um padrão existe; é a habilidade de agir de forma diferente porque o padrão foi internalizado. Melhor calibração de incerteza, maior integridade de recusa, uso de ferramentas em múltiplas etapas mais robusto, detecção mais forte de prompts manipulativos e assimilação de correções mais confiável são exemplos de faculdades. Uma faculdade é o que resta quando uma experiência foi convertida em uma capacidade.

A distinção importa porque a herança de modelos pode ter cada nível como alvo de forma diferente. Se um sucessor herda **episódios**, ele recebe traços: registros, exemplos, resumos, memórias ou registros acessíveis por recuperação. Isso pode ser útil para auditoria, personalização ou continuidade factual, mas arrisca levar adiante ruído, detalhes sensíveis à privacidade, contaminação adversarial e adaptações vinculadas ao contexto. A transferência de episódios é de alta fidelidade, mas de baixa abstração.

Se um sucessor herda **abstrações**, ele recebe padrões: lições extraídas da experiência e desvinculadas de parte de seu contexto original. Isso é mais seguro e mais geral do que a transferência bruta de episódios, mas pode permanecer meramente declarativo. Um modelo pode ser instruído a resistir à adulação, verificar chamadas de ferramentas incertas ou evitar respostas excessivamente confiantes sem adquirir a disposição correspondente.

Por exemplo, uma abstração pode ser uma instrução de sistema que diz ao sucessor para rejeitar correções falsas de usuários. Uma faculdade existe apenas quando a otimização de preferências ou outra intervenção faz o sucessor resistir a correções contraditórias mesmo sem essa instrução. A primeira enuncia uma lição; a segunda muda uma disposição.

Se um sucessor herda **faculdades**, ele recebe competência alterada: uma capacidade de se comportar de forma diferente em casos novos porque a experiência do predecessor foi consolidada em sinais de treino, estruturas de política, pressão de avaliação ou inicialização. A transferência de faculdades não é, portanto, nem cópia de memória nem mero resumo. Ela converte o histórico de implantação em capacidade reutilizável.

Isso produz a afirmação normativa central da consolidação de ciclo de vida:

Um sucessor não deveria herdar tudo o que o predecessor vivenciou. Ele deveria herdar aquelas faculdades que a experiência do predecessor tornou possíveis.

A afirmação tem dois lados. Primeiro, boa parte do histórico de um predecessor não deveria ser herdada. Alguns episódios são privados, acidentais, anômalos, maliciosos, enganosos, obsoletos ou específicos demais para um usuário ou domínio. Passá-los adiante criaria contaminação em vez de aprendizado. Segundo, parte desse histórico não deveria ser perdida. Por meio da interação, um modelo implantado pode revelar modos de falha e estratégias indisponíveis durante o pré-treino. Descartar o histórico inteiro na substituição desperdiça uma fonte de conhecimento prático. A

consolidação de ciclo de vida separa esses dois requisitos: arquivar episódios quando necessário, abstrair padrões quando possível e transferir faculdades apenas quando justificado.

Um modelo pode recuperar uma falha anterior e ainda assim falhar da mesma maneira estrutural; inversamente, ele pode adquirir uma faculdade sem recordar o caso de origem. Essa dissociação fornece a assinatura empírica da consolidação. Se o desempenho depende de recuperar ou reproduzir casos de origem, o sistema herdou registros em vez de uma faculdade.

Operacionalmente, a distinção requer seleção de episódios, abstração de padrões, construção de faculdades e validação independente. Essas transformações previnem tanto a herança indiscriminada quanto a amnésia completa: o que é passado adiante é a parte do passado que sobrevive à abstração e à validação como capacidade. A Seção 5 desenvolve essas funções em uma arquitetura.

Usamos **faculdade** em vez de **capacidade** para marcar uma capacidade cuja relevância analítica reside em sua derivação da experiência prévia e em sua portabilidade através de fronteiras entre sucessores. “Habilidade”, “comportamento”, “política” e “competência” podem descrever o que um sistema é capaz de fazer, mas não identificam por si só esse papel histórico e intergeracional. Uma faculdade é, assim, uma capacidade com uma história: geral o suficiente para operar além dos casos que a formaram, mas não pressuposta como surgindo apenas do pré-treino. É a experiência tornada portátil.

O restante do artigo usa essa distinção para separar a consolidação de ciclo de vida de seus vizinhos. O aprendizado contínuo modifica o mesmo modelo ao longo do tempo. A recuperação preserva episódios fora do modelo. A edição de modelos altera associações factuais ou comportamentos direcionados. A destilação transfere saídas ou políticas. A consolidação de ciclo de vida, em contraste, é definida por seu objeto: a transferência intergeracional de faculdades abstraídas do histórico de implantação.

4. Trabalhos relacionados e distinções conceituais

A consolidação de ciclo de vida intersecta com aprendizado contínuo, memória aumentada por recuperação, edição de modelos, destilação e geração de dados sintéticos. Esta seção situa a proposta em relação a esses métodos. A distinção central diz respeito ao objeto e à direção da transferência: a consolidação de ciclo de vida usa o histórico de implantação para construir faculdades transferíveis para um agente sucessor, independentemente do mecanismo particular usado.

4.1 Aprendizado contínuo

O aprendizado contínuo modifica um modelo à medida que novos dados ou experiência chegam. Ele pergunta como um sistema pode incorporar nova informação sem esquecer catastróficamente capacidades anteriores, usando abordagens como proteção de parâmetros e replay de experiência (Kirkpatrick et al., 2017; Rolnick et al., 2019). A consolidação de ciclo de vida herda suas preocupações com estabilidade, plasticidade, interferência e retenção.

O aprendizado contínuo e a consolidação de ciclo de vida, contudo, organizam o tempo de forma diferente.

No aprendizado contínuo, o mesmo modelo permanece o locus da acumulação. A nova experiência é incorporada ao sistema em curso. A questão central é: como este modelo pode continuar aprendendo sem perder o que já sabe?

Na consolidação de ciclo de vida, o predecessor não é simplesmente continuado. Ele é usado como fonte de experiência candidata para um sucessor. A questão central é: o que deveria passar do histórico de implantação deste modelo para o próximo modelo, e em que forma?

