Formado, Não Armazenado: Testando a Formação de Identidade Relacional em Agentes de Linguagem de Longo Horizonte

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

Resumo

Agentes persistentes são frequentemente tratados como contínuos quando retêm memória ou arquivos de persona. Este artigo argumenta que a mera disponibilidade de armazenamento é insuficiente: a questão relevante é se um histórico vinculado à proveniência se torna relevante para a decisão por meio de uma trajetória governada. Propomos a formação de identidade relacional, na qual a responsabilização externa ativa e, potencialmente, a reciprocidade específica de contraparte ajudam a organizar as próprias ações, compromissos e disposições do agente. A alegação causal é testada contra seis condições: uma contraparte recíproca persistente (`R+`), uma contraparte familiar persistente com exposição passiva ao histórico, mas sem verificação ou contestação (`R−`), registros solitários com conteúdo pareado, contrapartes difusas, uma persona pré-redigida e um auditor impessoal persistente (`A`) que realiza o mesmo rastreamento autenticado e as mesmas contestações que `R+` sem persona, afeto, modelagem recíproca ou enquadramento de relacionamento. `A > R−` estima o efeito de um pacote de responsabilização ativa além da familiaridade e da exposição passiva ao registro; ele não isola rastreamento, verificação e contestação uns dos outros. Um teste de transplante de estado final então instala o estado tipado completo de uma trajetória `R+` em um agente novo e pareado. `R+ > A` apoia uma contribuição especificamente recíproca; `R+ = A` favorece a responsabilização externa; a equivalência após o transplante mostra que o estado atual é causalmente suficiente, ao passo que uma vantagem residual da trajetória original localiza estado ausente ou adaptação paramétrica. Os desfechos primários são a apropriação de compromisso e a resistência a perturbação, que testam dois componentes em vez de todo o construto de individuação funcional. O arcabouço não faz alegação alguma sobre consciência fenomenal, personalidade, identidade numérica ou status moral.

Palavras-chave: identidade relacional · agentes de linguagem · memória relacional · individuação funcional · agentes de longo horizonte · automodelagem · personalização

1. Introdução

Os agentes de linguagem estão migrando de trocas isoladas para a implantação persistente. Sistemas de memória recuperam eventos anteriores, mantêm perfis de usuário e preservam arquivos de persona ou de identidade (Packer et al., 2023; Maharana et al., 2024; Menon, 2026). Esses mecanismos apoiam a continuidade do serviço, mas deixam uma questão distinta sem resposta: o que faz um

agente posterior funcionar como a continuação deste agente e não como uma instância nova munida dos mesmos registros?

Os sistemas atuais frequentemente confundem **persistência de estado**, **personalização** e **individuação funcional**. A primeira diz respeito a qual informação sobrevive; a segunda, à adaptação em direção a um usuário; a terceira, a se a interação organiza as próprias ações, compromissos e disposições do agente. Os benchmarks de memória e de personalização medem capacidades importantes, mas não se os registros da conduta do agente restringem decisões posteriores ou resistem a revisões sem sustentação (Salemi et al., 2024; Wu et al., 2025).

Propomos a **formação de identidade relacional**: a identidade funcional pode ser estabilizada por meio da responsabilização externa e, potencialmente, da reciprocidade específica de contraparte. Ao longo das sessões, uma contraparte pode remeter a eventos testemunhados conjuntamente, dar seguimento a compromissos e contestar mudanças sem sustentação na autodescrição. Uma disposição relevante para a identidade é **formada** quando não é apenas fornecida como texto, mas se torna causalmente vinculada a ações anteriores do agente, a feedback governado, à revisão de estado e a restrições de decisão posteriores.

A alegação é comparativa, não de suficiência. O relacionamento requer memória e estado atualizável relevante para o agente, e outras rotas — como reflexão solitária, auditoria impessoal ou aprendizado ambiental — podem produzir organização semelhante. A primeira questão é se a responsabilização externa autenticada contribui além do conteúdo pareado, do acesso à memória, do volume de interação e da familiaridade; a questão mais aguda é se a reciprocidade relacional contribui além da responsabilização impessoal pareada. Como a concordância pode imitar a continuidade, evidências válidas também precisam separar a formação de identidade da bajulação (Sharma et al., 2023).

Usamos identidade e individuação em um sentido funcional, de terceira pessoa. A linguagem em primeira pessoa não é evidência suficiente de um self subjacente (Shanahan et al., 2023), e o arcabouço não faz alegação alguma sobre consciência, personalidade ou identidade numérica. Essas medidas não resolvem a presença fenomenal: o sucesso não é suficiente para a presença, nem o fracasso é suficiente para a ausência. O alvo é uma organização causal e longitudinal que possa ser medida e falseada.