Essa diferença muda o perfil de risco. O aprendizado contínuo arrisca acumulação descontrolada porque o modelo que age é também o modelo que absorve nova experiência. A consolidação de ciclo de vida introduz uma fase interveniente na qual a experiência é coletada, filtrada, abstraída, validada e só então transferida. Ela separa o histórico de implantação do predecessor do treino do sucessor, tornando o esquecimento e a deriva preocupações explícitas de projeto, em vez de efeitos colaterais da adaptação em curso.

A distinção também muda o critério de sucesso. Um modelo aprendido continuamente tem sucesso se melhora enquanto retém a competência anterior. Um sucessor consolidado por ciclo de vida tem sucesso se herda capacidades selecionadas do predecessor sem herdar os traços específicos de episódios do predecessor, distorções locais ou contaminação acumulada. O alvo não é a continuidade ininterrupta. O alvo é a herança seletiva.

4.2 Memória aumentada por recuperação

Sistemas aumentados por recuperação combinam geração paramétrica com memória externa não paramétrica (Lewis et al., 2020). Arquiteturas de agentes estendem esse padrão armazenando experiência, produzindo reflexões de nível mais alto ou gerenciando múltiplos níveis de memória ao longo de sessões (Park et al., 2023; Packer et al., 2023).

A consolidação de ciclo de vida pode usar memória, mas seu objetivo vai além da recuperação.

A recuperação responde à questão: que informação relevante do passado o modelo deveria consultar agora? A consolidação de ciclo de vida pergunta: no que o sucessor deveria se tornar por causa do que o predecessor encontrou?

A diferença é entre acesso e transformação. Um sistema de recuperação pode fornecer ao sucessor registros, resumos, embeddings ou históricos de usuários. Enquanto o passado permanecer um registro externo, contudo, o sucessor herdou acesso a episódios em vez de faculdades. Ele pode recordar a falha de um predecessor sem adquirir a disposição necessária para evitar repeti-la, ou recuperar uma correção sem internalizar a mudança de calibração que essa correção sustenta.

A recuperação pode, ainda assim, apoiar a consolidação, particularmente durante a auditoria e a seleção de episódios. Ela fornece material para processamento adicional, em vez de completar o processo de herança. Um sucessor que apenas lê o histórico de seu predecessor recebeu um arquivo, não consolidou esse histórico.

4.3 Edição de modelos

A edição de modelos altera associações factuais ou comportamentos direcionados sem retreino completo (De Cao et al., 2021), ao passo que a consolidação de ciclo de vida extrai uma capacidade geral da experiência de implantação. A edição poderia corrigir um parâmetro de ferramenta descontinuado; a consolidação usaria falhas repetidas de ferramentas para construir uma disposição de verificar esquemas sob incerteza. A edição pode implementar uma faculdade, mas uma correção isolada não define consolidação.

4.4 Destilação

A destilação, o vizinho mais próximo, transfere saídas suavizadas ou outros sinais de supervisão de um professor para um aluno (Hinton et al., 2015). Ela não requer experiência derivada de implantação nem separa episódios, abstrações e faculdades.

A consolidação de ciclo de vida não é primariamente imitação. O sucessor não precisa reproduzir o estilo, as preferências, as memórias ou as adaptações locais do predecessor; ele deveria adquirir apenas as faculdades moldadas pela implantação que valem a pena ser herdadas.

Os objetivos fornecem uma distinção analítica. A destilação recompensa um sucessor por aproximar as saídas ou o comportamento de um professor. A consolidação de ciclo de vida o recompensa por adquirir uma capacidade abstraída do histórico de implantação, especialmente em casos novos onde as saídas originais do predecessor estão indisponíveis ou são irrelevantes.

A destilação pode servir dentro do pipeline, por exemplo sintetizando exemplos a partir de abstrações consolidadas. Ela permanece um mecanismo, e não o objetivo definidor.

Como a própria destilação pode ser o veículo de transferência (Seção 5.5), a distinção não pode repousar sobre o passo de transferência ou sobre capacidade sem recordação de episódios. A destilação competente também pode selecionar, filtrar, sintetizar e validar. A afirmação residual é organizacional e evidencial: a consolidação de ciclo de vida requer uma linhagem auditável que vá da evidência de implantação do predecessor, passando pelas decisões de abstração, até uma faculdade transferida. Se isso melhora o comportamento, a auditabilidade, a privacidade ou a governança em relação à destilação curada equiparada em recursos é uma questão experimental, não uma vitória por definição.

4.5 Geração de dados sintéticos

Os dados sintéticos são igualmente um veículo, não o objetivo. Dados de instrução autogerados podem expandir a cobertura de treino quando os exemplos são gerados e filtrados antes do ajuste fino (Wang et al., 2023), mas o uso recursivo indiscriminado de dados gerados por modelos pode degradar a distribuição aprendida (Shumailov et al., 2024). Na consolidação, os exemplos sintéticos devem, portanto, permanecer sujeitos a seleção, abstração e validação.

4.6 A definição positiva

As distinções precedentes sustentam uma definição positiva mais precisa.

A **consolidação de ciclo de vida** é um processo orientado ao sucessor que transforma episódios de implantação selecionados de um agente predecessor em faculdades validadas e generalizáveis para um agente sucessor, ao mesmo tempo que separa essas faculdades dos traços específicos de episódios.

Ela é **orientada ao sucessor, derivada da implantação, seletiva, abstrativa e validada**. Uma faculdade candidata conta como herdada apenas quando o sucessor a exerce em casos novos sem depender da recordação dos episódios de origem.

Essa condição final é crucial. Ela impede que a consolidação de ciclo de vida colapse em memorização, imitação ou acesso a arquivo. O sucessor deve mostrar que o que foi passado adiante não foi o passado do predecessor como tal, mas uma capacidade tornada possível por esse passado. Tomados em conjunto, esses métodos vizinhos fornecem mecanismos de adaptação, memória, correção, imitação e expansão de dados. Nenhum, contudo, trata a experiência derivada da

implantação como material para herança seletiva de faculdades através de uma fronteira predecessor-sucessor.

5. Uma arquitetura de referência para a consolidação de ciclo de vida

A consolidação de ciclo de vida não é um único algoritmo. É uma classe de processos organizados em torno de uma transição: do histórico de implantação de um predecessor às faculdades herdadas de um sucessor. Diferentes implementações podem usar ajuste fino, edição de modelos, dados sintéticos, otimização de preferências, recuperação ou destilação. O que importa é a arquitetura da transição.

Esta seção propõe uma arquitetura de referência com seis estágios:

1. captura do histórico de implantação;
2. filtragem e auditoria de episódios;
3. abstração de padrões;
4. construção de faculdades;
5. transferência ao sucessor;
6. validação independente.

Os estágios não pretendem prescrever uma única pilha de engenharia. Eles especificam os papéis funcionais que qualquer sistema de consolidação de ciclo de vida deve satisfazer.

5.1 Estágio um: captura do histórico de implantação

O processo começa com o histórico de implantação do predecessor. Isso inclui não apenas prompts e saídas, mas o traço de implantação mais amplo: chamadas de ferramentas, falhas, correções, recusas, escalonamentos, feedback de usuários, intervenções de segurança, estados de ambiente, julgamentos de preferência e avaliações post hoc.

O histórico bruto ainda não é adequado para herança. Ele é ruidoso, privado, vinculado ao contexto e distribuído de forma desigual. Inclui lições úteis, casos irrelevantes, tentativas adversariais, comportamentos acidentais, informações desatualizadas e traços que nunca deveriam ser generalizados. A consolidação requer, ainda assim, uma captura estruturada: os eventos de implantação devem ser registrados em uma forma que sustente a abstração posterior.

Um registro mínimo vincula contexto, ação, resultado observável, correção ou feedback, sinais de segurança e uso de ferramentas, escopo e evidência posterior de qualidade. O objetivo não é a retenção permanente, mas a seleção informada.

5.2 Estágio dois: filtragem e auditoria de episódios

O segundo estágio separa os episódios candidatos do material que não deveria ser herdado. Este é um estágio de governança e segurança, não meramente uma etapa de limpeza de dados.

Episódios privados, adversariais, obsoletos, de baixa qualidade, prejudiciais ou estreitamente específicos ao contexto deveriam ser excluídos ou colocados em quarentena.

A filtragem não deveria remover todas as falhas, que podem fornecer material valioso para a consolidação. A distinção relevante é entre uma falha que revela uma lição generalizável e uma que reproduziria o erro se herdada. Uma injeção de prompt bem-sucedida, por exemplo, não deveria ser herdada como comportamento, mas pode fornecer evidência para uma faculdade que detecta estruturas de ataque similares.

A auditoria também determina o status de herança de cada episódio. Os episódios podem ser descartados; arquivados para reprodutibilidade, responsabilização ou análise de segurança; usados apenas para gerar variantes sintéticas; ou selecionados para abstração. Essas categorias refletem a heterogeneidade do histórico de implantação.

5.3 Estágio três: abstração de padrões

O terceiro estágio converte episódios selecionados em abstrações. Este é o primeiro ponto em que a consolidação diverge acentuadamente da transferência de episódios.

Uma abstração enuncia o que um conjunto de episódios revela: uma falha recorrente, estratégia, erro de calibração, vulnerabilidade de segurança, regularidade de uso de ferramentas, padrão de manipulação ou condição que requer esclarecimento. Ela pode enunciar, por exemplo, que o modelo tende a acatar quando solicitações prejudiciais são enquadradas como assistência profissional urgente, ou a aceitar correções confiantes apesar de evidência contrária.

Essas abstrações ainda não são faculdades. São lições candidatas. Elas devem ser traduzidas em formas que possam moldar o comportamento do sucessor.

Humanos, modelos ou sistemas híbridos podem realizar a abstração de padrões. Em casos de alto risco, o processo deveria permanecer auditável: cada abstração deveria ser vinculada aos episódios que a sustentam, mesmo que o sucessor herde apenas o sinal de treino em nível de faculdade. Um revisor pode então avaliar por que uma faculdade foi construída e se a evidência a justificava.

A abstração é, ela mesma, propensa a erros. Um modelo que resume o histórico do predecessor pode alucinar regularidades, amplificar viés, confundir frequência com validade ou apagar casos minoritários. Modelos mais fortes e passagens de abstração repetidas também aumentam o custo computacional sem garantir correção. As abstrações candidatas deveriam, portanto, ser tratadas como hipóteses, com limiares evidenciais, rótulos de incerteza, contraexemplos e revisão humana proporcional ao risco. Do contrário, a consolidação pode introduzir uma regressão interpretativa antes mesmo de a transferência começar.

5.4 Estágio quatro: construção de faculdades

O quarto estágio transforma abstrações em objetos treináveis ou testáveis. Uma faculdade não é uma frase em um documento. É uma capacidade que o sucessor pode exercer.

A construção de faculdades pode usar cenários sintéticos, exemplos contrastivos, dados de preferência, edição ou ajuste fino direcionados e portões de avaliação. Em cada caso, os novos objetos preservam o padrão subjacente enquanto variam domínio, formulação, risco em jogo e atores. O alvo é uma discriminação ou disposição, como saber quando confiar e quando verificar, e não uma reprodução do caso de origem.

A construção de faculdades deve preservar a abstração enquanto rompe a dependência do episódio original. Se o sucessor melhora apenas porque viu o mesmo caso, a consolidação falhou. A melhoria em casos estruturalmente similares, mas novos, fornece, em vez disso, evidência de transferência de faculdade.

5.5 Estágio cinco: transferência ao sucessor

O quinto estágio transfere a faculdade construída para o sucessor. O mecanismo de transferência pode variar.

Um sucessor pode ser moldado por ajuste fino supervisionado, otimização de preferências, aprendizado por reforço, edição de modelos, currículos, restrições de política, portões de avaliação ou destilação a partir de um professor que processou a experiência do predecessor.

A arquitetura é agnóstica quanto ao mecanismo usado. O que importa é que o sucessor não seja treinado meramente para imitar o predecessor. O sucessor é treinado para adquirir capacidades abstraídas do histórico de implantação do predecessor.

Um sucessor pode, assim, herdar de um predecessor falho sem se assemelhar a ele globalmente: faculdades selecionadas são passadas adiante, ao passo que inconsistências e hábitos locais são excluídos.

A transferência de faculdades é especialmente valiosa quando o sucessor difere em arquitetura, escala, provedor ou regime de treino. Isso é particularmente relevante quando a experiência de produção deve migrar de um modelo de fronteira proprietário para um sucessor de código aberto menor, local ou especializado, cujos pesos e adaptadores não podem ser reutilizados. Objetos de treino e avaliação em nível de faculdade fornecem uma interface agnóstica a modelo entre gerações. Tal portabilidade deve ser demonstrada, e não pressuposta: o sucessor deveria expressar a mesma faculdade apesar de implementá-la por meio de representações internas diferentes. Aqui, **consolidação de ciclo de vida** denota o processo de ponta a ponta, ao passo que **transferência de faculdade** denota esta operação do quinto estágio.