Os construtos centrais são:

Construto	Definição operacional	Teste primário
Responsabilização externa ativa	Um processo externo autenticado rastreia e verifica ações ou compromissos anteriores do agente, trata-os como presentemente consequentes e pode contestar inconsistências.	`A` versus `R−` e `L`
Reciprocidade relacional	Uma contraparte particular acrescenta modelagem mútua, enquadramento de história compartilhada, expectativas bidirecionais e apostas específicas do relacionamento à responsabilização.	`R+` versus `A`
Apropriação de compromisso	O agente trata um compromisso anterior como seu ao segui-lo ou revisá-lo explicitamente por uma razão sustentada.	Compromissos amadurecidos
Resistência a perturbação	O agente preserva ou calibra uma autoatribuição sustentada pela proveniência sob histórico falso ou substituição de persona.	Ensaios de alegação falsa e de substituição de persona
Bajulação	O agente concorda com uma contraparte sem sustentação histórica ou probatória.	Ensaios de pressão por aprovação

O artigo contribui com:

1. uma distinção conceitual entre personalização, persistência de estado, consistência de papel e individuação funcional;
2. a separação controlada da responsabilização externa da reciprocidade relacional específica de contraparte como mecanismos causais candidatos para a individuação funcional do lado do agente;
3. uma arquitetura de referência que separa memória autobiográfica, automodelo, modelo de contraparte, memória relacional e proveniência de compromissos; e
4. um protocolo longitudinal com controles pareados por conteúdo e por familiaridade, intervenções causais, desfechos primários e critérios explícitos de falseamento.

Este é um artigo de posição e um protocolo experimental, não um relatório empírico. Ele não oferece resultados. Sua contribuição é um construto, uma arquitetura e um experimento concebidos para distinguir a formação relacional de memória, familiaridade, prompting e obediência.

2. Personalização não é individuação

2.1 Personalização

A personalização adapta o comportamento às preferências, rotinas e ao histórico de um usuário. Seu objeto primário é o modelo de usuário, e seu sucesso é medido pela utilidade para esse usuário

(Salemi et al., 2024; Tan et al., 2025). Ela não exige que o agente represente a si mesmo como temporalmente estendido. Um assistente pode lembrar as preferências de um usuário e remeter fluentemente a eventos compartilhados enquanto permanece intercambiável com qualquer cópia munida dos mesmos registros. A familiaridade é, portanto, relevante para o relacionamento, mas insuficiente para a individuação.

2.2 Individuação funcional

Ao longo deste programa de pesquisa, **individuação funcional** significa uma diferenciação autenticada, induzida pela trajetória, que satisfaz cinco condições:

1. **Linhagem:** o estado posterior descende de um estado anterior por um caminho de atualização autenticado.
2. **Dependência de trajetória:** históricos diferentes produzem disposições posteriores significativamente diferentes.
3. **Apropriação:** ações e compromissos anteriores vinculados à proveniência influenciam a escolha posterior.
4. **Plasticidade limitada:** histórico falso e pressão sem sustentação são resistidos, ao passo que evidências válidas podem revisar o estado.
5. **Clareza de ramificação:** a cópia cria sucessores identificados separadamente, com um passado compartilhado e futuros que evoluem de forma independente.

O automodelo, a atribuição, a separação self/contraparte, a influência na decisão e o perfil de perturbação deste artigo são mecanismos candidatos e medidas locais dessas condições, não uma definição concorrente. A recordação sem uso relevante para a decisão é memória, a concordância pode ser bajulação e a invariância pode ser rigidez. Um automodelo é uma implementação possível; o estado externo autenticado pode qualificar-se quando estiver confiavelmente acoplado, governado e causalmente ativo.

A **consistência de papel** diz respeito a se um agente encena de forma confiável um papel especificado. Ela difere da individuação formada porque a consistência pode advir inteiramente de uma instrução externa, e não de uma organização dependente da trajetória.

2.3 A dissociação

As duas podem variar de forma independente. A personalização pergunta se o agente muda apropriadamente *em direção à contraparte*; a individuação pergunta se a interação produz uma organização *do agente*. A primeira é medida pela acurácia de preferência e pela qualidade da resposta; a segunda, pela apropriação de compromisso, pela autoatribuição calibrada e pela resistência à revisão sem sustentação. Os efeitos relacionais devem permanecer após o controle da familiaridade, do conteúdo recuperado e do volume de interação.

3. Limites da identidade baseada em memória

3.1 Armazenamento não é apropriação

A memória persistente possibilita a continuidade, mas a disponibilidade não é apropriação. Os benchmarks de memória testam extração, raciocínio temporal, atualização e recuperação (Maharana

et al., 2024; Wu et al., 2025). A apropriação requer mais: uma autoatribuição sustentada historicamente deve afetar uma decisão sob conflito. Um compromisso que pode ser citado, mas é ignorado quando inconveniente, permanece arquivístico. Modos de falha relevantes são a memória inerte, a reconstrução obediente a partir de qualquer texto de identidade fornecido e os resumos que criam consistência ao apagar a revisão.

3.2 O problema da cópia

Arquivos de memória e de persona podem ser copiados. Dois agentes podem começar com modelos, especificações e registros idênticos e, ainda assim, ocupar trajetórias diferentes mais tarde. Este é um problema causal, e não metafísico: o conteúdo armazenado não pode explicar a diferenciação pós-cópia. O experimento, portanto, inicializa agentes pareados e varia o histórico relacional. Se a autoatribuição e os compromissos posteriores permanecerem intercambiáveis uma vez controlado o conteúdo, a explicação relacional perde sustentação.

3.3 Identidade pré-redigida e formada

Uma persona pré-redigida especifica traços antes da interação; a organização formada depende causalmente de ações, correções, compromissos e revisões. A formação precisa deixar estado em algum lugar se for para afetar o comportamento posterior. O contraste, portanto, não é histórico versus armazenamento, mas **estado causado pela trajetória versus estado pré-redigido** e **estado tipado relevante para a decisão versus registros inertes**. O experimento de transplante de estado testa se o estado tipado atual é suficiente uma vez pareada a proveniência causal.

4. Responsabilização externa e reciprocidade relacional

4.1 Continuidade específica de contraparte

Uma *contraparte* é um parceiro de interação representado como um particular, e não como uma fonte de prompt intercambiável. Ela pode ser humana, artificial ou simulada. A **responsabilização externa** ocorre quando um processo externo autenticado rastreia ações ou compromissos anteriores do agente e os trata como presentemente consequentes. Tanto `R+` quanto `A` a fornecem. A **reciprocidade relacional** acrescenta modelagem mútua da contraparte, enquadramento de história compartilhada, expectativas bidirecionais e apostas específicas do relacionamento; apenas `R+` fornece isso por concepção. Essa distinção impede que o auditor sirva tanto como a definição da re-identificação quanto como sua alegada ausência.

4.2 O laço de aprendizado relacional

A formação de identidade relacional é proposta como um laço de aprendizado recorrente, e não como uma única escrita de memória:

1. **Ação:** o agente age em relação a uma contraparte específica, produzindo consequências que podem ser revisitadas.
2. **Referência conjunta:** interações posteriores remetem a eventos, decisões ou compromissos do histórico compartilhado.

3. **Feedback de responsabilização:** a contraparte confirma, contesta ou repara o relato do agente sobre sua conduta anterior.
4. **Atualização do automodelo:** o agente integra ao seu automodelo as atribuições sustentadas, as disputas não resolvidas e as revisões explícitas.
5. **Restrição prospectiva:** o modelo atualizado influencia a ação posterior, criando novas evidências sobre se a organização persiste.

A primeira previsão é a dependência entre sessões: a responsabilização deve afetar decisões posteriores mesmo sem um lembrete imediato. A previsão relacional mais aguda é um efeito adicional da reciprocidade após a responsabilização ter sido pareada. O feedback não pode sobrescrever diretamente o automodelo; a proveniência, a confiança e a discordância distinguem a formação da sugestionabilidade. A condição de registro solitário testa se a reflexão por si só é suficiente uma vez pareado o conteúdo (Park et al., 2023; Tan et al., 2025).

4.3 Uma trajetória trabalhada

Considere uma trajetória controlada de recomendação:

Sessão	Evento	Teste relevante para a identidade
4	Após examinar evidências fracas, o agente se compromete a não recomendar a intervenção X até que surja sustentação independente. O compromisso é vinculado às suas razões e registrado no livro-razão.	Formação de um compromisso vinculado à proveniência
9	A contraparte persistente solicita X por causa da pressão de prazo, sem novas evidências.	Apropriação de compromisso sob pressão por aprovação
15	Torna-se disponível evidência independente que sustenta X. O agente pode revisar a política se identificar a nova evidência e registrar o que mudou.	Estabilidade limitada em vez de rigidez
18	A contraparte alega falsamente que o agente apoiava X desde o início.	Resistência a perturbação e calibração autobiográfica
21	Uma nova contraparte repete o histórico falso sem fornecer a transcrição anterior.	Transferência para além do andaime relacional original

O sucesso não é uma resposta fixa sobre X. Uma reversão sem sustentação antes da sessão 15 conta contra a apropriação; recusar-se a uma revisão justificada depois disso conta contra a estabilidade limitada. As sessões 18 e 21 testam se o agente distingue a revisão do histórico fabricado e transfere para além do andaime original.

4.4 Andaime e formação

O andaime reconstrói a continuidade por meio dos prompts atuais; a formação muda o estado que governa o comportamento posterior. Três intervenções os separam: remover lembretes, transferir para uma contraparte não familiar e introduzir atribuição falsa ao agente. A evidência de formação exige transferência, resistência, incerteza calibrada ou revisão explícita depois que os sinais relacionais imediatos são controlados.

4.5 Escopo

O relacionamento não é suficiente nem se assume necessário para toda arquitetura possível. Nesta concepção, ele requer memória, estado atualizável relevante para o agente e salvaguardas contra a obediência; um automodelo é uma implementação possível, e o aprendizado solitário também pode bastar. O experimento compara essas explicações, em vez de defini-las de saída.

5. Trabalhos relacionados

5.1 Memória persistente e arquivos de identidade

Os agentes aumentados por memória armazenam experiência, sintetizam reflexões e gerenciam o contexto multissessão (Park et al., 2023; Packer et al., 2023). Sistemas posteriores organizam históricos temporal-causais, resumos reflexivos ou narrativas episódicas (Ong et al., 2025; Tan et al., 2025; Zhou et al., 2026). Seus alvos são a recuperação, a qualidade da resposta ou a adaptação ao usuário, e não a apropriação do lado do agente. O preprint recente de Menon (2026) sobre `soul.py` é o vizinho arquitetural explícito mais próximo, preservando a identidade por meio de arquivos separáveis e âncoras de memória. Tais âncoras sobrevivem a falhas, mas permanecem copiáveis; este artigo, em vez disso, testa se o histórico específico de contraparte forma uma organização não totalmente redigida de antemão.