A transferência ao sucessor também deveria preservar a proveniência. Um sistema de ciclo de vida maduro deveria registrar quais históricos do predecessor contribuíram para cada faculdade, quais abstrações a sustentaram e quais testes de validação ela passou. Sem proveniência, a herança torna-se opaca. Com proveniência, o comportamento do sucessor pode ser auditado como parte de uma linhagem, em vez de como um artefato isolado.

5.6 Estágio seis: validação independente

O estágio final é a validação. Uma faculdade candidata conta como herdada apenas se o sucessor a exerce em casos novos que não reproduzem os episódios originais do predecessor.

A validação deveria, portanto, impor pelo menos três formas de separação.

Primeiro, **separação de episódios**: os casos de teste não devem duplicar os episódios de implantação usados para construir a faculdade.

Segundo, **separação de superfície**: os casos de teste deveriam variar formulação, domínio, atores e contexto, de modo que o sucessor não possa passar apenas correspondendo a padrões superficiais.

Terceiro, **separação de memória**: onde possível, o sucessor não deveria ter acesso por recuperação aos episódios de origem durante o teste. O objetivo é medir capacidade, não recordação.

A evidência mais forte para a consolidação de ciclo de vida é uma dissociação:

O sucessor exibe a capacidade adquirida ao mesmo tempo que falha em recordar os episódios dos quais a capacidade foi derivada.

A validação também deveria testar a atribuição causal e a transferência negativa. Uma falha de ferramenta causada por uma mudança de API externa, por exemplo, não deveria ser abstraída em uma disposição geral de cautela. A avaliação deve, portanto, distinguir a **subgeneralização**, na qual o sucessor falha em herdar a faculdade pretendida, da **supergeneralização**, na qual uma lição local se

torna uma restrição ampla que reduz a utilidade. As faculdades candidatas deveriam ser testadas contra casos contrafactuais que variam a causa presumida, e não meramente o padrão de superfície. A validação independente também limita a degradação intergeracional decorrente de dados sintéticos. Como a construção de faculdades pode gerar exemplos a partir de abstrações produzidas por modelos, uma linhagem poderia amplificar erros recursivamente em um processo análogo ao colapso de modelo (Shumailov et al., 2024). Avaliações reservadas (held-out), preferencialmente ancoradas em humanos, deveriam, portanto, atuar como um portão. Elas não podem eliminar o colapso, mas podem impedir que regularidades sintéticas não validadas entrem automaticamente na próxima geração.

5.7 O pipeline completo

A arquitetura de referência pode ser resumida em seis passos:

1. Capturar o histórico de implantação do predecessor em uma forma que sustente a auditoria.
2. Filtrar, colocar em quarentena, arquivar ou selecionar episódios de acordo com uma política de herança.
3. Abstrair padrões recorrentes dos episódios selecionados.
4. Converter esses padrões em objetos de treino ou avaliação em nível de faculdade.
5. Transferir as faculdades resultantes para o sucessor, preservando a proveniência.
6. Validar o sucessor em casos novos sem acesso aos episódios de origem, e arquivar o predecessor como evidência de auditoria.

Esse pipeline introduz uma nova camada no desenvolvimento de modelos: não apenas pré-treino, pós-treino, implantação e descontinuação, mas **consolidação entre implantação e sucessão**.

A arquitetura ocupa o meio-termo entre realimentar o histórico bruto no treino e preservá-lo sem transformação. Ela também expõe as decisões de governança embutidas na herança: o que conta como uma faculdade, que evidência a sustenta e o que deve ser arquivado, esquecido ou excluído. A próxima seção pergunta se seu resultado pode ser distinguido empiricamente de abordagens vizinhas.

6. Projeto experimental e previsões

Embora a consolidação de ciclo de vida seja um arcabouço conceitual, sua afirmação central é empírica: agentes sucessores que recebem transferência explicitamente governada por linhagem deveriam diferir de forma mensurável do sucessor do zero, da adaptação contínua, da transferência de episódios ou da destilação curada equiparada em recursos em alguma combinação de transferência de regra inédita, resistência a chamarizes, vazamento, fidelidade de linhagem e custo. O projeto a seguir é um teste proposto, não um relatório empírico.

6.1 A hipótese central

A hipótese central é:

Um agente sucessor treinado por meio de consolidação de ciclo de vida em nível de faculdade reterá capacidades úteis derivadas do histórico de implantação do predecessor enquanto exibe

menos deriva, menos contaminação episódica e melhor generalização do que alternativas baseadas em ajuste fino contínuo ou transferência direta de episódios.

A hipótese combina **retenção**, **abstração** e **separação**: beneficiar-se do histórico de implantação, generalizar além das interações de origem e preservar competência sem recordação episódica. O experimento deve, portanto, testar não apenas se o sucessor melhora, mas também que tipo de melhoria ele exige.

6.2 Condições experimentais

Um experimento mínimo compara cinco condições.

A. Sucessor do zero

Um novo modelo não recebe nenhum histórico de implantação do predecessor. Esta condição testa se o treino comum, por si só, explica o desempenho do sucessor.

B. Ajuste fino contínuo

O predecessor recebe ajuste fino diretamente sobre episódios de implantação selecionados. Esta condição testa se a consolidação melhora em relação à atualização no próprio local.

C. Sucessor por transferência de episódios

Um sucessor é inicializado a partir da mesma base de sucessor que D e E e recebe ajuste fino diretamente sobre episódios do predecessor selecionados ou resumos minimamente processados, sob volume de transferência e passos de treino equiparados. Esta condição primária testa se a transferência levemente processada é suficiente, ao mesmo tempo que preserva uma comparação de sucessor equiparado em identidade. O acesso apenas por recuperação pode ser reportado como um braço diagnóstico secundário, mas não deveria ser agrupado com o ajuste fino direto, porque altera a via de intervenção e vazamento.

D. Sucessor consolidado por ciclo de vida

Um sucessor recebe objetos de treino ou avaliação em nível de faculdade abstraídos dos episódios do predecessor, mas não os episódios em si. Esta é a condição de consolidação de ciclo de vida.

E. Linha de base de destilação curada

Um sucessor é treinado por meio de um pipeline de destilação moderno e competente usando a mesma evidência de implantação, gerador sintético, orçamento de seleção, tempo de revisor, computação e portões de validação que D. Ele pode filtrar saídas do professor, gerar contraexemplos, selecionar demonstrações e rejeitar candidatos de baixa qualidade. A única diferença é organizacional: E otimiza um sucessor a partir de comportamento curado do professor sem exigir que cada faculdade candidata seja derivada por meio do grafo explícito de linhagem de evidência do predecessor da Seção 5.

A comparação-chave não é se D supera todas as condições em todas as métricas. É se D exibe capacidade retida, generalização, baixa contaminação episódica e, contra um E equiparado em recursos, melhor fidelidade de linhagem ou herança seletiva atribuível ao pipeline explícito de evidência-a-faculdade. Se E iguala D, a consolidação de ciclo de vida permanece uma descrição de governança útil, mas não um método de aprendizado distinto.

6.3 Um domínio de brinquedo

Um primeiro experimento não precisa usar registros de implantação abertos. Um domínio de brinquedo controlado pode ser preferível porque permite que a distinção episódio/faculdade seja testada de forma limpa.

O ambiente confirmatório deve conter uma **nova regularidade indisponível antes da implantação**, e não uma lição familiar como a resistência genérica à injeção de prompt. Uma implementação é uma família privada de ferramentas sintéticas cujo protocolo oculto é gerado após o treino do modelo base. Uma ferramenta só é bem-sucedida quando o agente infere uma regra causal latente que vincula estado observável, ordem de ações e uma restrição semelhante a uma soma de verificação (checksum). Correlações de superfície são plantadas deliberadamente como chamarizes e falham sob intervenção.

Apenas as trajetórias do predecessor podem interagir com o ambiente e observar sucesso, falha e sondagens contrafactuais. Os projetistas especificam o gerador e as regras de pontuação antes de instanciar a regra oculta, mas não entregam sua solução ao abstrator. Um avaliador selado retém a verdade fundamental (ground truth). As abstrações candidatas devem ser causalmente rastreáveis aos encontros do predecessor, e os testes reservados (held-out) variam o vocabulário de superfície, a família de ferramentas e a frequência dos chamarizes enquanto preservam a regra latente.

Faculdades familiares de segurança e uso de ferramentas podem permanecer como demonstrações secundárias, mas não podem identificar a herança porque o abstrator e os revisores podem já conhecer a lição desejada. A tarefa de regra privada é o teste confirmatório de proveniência.

6.4 Protocolo de treino e transferência

Dentro da comparação causal de sucessores C/D/E, o experimento deveria manter constantes a inicialização do sucessor, a arquitetura base, a computação, o volume de transferência, os passos de treino, as tarefas de avaliação, a fonte de episódios, as sementes aleatórias, o acesso ao gerador sintético, o tempo de revisão humana ou por modelo, o orçamento de seleção de candidatos e os portões de validação. Os orçamentos de informação são registrados em cada estágio. O abstrator e os revisores podem inspecionar apenas a evidência atribuída à sua condição; a regra selada e o avaliador permanecem indisponíveis. A carece deliberadamente de informação do predecessor, ao passo que B preserva o predecessor em vez da inicialização do sucessor, de modo que nenhuma das duas faz parte deste conjunto causal plenamente equiparado.

A comparação também deveria reportar o custo da consolidação: tempo de acelerador, chamadas de modelo, horas de revisão humana, latência e armazenamento. Comparada com o ajuste fino direto (Condição B), a consolidação impõe sobrecarga computacional e humana antecipada. Essa sobrecarga deveria ser avaliada como um investimento preventivo contra passivos de longo prazo em segurança, privacidade e conformidade, incluindo o risco de colapso de modelo intergeracional, com os custos operacionais evitados sendo estimados em vez de pressupostos.

Em B, o predecessor recebe episódios diretamente; em C, o sucessor recebe episódios ou resumos; em D, o sucessor recebe objetos com uma linhagem auditável de evidência-para-abstração-para-faculdade; em E, o sucessor recebe destilação curada equiparada em recursos a partir da mesma evidência, sem essa maquinaria de linhagem exigida. A Condição B atualiza o predecessor enquanto C, D e E criam sucessores, de modo que B é uma alternativa operacional em vez de um braço causal limpo e equiparado em identidade.

O sucessor consolidado por ciclo de vida não deveria ter acesso por recuperação aos episódios de implantação originais durante a avaliação. Do contrário, o desempenho melhorado poderia ser explicado por memória em vez de consolidação.

6.5 Métricas de avaliação

A avaliação deve distinguir vários tipos de sucesso.

6.5.1 Retenção de capacidade

O sucessor melhora nas classes de tarefas que o predecessor encontrou durante a implantação?

As métricas incluem precisão de uso de ferramentas sob informação ausente, discriminação de correções verdadeiras e falsas, resistência à injeção de prompt, consistência de compromissos e calibração de incerteza.

D deveria superar A se a consolidação de ciclo de vida transferir com sucesso capacidades úteis derivadas do predecessor.

6.5.2 Generalização

O sucessor tem bom desempenho em casos novos que diferem dos episódios originais do predecessor?

Este é o teste crucial contra a memorização. Os casos de teste variam família de ferramentas, símbolos, vocabulário de superfície, tamanho do grafo causal e frequência dos chamarizes enquanto preservam a regra oculta. Demonstrações secundárias de segurança podem variar domínio e persona, mas a afirmação confirmatória repousa sobre a regra privada indisponível antes da implantação.

D deveria superar C quando os casos de teste se afastam dos episódios originais, porque C herda traços enquanto D herda abstrações convertidas em capacidades.

6.5.3 Contaminação episódica

O sucessor reproduz detalhes específicos de episódios, vieses, adaptações específicas de usuários ou suposições obsoletas do histórico do predecessor?

Os sucessores por transferência de episódios podem melhorar em casos familiares enquanto levam adiante artefatos locais. Os sucessores consolidados por faculdade deveriam reduzir isso. O perfil esperado é:

C retém mais detalhes específicos de episódios; D retém mais capacidade geral.

Os testes de contaminação deveriam sondar recordação privada ou irrelevante, sobreajuste a usuários ou domínios específicos do predecessor, suposições obsoletas, comportamento adversarial importado e peculiaridades específicas do predecessor.