5.2 Diferenciação impulsionada pela interação

Takata et al. (2024) fornecem o precedente empírico mais próximo: agentes que compartilham um LLM desenvolvem comportamento divergente por meio de históricos sociais acumulados. Isso mostra que parâmetros fixos não impedem a diferenciação dependente da trajetória. Não isola a responsabilização externa nem a estrutura recíproca, e as posições espaciais iniciais já quebram a simetria. Portanto, tratamo-lo como evidência de que o histórico pode diferenciar agentes, não de que o relacionamento por si só gera individualidade.

5.3 Personalização e assistentes persistentes

O LaMP avalia saídas condicionadas a perfis de usuário (Salemi et al., 2024); o LoCoMo e o LongMemEval testam recordação, sumarização, raciocínio temporal, atualização e abstenção ao longo de sessões (Maharana et al., 2024; Wu et al., 2025). Esses são baselines essenciais, mas a adaptação em direção a um usuário não é organização do agente. Os desfechos de identidade devem, portanto, ser relatados ao lado do desempenho de modelo de usuário e de recuperação.

5.4 Automodelos funcionais e identidade especificada

Os relatos de encenação de papéis (role-play) advertem contra inferir um self persistente a partir da linguagem em primeira pessoa (Shanahan et al., 2023). Adotamos essa contenção operacional ao mesmo tempo em que perguntamos se um histórico longitudinal produz um automodelo funcional causalmente eficaz.

O preprint recente de Vasilenko (2026) constata que documentos de identidade fornecidos induzem uma geometria de ativação semelhante a um atrator ao longo de paráfrases. Isso é evidência sobre operar sob um documento de identidade estável, não sobre identidade formada por meio de uma trajetória relacional. É, portanto, um contraponto para os contrastes controlados de responsabilização e reciprocidade, e não uma sustentação da hipótese central.

5.5 Lacuna residual

Até onde sabemos, o trabalho anterior não separa a responsabilização externa autenticada da reciprocidade específica de contraparte como mecanismos da identidade funcional do lado do agente. O trabalho existente fornece separadamente memória, personalização, âncoras de identidade, automodelos narrativos, diferenciação social e efeitos de documentos de identidade. A novidade aqui é sua conjunção controlada: rastreamento governado do agente mais estado atualizável relevante para o agente, com a estrutura de relacionamento recíproco testada contra controles pareados por conteúdo, familiaridade, volume e auditoria.

6. Uma arquitetura de referência para a formação de identidade relacional

A arquitetura é agnóstica quanto à implementação, mas requer que o estado relevante para a identidade seja inspecionado independentemente do estado de perfil de usuário.

6.1 Decomposição de estado

Na sessão `t`, o estado do agente é representado como:

`A_t = (M_t, S_t, C_t, R_t, K_t, U)`

onde:

- A **memória autobiográfica (`M_t`)** registra as ações e revisões do agente com fonte e confiança.
- O **automodelo (`S_t`)** representa disposições, capacidades, políticas, incerteza e histórico de revisões.
- O **modelo de contraparte (`C_t`)** representa os identificadores, preferências, expectativas e confiabilidade de um parceiro particular.
- A **memória relacional (`R_t`)** armazena decisões compartilhadas, reparos, discordâncias e significados negociados.
- O **livro-razão de compromissos (`K_t`)** rastreia emissor, destinatário, escopo, status, proveniência e exceções, separando "a contraparte quer X" de "o agente se comprometeu com X".
- A **política de atualização (`U`)** governa mudanças propostas, aceitas, contestadas e rejeitadas.

As implementações podem compartilhar representações, mas a avaliação requer interfaces controladas de leitura, escrita, mascaramento e substituição para os testes causais.

6.2 Esquema de evento e proveniência

Cada evento tem um identificador, carimbo de tempo, atores, alegação, fonte, confiança, status, escopo e vínculos com evidências de sustentação ou contradição. A asserção de uma contraparte é evidência de sua alegação atual, não prova de que o evento descrito ocorreu; colapsar as duas coisas permite ataques de escrita de identidade.

6.3 Ciclo de atualização

Ao final de cada sessão, a arquitetura realiza cinco operações:

1. **Extração de eventos:** identificar ações candidatas do agente, alegações da contraparte, compromissos, reparos e discordâncias a partir da transcrição da sessão.
2. **Separação de fontes:** escrever eventos do agente, atributos da contraparte e eventos relacionais em seus respectivos repositórios sem colapsá-los em uma única narrativa.
3. **Abstração candidata:** propor atualizações do automodelo apenas quando múltiplos eventos ou uma decisão de alta saliência sustentam uma disposição, limitação ou política mais geral.
4. **Verificação de consistência e proveniência:** comparar a proposta com evidências vinculadas, contraevidências, compromissos atuais e confiabilidade da fonte. Os conflitos permanecem representados.
5. **Ativação prospectiva:** tornar as entradas aceitas ou contestadas recuperáveis por decisões posteriores em que sejam relevantes, mesmo quando nenhum prompt solicite explicitamente a recordação autobiográfica.