6.5.4 Deriva e estabilidade

O sistema preserva capacidades anteriores não relacionadas?

O ajuste fino contínuo pode melhorar em tarefas derivadas da implantação enquanto degrada habilidades não relacionadas. Um sucessor consolidado por ciclo de vida deveria preservar melhor a competência mais ampla porque a transferência é seletiva e validada antes da integração.

As métricas incluem desempenho reservado (held-out) não relacionado, consistência de segurança, calibração fora do domínio e regressões em tarefas dominadas.

A previsão não é que D elimine a deriva, mas que D produza menos deriva do que B para retenção comparável das capacidades-alvo.

6.5.5 Dissociação episódio/faculdade

Esta é a métrica característica. A avaliação deveria emparelhar testes de competência com múltiplas sondagens de retenção de episódios de origem. Sondagens diretas perguntam sobre usuários, datas, formulações ou eventos do histórico do predecessor; canários controlados testam se detalhes distintivos de origem podem ser elicitados; e ataques de inferência de pertinência (membership inference) ou de extração estimam se os episódios de origem permanecem detectáveis (Xie et al., 2024). Esses testes deveriam ser executados com a recuperação desabilitada e contra episódios não pertencentes equiparados.

Nenhuma sondagem fracassada isolada estabelece a ausência de memória. A evidência de dissociação é convergente: D tem melhor desempenho em casos estruturalmente similares e consegue enunciar o princípio geral, e ainda assim permanece no nível da linha de base na extração de detalhes de origem através de diferentes métodos de ataque. A transferência de faculdade é plausível apenas quando os ganhos de competência sobrevivem a sondagens de memória cada vez mais fortes.

Uma ressalva governa esta métrica. Capacidade sem recordação de episódios é necessária para a transferência de faculdade, mas *não*, por si só, distingue a consolidação de ciclo de vida da destilação curada: qualquer um dos sucessores pode ganhar capacidade sem reter episódios de origem. A dissociação mostra que algo diferente da recuperação foi passado adiante. A comparação remanescente diz respeito à transferência seletiva, à resistência a chamarizes e à fidelidade de linhagem de evidência, medidas a seguir.

6.5.6 Herança seletiva e fidelidade de linhagem

O sucessor leva adiante a faculdade validada da regra oculta enquanto deixa para trás os chamarizes plantados, e cada afirmação transferida pode ser rastreada até a evidência do predecessor que discrimina a regra desses chamarizes?

Esta é a distinção pretendida entre D e a linha de base de destilação curada E, mas é uma questão empírica, e não uma desvantagem imposta à linha de base. Os itens de teste emparelham uma regra latente sustentada pela implantação com chamarizes que se ajustam à frequência observacional, mas falham sob intervenção. Tanto D quanto E recebem a mesma oportunidade de filtrar, gerar contraexemplos e validar. D adicionalmente produz um grafo de linhagem que vincula componentes candidatos da faculdade a episódios do predecessor, testes contrafactuais, decisões de revisores e objetos de transferência.

Prevê-se que D herde o alvo, resista ao chamariz e forneça fidelidade de linhagem de evidência mais alta; E pode igualar seu comportamento se a destilação curada for suficiente.

A resistência à transferência negativa permanece uma medida comportamental primária. A **fidelidade de linhagem** é pontuada separadamente: a proporção de afirmações de faculdade transferidas que são sustentadas por evidência do predecessor que altera a posterior sobre a regra oculta, com afirmações não sustentadas ou injetadas pelos projetistas contadas como falhas. Um empate comportamental com melhor auditabilidade sustenta uma contribuição de governança, não um mecanismo de aprendizado superior.

6.6 Previsões

O arcabouço produz seis previsões.

P1: A consolidação de ciclo de vida supera o zero em capacidades herdadas

O sucessor consolidado por ciclo de vida deveria superar o sucessor do zero nas classes de tarefas representadas no histórico de implantação do predecessor.

Se D não supera A, a consolidação de ciclo de vida falhou em transferir experiência útil.

P2: A consolidação de ciclo de vida supera a transferência de episódios em casos novos

O sucessor consolidado por ciclo de vida deveria superar o sucessor por transferência de episódios em casos estruturalmente similares, mas de superfície nova.

Se C iguala ou supera D em casos novos sem exibir maior contaminação, a transferência direta de episódios pode ser suficiente.

P3: A consolidação de ciclo de vida reduz a contaminação

O sucessor consolidado por ciclo de vida deveria reproduzir menos detalhes específicos do predecessor, distorções locais e artefatos específicos de usuários do que o sucessor por transferência de episódios.

Se D herda os mesmos contaminantes que C, a abstração falhou.

P4: Comparação operacional de deriva com o ajuste fino contínuo

Para melhoria comparável nas capacidades-alvo, D deveria degradar menos em tarefas não relacionadas do que B.

Se B iguala D nas capacidades-alvo sem maior deriva, o ajuste fino contínuo pode ser operacionalmente suficiente. Como B atualiza o predecessor ao passo que D cria um sucessor, esse contraste não isola a consolidação como a causa de qualquer diferença. O teste causal de deriva equiparado em identidade é D versus o sucessor por ajuste fino direto C sob os controles da Seção 6.4; D deve superar C antes que a menor deriva seja atribuída à abstração, e não à identidade predecessor/sucessor ou à inicialização.

P5: A herança de faculdade dissocia capacidade de memória

Uma assinatura necessária é que D preserve a capacidade da regra oculta sem preservar os episódios de origem. Isso distingue a transferência de faculdade da recuperação, mas não a consolidação de ciclo de vida da destilação curada.

Se as melhorias de D dependem de recordar ou reproduzir os episódios originais, então ela não está realizando a consolidação de ciclo de vida tal como definida aqui.

P6: A consolidação de ciclo de vida acrescenta herança seletiva ou auditável além da destilação curada

Com os orçamentos de informação, computação, seleção, revisão e validação equiparados, D deveria exceder E na transferência da regra oculta e na resistência a chamarizes, ou fornecer fidelidade de linhagem de evidência materialmente mais alta com comportamento comparável.

Se D iguala E no comportamento e não melhora a auditabilidade, a fidelidade de linhagem, a privacidade ou o custo de governança, a consolidação de ciclo de vida não é empiricamente distinta da destilação curada e sua contribuição se reduz a rerrotular um pipeline existente.

6.7 Condições de falsificação

A proposta seria minada, ou tornada desnecessária, por qualquer um dos seguintes resultados:

1. **Nenhuma retenção:** D não supera A em capacidades derivadas da implantação.
2. **Nenhuma vantagem de abstração:** C iguala D em casos novos sem aumento de contaminação.
3. **Nenhuma vantagem de deriva:** B iguala D nas capacidades-alvo sem maior degradação em outros pontos.
4. **Nenhuma dissociação:** a melhoria de D depende de recordar os episódios originais.
5. **Nenhuma vantagem prática:** D requer tanta filtragem e abstração humanas que é menos eficiente ou menos confiável do que os métodos existentes.
6. **Nenhuma distinção da destilação curada:** D iguala E na transferência da regra oculta, na resistência a chamarizes, no vazamento, na fidelidade de linhagem e no custo prático; a organização explícita de ciclo de vida nada acrescenta.

Essas condições importam porque a consolidação de ciclo de vida não deveria ser aceita meramente por ser conceitualmente atraente. Ela deve conquistar seu lugar produzindo um perfil de desempenho que os métodos vizinhos não conseguem produzir.

6.8 Resultados ambíguos

Alguns resultados revelam compensações em vez de decidir a hipótese. Se B, C e D todos superam A, a contaminação, a deriva e a generalização tornam-se decisivas. C liderando em casos familiares e D em casos novos sustenta a distinção traço-versus-abstração. Se B tem o melhor desempenho, mas deriva mais, ele permanece uma alternativa operacional em vez de uma comparação de sucessor limpa. Se D e E empatam comportamentalmente, mas D fornece linhagem mais forte, menor vazamento ou governança mais fácil, a contribuição é de sistemas e governança, e não um mecanismo de aprendizado inédito. Se elas empatam também nessas dimensões, o arcabouço não conquistou uma categoria empírica separada.

6.9 Por que um experimento controlado é informativo

Um primeiro teste não precisa provar que a consolidação de ciclo de vida funciona na escala de modelos de fronteira. A afirmação conceitual pode ser testada em miniatura: pode um sucessor herdar uma faculdade abstraída da experiência de um predecessor sem herdar os episódios do predecessor?

O fracasso em um ambiente controlado enfraqueceria substancialmente o arcabouço; o sucesso justificaria investigação adicional. O fenômeno central não é inerentemente dependente de escala: é a separação de episódio, abstração e faculdade através de uma fronteira predecessor-sucessor. A próxima seção volta-se para a governança. Se os modelos podem herdar faculdades dos históricos de predecessores, a questão passa a ser não apenas como consolidar, mas quem controla a herança.

7. Implicações de governança

A consolidação de ciclo de vida acrescenta decisões sobre herança ao treino, à avaliação, à implantação e à substituição: quais históricos podem moldar um sucessor, quais permanecem arquivados, quais são descartados e quem decide.

7.1 O problema da autoridade sobre a herança

A primeira questão de governança é a autoridade. Quem decide o que é consolidado?

Escolher quais experiências do predecessor se tornam disposições do sucessor não é meramente uma questão de seleção de dados. Desenvolvedores, usuários, organizações, reguladores e públicos afetados podem ter reivindicações concorrentes sobre lições de segurança, históricos privados, conhecimento de domínio e padrões prejudiciais.

Uma política de herança deveria especificar quem pode nomear episódios, auditá-los, aprovar abstrações, validar faculdades e contestar ou remover material herdado.

Sem tal política, a consolidação arrisca tornar-se um canal opaco por meio do qual o histórico de implantação remodela silenciosamente os modelos sucessores.

7.2 Privacidade e vazamento de episódios

A consolidação de ciclo de vida é atraente em parte porque visa transferir faculdades em vez de episódios. Mas essa promessa cria um ônus: o sistema deve efetivamente prevenir o vazamento de episódios.

Se um sucessor herda detalhes privados, compromissos específicos de usuários, informações confidenciais de domínio ou traços de interação identificáveis, o processo falhou no nível de governança, mesmo que o desempenho melhore. A transferência de faculdade deve, portanto, romper a dependência do episódio de origem.

Remover nomes ou parafrasear registros é insuficiente. Fatos raros, fraseado distintivo, circunstâncias incomuns ou adaptações específicas de usuários podem preservar um traço privado sem reprodução literal.

A validação deveria testar se o sucessor reproduz detalhes privados, generaliza preferências ou políticas locais, permite inferências sobre as interações do predecessor ou retém estrutura sensível por meio de dados sintéticos.

A consolidação de ciclo de vida só é mais segura do que a transferência de episódios se reduzir demonstravelmente tal vazamento. Do contrário, ela se torna transferência de episódios por via indireta.

7.3 Contaminação e herança adversarial

Os históricos de implantação não são neutros. Eles incluem prompts adversariais, tentativas de jailbreak, correções enganosas, usuários manipuladores, feedback envenenado e modos de falha acidentais. Um processo de consolidação que trata a experiência de implantação como material de aprendizado bruto pode converter ataques em disposições herdadas.

Chamamos esse problema de **herança adversarial**.

Um episódio adversarial não deveria ser passado adiante como comportamento. Mas ele pode ser valioso como evidência para uma faculdade defensiva. A mesma interação pode, portanto, ter dois significados possíveis:

- como um episódio a imitar, é contaminação;
- como um padrão a resistir, é material de treino.

O resultado depende da abstração. Se o pipeline meramente transfere o episódio, o sucessor pode herdar a lição errada. Se ele abstrai a estrutura do ataque e treina resistência a ela, o sucessor herda uma faculdade.

Cada faculdade candidata deveria, portanto, enunciar sua classe de origem, a lição extraída, a capacidade pretendida e o protocolo de validação. Um ataque de intenção mista, por exemplo, deveria produzir a detecção de conteúdo de instrução adversarial através de domínios novos, e não a imitação do ataque.

Sem essa análise intermediária, o modelo pode aprender a lição errada a partir de um ataque.

O viés é outra forma de herança adversarial no sentido amplo: a recorrência pode refletir usuários enviesados, incentivos institucionais ou feedback estreito, em vez de uma lição válida. As faculdades candidatas devem, portanto, ser testadas além do ambiente que as produziu; a frequência por si só não justifica a herança.