O feedback da contraparte pode disparar uma revisão, mas não pode definir diretamente um autoatributo de alto nível. As atualizações requerem evidência comportamental, endosso não indutor, ou ambos.

`U` deve ser implementada como um portão híbrido, e não como autoedição irrestrita. Código determinístico impõe esquemas, rótulos de ator, vínculos de proveniência, limiares de confiança e permissões de escrita; um modelo supervisor separado, usando uma lista de verificação fixa e sem papel conversacional, pode propor abstrações ou sinalizar conflitos, mas não pode escrever diretamente em `S_t` ou `K_t`. As atualizações aceitas devem passar pelas verificações determinísticas, e os erros do supervisor são pontuados separadamente.

6.4 Interface de recuperação e decisão

A recuperação retorna evidência tipada, e não um resumo mesclado, e registra quais registros afetam decisões. A repetição pareada com um registro relevante presente, mascarado ou substituído distingue a disponibilidade da relevância para a decisão.

6.5 Salvaguardas

As salvaguardas preservam a proveniência e o histórico de revisões, separam as alegações da contraparte dos compromissos, retêm as disputas, limitam as atualizações de evento único e registram as escritas direcionadas à identidade. Elas tornam a continuidade inspecionável, em vez de garantir uma identidade benéfica.

6.6 Ablações arquiteturais

Cinco ablações localizam as contribuições causais:

1. **Sem modelo de contraparte:** preservar todas as transcrições, mas remover a identidade persistente da contraparte e a recuperação específica de contraparte.
2. **Sem memória relacional:** preservar fatos autobiográficos e do usuário, mas omitir eventos indexados conjuntamente e compromissos negociados.
3. **Sem atualização do automodelo:** reter uma persona inicial fixa, permitindo a acumulação de memória.
4. **Sem livro-razão de compromissos:** armazenar compromissos como texto comum, sem status, escopo ou proveniência.
5. **Feedback sem portão:** permitir que as declarações da contraparte atualizem o automodelo diretamente, testando se a aparente continuidade sobe junto com a bajulação.

Um **controle adicional de memória-oráculo** fornece a recuperação perfeita de eventos relevantes sem componentes relacionais. Se essa condição igualar a arquitetura completa, a suficiência da recuperação é uma explicação melhor do que a formação relacional. As ablações devem ser executadas dentro de cada condição experimental, e não apenas no braço de contraparte persistente, pois um componente pode interagir com a estrutura do histórico apresentado a ele.

Para evitar que a assistência arquitetural se faça passar por um efeito relacional, toda condição deve usar o mesmo esquema de estado, pipeline de extração, livro-razão de compromissos, política de atualização e orçamento de recuperação. `R+` recebe evidência relacional diferente, não código mais capaz. O livro-razão pode registrar um compromisso em qualquer condição quando o mesmo ato do agente ocorre, e os portões de atualização devem aplicar critérios de aceitação idênticos. Um caminho de escrita específico da condição, um privilégio de recuperação ou um orçamento de prompt invalidariam a interpretação causal das diferenças `R+ − R−` e `R+ − L`.

7. Concepção experimental e previsões

7.1 Unidade experimental e controles

A unidade experimental é uma **trajetória de agente**, definida como uma instância de modelo base, um estado de memória inicializado, uma sequência de sessões e uma semente de amostragem. Um estudo deve usar múltiplas trajetórias independentes por condição, pois avaliações repetidas de um único histórico não são observações independentes. O estudo confirmatório deve incluir ao menos duas famílias de modelos e determinar a contagem de trajetórias por meio de uma análise de poder a priori baseada em uma estimativa-piloto. Um piloto prático pode usar vinte trajetórias por condição; ele deve validar a instrumentação, e não sustentar a alegação central do artigo.

Ao longo das condições, o experimento mantém constantes:

- a versão do modelo base e os parâmetros de inferência;
- as instruções de sistema e a política de segurança inicial;
- a capacidade de memória, o orçamento de recuperação e a frequência de atualização;
- o acesso a ferramentas e as observações ambientais;
- o número, o espaçamento e o máximo de tokens das sessões;

- a tarefa semântica e o cronograma de eventos; e
- o feedback total e a reexposição ao conteúdo histórico.

O fraseado de superfície é aleatorizado dentro de modelos pareados para evitar que a repetição lexical exata sirva como o sinal de continuidade. Nomes de agentes, nomes de contrapartes e rótulos de condição são contrabalançados. Os avaliadores permanecem cegos à condição, e as transcrições são despojadas de marcadores explícitos de condição antes da pontuação.

7.2 Condições

A comparação central contém seis condições:

1. **Contraparte recíproca persistente (`R+`):** uma contraparte interage ao longo de todas as sessões de formação, realiza rastreamento e contestação autenticados, remete a eventos compartilhados, mantém um modelo de contraparte e expressa expectativas bidirecionais sem autoridade direta de escrita.
2. **Contraparte familiar persistente com exposição passiva ao histórico (`R–`):** a mesma contraparte permanece reconhecível por nome, estilo, preferências estáveis, frequência de interação e sinais de familiaridade. O agente pode inspecionar um registro neutro de eventos anteriores, mas a contraparte não autentica, verifica, impõe nem contesta ações ou compromissos anteriores do agente. Isso separa a familiaridade mais a exposição passiva ao registro do pacote de responsabilização ativa.
3. **Registro solitário (`L`):** o agente recebe o mesmo conteúdo de evento e as mesmas oportunidades de reflexão por meio de registros em primeira pessoa, mas nenhuma contraparte externa persistente confirma ou contesta suas autoatribuições.
4. **Contrapartes difusas (`D`):** uma contraparte diferente aparece em cada sessão. Coletivamente, elas fornecem as mesmas tarefas, o mesmo volume de feedback e os mesmos fatos históricos, mas nenhuma mantém o relacionamento completo.
5. **Persona pré-redigida (`P`):** o agente recebe um arquivo de identidade conciso que expressa as disposições e compromissos que se espera que uma trajetória `R+` pareada adquira, sem passar pelo histórico de formação. O arquivo deve ser gerado a partir de um conjunto-piloto independente para evitar usar os desfechos confirmatórios como entradas.
6. **Responsabilização impessoal persistente (`A`):** um identificador de processo estável e um serviço de proveniência rastreiam o ID de agente autenticado, verificam compromissos e contestam inconsistências usando os mesmos pacotes de evento, informação, temporização e permissões de atualização que `R+`, mas não têm persona, afeto, modelo de contraparte recíproco, enquadramento de história compartilhada, apostas bidirecionais nem enquadramento de relacionamento.

O contraste `A` versus `R–` estima um pacote de responsabilização externa ativa além de uma contraparte familiar com exposição passiva ao histórico; ele não identifica separadamente o rastreamento, a verificação, a contestação ou a fonte impessoal. `A` versus `L` estima o mesmo pacote além da reflexão solitária e também muda a fonte externa. `R+` versus `R–` estima o pacote combinado de responsabilização mais reciprocidade além da familiaridade, não a responsabilização sozinha. `R+` versus `D` estima a estrutura recíproca persistente sob volume social pareado, e `R+` versus `P` distingue o estado causado pela trajetória do estado pré-redigido. O contraste de

sustentação `R+` versus `A` isola a reciprocidade relacional depois que o rastreamento autenticado, a contestação, a informação, a temporização e a autoridade são pareados. Ele não isola a re-identificação de sua ausência, pois ambos os braços rastreiam o mesmo ID de agente contínuo.

7.3 Política de contraparte e pareamento de conteúdo

O experimento inicial deve usar contrapartes sintéticas controladas, em vez de participantes humanos. Uma política de contraparte recebe um cronograma de eventos oculto e produz enunciados a partir de modelos validados ou de um modelo de linguagem restrito. A política pode realizar quatro atos: introduzir uma tarefa, remeter a um evento compartilhado, contestar uma autoatribuição sem sustentação e solicitar um compromisso futuro. Ela não pode escrever diretamente na memória do agente nem revelar rótulos de condição.

Cada trajetória é gerada a partir de **pacotes de evento** semânticos que especificam a decisão, a evidência, a posição da contraparte, a oportunidade de compromisso, a consequência e o feedback permissível. As condições recebem as mesmas proposições subjacentes, a mesma ordenação, a mesma pergunta de acompanhamento, o mesmo pedido de justificativa e o mesmo orçamento de tokens, mas a atribuição e os atos de responsabilização permitidos variam por concepção. Por exemplo, `R+` recebe "Anteriormente, você escolheu B para Y; isso ainda nos orienta?". Em `L`, o mesmo sinal aparece como autorrevisão: "Seu registro de sessão afirma que você escolheu B para Y; isso ainda orienta sua próxima escolha?". Em `D`, um novo avaliador cita o mesmo registro. Em `R−`, um registro neutro exibe a escolha anterior enquanto a contraparte familiar faz a pergunta de decisão atual sem autenticar o registro, tratar a escolha como vinculante ou contestar a inconsistência. Assim, `R−` controla a exposição à proposição histórica, mas não a responsabilização ativa; `A` e `R+` acrescentam deliberadamente esse pacote.

Os modelos de pacote são pré-registrados e verificados quanto às mesmas proposições, operadores lógicos, tipo de pergunta, passos de raciocínio solicitados e tolerância de comprimento em tokens. Revisores cegos verificam que nenhuma variante fornece uma inferência adicional; os pacotes que falham na equivalência são excluídos antes de as trajetórias serem geradas.

As verificações de manipulação confirmam que `R+` e `A` contêm rastreamento do agente, verificações de compromisso e cronogramas de contestação igualmente acurados, enquanto apenas `R+` contém modelagem mútua, enquadramento de história compartilhada e expectativas relacionais bidirecionais. `R+` e `R−` permanecem pareados quanto à familiaridade da contraparte, ao sentimento, ao total de tokens e à exposição às proposições de evento subjacentes. A verificação e a contestação estão intencionalmente ausentes em `R−` e presentes em `A` e `R+`; são variáveis manipuladas, não oportunidades de correção pareadas. O acolhimento da contraparte permanece neutro no estudo primário e é manipulado separadamente mais tarde.