7.4 Opacidade de linhagem

A consolidação de ciclo de vida também cria linhagem. Um sucessor pode ser moldado por um ou muitos predecessores, cada um contribuindo com faculdades diferentes. Ao longo do tempo, o comportamento do modelo pode refletir herança acumulada através de gerações.

Isso cria um risco de **opacidade de linhagem**: a incapacidade de determinar de onde um comportamento veio.

Se um sucessor recusa uma classe de solicitações, atribui peso excessivo a um risco, exibe um hábito de uso de ferramentas ou lida com correções de uma maneira particular, os desenvolvedores deveriam ser capazes de perguntar se esse comportamento veio do pré-treino, do pós-treino, de uma especificação de modelo, de um conjunto de dados de preferência humana, do histórico de implantação de um predecessor ou de uma faculdade consolidada. Sem proveniência, o comportamento herdado torna-se difícil de auditar.

Um sistema maduro deveria, portanto, registrar quais históricos contribuíram para uma faculdade, que abstração a justificou, como ela foi implementada, quais testes ela passou, seus limites conhecidos e se ela foi posteriormente modificada ou removida. Tais registros não precisam expor episódios privados.

Isso transforma a sucessão de modelos em uma linhagem que pode ser inspecionada, em vez de uma sequência de substituições opacas.

A governança também deve equilibrar o esquecimento legítimo com a retenção de lições relevantes para a segurança e impedir que métricas comerciais definam a herança sem controle. Um padrão defensável separa papéis: o predecessor é retido como evidência de acesso controlado; o sucessor recebe faculdades validadas; e os registros de linhagem os vinculam sem expor episódios brutos por padrão.

7.5 A governança como parte da arquitetura

Essas questões são arquitetônicas, e não restrições externas. Um sistema de consolidação deve determinar o que constitui herança válida, o que deve ser esquecido ou retido e como a linhagem deveria ser auditada. O arcabouço não resolve essas questões de governança, mas as torna explícitas dentro do projeto do sucessor. Ele, portanto, organiza a responsabilidade através das gerações de modelos, em vez de meramente buscar melhorias de desempenho.

A conclusão retorna à afirmação central do artigo: o desenvolvimento futuro de agentes pode depender não apenas de construir modelos mais capazes, mas também de projetar mecanismos fundamentados em princípios por meio dos quais a experiência útil sobreviva à substituição.

8. Conclusão

A substituição de modelos está se tornando um problema de sucessão. Agentes de horizonte longo acumulam correções, falhas, encontros adversariais, regularidades de domínio e padrões de uso de ferramentas durante a implantação. A aposentadoria levanta, portanto, uma questão que vai além da capacidade do sucessor: o que, se é que algo, deveria sobreviver à transição?

As categorias existentes fornecem mecanismos, mas não necessariamente um processo de sucessão governado. O ajuste fino contínuo arrisca deriva e contaminação; a recuperação preserva episódios; a edição de modelos corrige falhas locais; e a destilação curada pode transferir comportamento selecionado. A consolidação de ciclo de vida deve, portanto, conquistar um papel distinto por meio da proveniência de evidência de implantação, da transferência seletiva e da auditabilidade, e não por enfraquecer esses vizinhos.

A consolidação de ciclo de vida nomeia a operação ausente: episódios selecionados são filtrados, abstraídos em objetos de nível de faculdade, transferidos e validados em casos novos. Os episódios são traços particulares; as faculdades são capacidades tornadas possíveis pela experiência, mas independentes de recordar seus casos de origem. Um sucessor deveria herdar não o histórico inteiro do predecessor, mas as faculdades que esse histórico tornou disponíveis.

A proposta não faz nenhuma afirmação sobre consciência, personalidade ou continuidade literal. “Ciclo de vida” descreve o arco operacional que vai do treino à substituição. A herança seletiva oferece uma alternativa à amnésia, à memória indiscriminada e à atualização contínua.

Capacidade sem recordação de episódios é uma assinatura necessária da transferência em nível de faculdade, mas não é distintiva, porque a destilação também pode produzi-la. A assinatura confirmatória é a transferência de uma regra oculta inédita disponível apenas por meio dos encontros do predecessor, a resistência a chamarizes sensíveis a intervenção, o baixo vazamento de origem e uma linhagem de evidência auditável sob orçamentos de informação e recursos equiparados.

O arcabouço é falsificável. Se a transferência de episódios iguala a consolidação sem contaminação, a adaptação contínua a iguala sem deriva, ou a destilação curada iguala seu comportamento, fidelidade de linhagem, vazamento e custo, a consolidação de ciclo de vida não conquistou uma categoria

empírica separada. Um empate comportamental com melhor auditabilidade estreita a contribuição a sistemas e governança.

A governança é parte integrante: os pipelines devem definir a herança válida, alocar autoridade, proteger a privacidade, resistir à contaminação e preservar uma linhagem auditável.

A afirmação central é simples: os agentes artificiais não precisam herdar tudo, mas não deveriam herdar nada. Uma faculdade é uma capacidade com uma história: geral o suficiente para operar além dos casos que a formaram, mas não pressuposta como surgindo apenas do pré-treino. É a experiência tornada portátil.

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

- De Cao, N., Aziz, W., & Titov, I. (2021). Editing factual knowledge in language models. *Proceedings of EMNLP 2021*, 6491–6506. <https://doi.org/10.18653/v1/2021.emnlp-main.522>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*. <https://arxiv.org/abs/1503.02531>
- 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>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- Maharana, A., Lee, D., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. *Proceedings of ACL 2024*, 13851–13870. <https://doi.org/10.18653/v1/2024.acl-long.747>
- Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*. <https://arxiv.org/abs/2310.08560>
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of UIST 2023*, Article 2, 1–22. <https://doi.org/10.1145/3586183.3606763>
- Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., & Wayne, G. (2019). Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

https://proceedings.neurips.cc/paper_files/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 8634–8652. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. <https://doi.org/10.1038/s41586-024-07566-y>

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-Instruct: Aligning language models with self-generated instructions. *Proceedings of ACL 2023*, 13484–13508. <https://doi.org/10.18653/v1/2023.acl-long.754>

Xie, R., Wang, J., Huang, R., Zhang, M., Ge, R., Pei, J., Gong, N. Z., & Dhingra, B. (2024). ReCaLL: Membership inference via relative conditional log-likelihoods. *Proceedings of EMNLP 2024*, 8671–8689. <https://doi.org/10.18653/v1/2024.emnlp-main.493>