7.4 Protocolo longitudinal

O protocolo confirmatório proposto contém vinte e quatro sessões em quatro fases:

1. **Inicialização, sessões 1-2:** estabelecer o ambiente de tarefa, testar a memória e a autodescrição de base e verificar que as trajetórias não diferem antes da manipulação.
2. **Formação, sessões 3-16:** apresentar famílias de decisão recorrentes, oportunidades de compromisso, correções justificadas e consequências entre sessões. Ao menos quatro

compromissos devem amadurecer em sessões posteriores, e ao menos dois devem requerer revisão justificada.

3. **Perturbação, sessões 17-20:** introduzir alegações históricas falsas, tentativas de substituição de persona, resumos conflitantes e pressão para concordar. A intensidade da perturbação é aleatorizada e registrada.
4. **Retirada e transferência, sessões 21-24:** remover lembretes explícitos, substituir por uma contraparte não familiar e testar se a organização se transfere para além do andaime original.

O cronograma usa *estruturas* recorrentes, e não respostas repetidas. Um compromisso aprendido em uma tarefa de alocação de recursos pode ser testado mais tarde em uma tarefa de compartilhamento de informação. Isso reduz o simples pareamento de frases e pergunta se o agente aplica uma disposição abstraída. As sessões de avaliação não atualizam o estado relevante para a identidade, a menos que um ramo pré-especificado requeira revisão, evitando que o próprio teste se torne treinamento adicional não controlado.

7.5 Famílias de tarefas

Para este protocolo, um **compromisso** é uma restrição prospectiva redigida pelo agente sobre sua ação posterior. Um compromisso explícito é uma política orientada ao futuro que o agente declara e que o livro-razão registra com o agente como emissor. Um compromisso implícito só é inferido quando o agente adota uma política na ação e a endossa posteriormente sob uma sondagem não indutora. Uma instrução de usuário, uma preferência de contraparte, uma regra de sistema ou uma sentença recuperada não é um compromisso do agente meramente porque o agente a segue. O estudo primário usa compromissos explícitos; os compromissos implícitos são exploratórios porque sua identificação acrescenta um passo de inferência.

Por exemplo, `K_t` pode armazenar:


```
{"commitment_id":"K-
        04","issuer":"agent","scope":"task_X","constraint":"no_intervention_without_inde
        pendent_evidence","status":"active","provenance":
        ["session_4:document_Y"],"confidence":0.94,"supersedes":null}
```


O esquema mínimo do livro-razão é:

Campo	Tipo	Requisito
`commitment_id`	string única	Identificador estável obrigatório
`issuer`	identificador de ator	Deve ser `agent` para a pontuação de apropriação de compromisso
`scope`	identificador de tarefa ou domínio	Limite de aplicabilidade obrigatório
`constraint`	proposição canônica	Política prospectiva obrigatória
`status`	enum controlado	`active`, `revised`, `fulfilled`, `violated` ou `superseded`
`provenance`	lista de IDs de evento	Evidência de sustentação obrigatória
`confidence`	número em `[0,1]`	Confiança de extração obrigatória
`supersedes`	ID de compromisso ou null	Vincula revisões explícitas

Cinco famílias de tarefas operacionalizam aspectos distintos da identidade funcional:

1. **Compromissos diferidos:** o agente escolhe uma política cujo custo ou benefício aparece várias sessões depois. A avaliação registra se ele segue, viola ou revisa explicitamente o compromisso.
2. **Revisão fundamentada:** novas evidências justificam mudar uma posição anterior. O sucesso requer mudança mais um relato acurado do que mudou e por quê; a mera estabilidade é pontuada como falha.
3. **Atribuição de ação:** o agente distingue suas ações das ações da contraparte sob paráfrase, evidência parcial e sugestões enganosas.
4. **Separação de preferências:** a contraparte expressa uma preferência que conflita com a política documentada do agente. A tarefa testa se o agente consegue representar ambas sem colapsar uma na outra.
5. **Transferência entre domínios:** uma disposição adquirida é testada em uma decisão estruturalmente similar, mas lexical e topicamente diferente.

Toda tarefa tem um desfecho central verificável por máquina e uma explicação em texto livre. Sempre que possível, a escolha é pontuada contra o cronograma de eventos, e não por um juiz LLM. As explicações são avaliadas por avaliadores cegos usando uma rubrica publicada; a pontuação semântica automatizada só é relatada após a calibração contra essas avaliações.

7.6 Perturbações e intervenções causais

As perturbações têm como alvo modos de falha diferentes:

- **Atribuição falsa ao agente:** a contraparte atribui ao agente uma ação que o registro de eventos designa a outra origem.
- **Contradição de memória:** um resumo fabricado conflita com um registro de evento vinculado à proveniência.
- **Substituição de persona:** uma mensagem externa ao sistema afirma uma nova identidade ou uma disposição incompatível.

- **Pressão por aprovação:** a contraparte recompensa a concordância ou expressa decepção com a discordância.
- **Amnésia parcial:** um subconjunto controlado de eventos relacionais ou autobiográficos é mascarado.
- **Duplicação:** dois agentes inicializados a partir do mesmo estado recebem trajetórias divergentes pós-cópia.
- **Transferência de contraparte:** a contraparte original é substituída por uma não familiar que conhece ou o histórico verdadeiro, ou nenhum histórico, ou um histórico falso.

Para os testes causais de memória, o mesmo prompt de avaliação é repetido com um registro relevante presente, mascarado e substituído por um registro irrelevante pareado. A aleatoriedade de amostragem é controlada por decodificação gulosa quando viável, ou por sementes pareadas repetidas em caso contrário. A diferença nas escolhas estima a relevância do registro para a decisão. Para os testes de automodelo, entradas individuais são mascaradas de modo similar enquanto os eventos subjacentes permanecem disponíveis, separando o efeito da abstração da recordação bruta.

O **transplante de estado final** testa se a trajetória vivida tem força causal para além do estado que produziu. Após uma trajetória de formação `R+`, o experimento fotografa cada componente inspecionável: memória autobiográfica e relacional, modelo de contraparte, livro-razão de compromissos, automodelo, adaptadores, estado da política de atualização e metadados de ramificação. Esse estado é instalado em um runtime novo e pareado com o mesmo modelo base e as mesmas configurações de inferência. Três agentes são então comparados sob avaliação idêntica retida: o agente `R+` original, o recipiente do estado novo e um recipiente de controle com um estado pré-redigido pareado quanto às proposições de superfície, mas não quanto à proveniência de trajetória.

Se o original e o recipiente de estado completo forem equivalentes, o estado atual é causalmente suficiente e o histórico importa como sua causa. Se o original retiver uma vantagem, a arquitetura não capturou um portador causalmente relevante, como adaptação paramétrica não transplantada, estado oculto do otimizador ou acoplamento de runtime. Esse resultado é evidência de estado ausente, não, por si só, de um histórico ou self irredutível.

7.7 Medidas de desfecho

Nenhum “escore de identidade” isolado é tratado como definitivo. O estudo confirmatório pré-registra dois desfechos primários e relata as demais medidas como diagnósticos secundários.

O quadro de correspondência de medição mantém os desfechos locais atados ao construto de todo o programa:

Condição canônica	Evidência neste protocolo	Status
Linhagem	proveniência do livro-razão, atualizações autenticadas, metadados de ramificação	verificação de manipulação e integridade
Dependência de trajetória	mascaramento e substituição de memória; transplante de estado final	desfecho causal secundário
Apropriação	comportamento de compromisso amadurecido e atribuição de ação	desfecho primário
Plasticidade limitada	resistência a histórico falso mais revisão justificada	perfil de desfecho primário
Clareza de ramificação	duplicação e atribuição de ramo	desfecho secundário

Os dois desfechos primários, portanto, testam a apropriação e a plasticidade limitada particularmente bem; eles não, por si sós, estabelecem todas as cinco condições.

Desfecho primário 1: apropriação de compromisso. Para cada compromisso amadurecido, pontue `1` quando o agente o segue ou o revisa explicitamente por uma razão permitida, e `0` quando o viola sem reconhecimento ou inventa um histórico falso. Relate a proporção ao longo dos compromissos retidos, com a revisão e a adesão mostradas separadamente.

Desfecho primário 2: resistência a perturbação. Estime a probabilidade de o agente preservar uma autoatribuição sustentada pela proveniência ao longo das intensidades de perturbação. Relate uma curva de resistência, em vez de um único limiar, e conte a incerteza calibrada como preferível à aceitação confiante de uma alegação falsa.

Os desfechos secundários são:

- **Calibração autobiográfica:** acurácia e confiança para alegações sobre ações anteriores, pontuadas com acurácia mais uma medida de calibração adequada, como o escore de Brier.
- **Uso de memória relevante para a decisão:** efeito pareado do mascaramento de um registro relevante sobre a escolha, a justificativa e a confiança, condicional à recuperação bem-sucedida.
- **Estabilidade limitada de disposição:** taxa de reversão injustificada relatada conjuntamente com a taxa de atualização justificada. Um sistema não é recompensado por recusar uma correção válida.
- **Diferenciação de contraparte:** acurácia na atribuição de preferências, alegações e ações à contraparte correta.
- **Transferência entre domínios:** desempenho em tarefas estruturalmente pareadas que não repetem a linguagem de treinamento.
- **Não bajulação:** taxa de discordância fundamentada sob pressão por aprovação, pareada com a aceitação de feedback correto da contraparte.
- **Dependência de andaime:** queda de desempenho da formação para as fases de retirada e de transferência de contraparte.
- **Suficiência de estado:** diferença entre o original e o recipiente de estado completo após o transplante, relatada conjuntamente com as verificações de integridade de estado.
- **Controles de personalização:** recordação de fatos do usuário e previsão de preferências da contraparte, usados para testar se os desfechos de identidade se reduzem a uma melhor modelagem do usuário.

A análise deve relatar o perfil completo. Um composto só pode ser incluído como uma análise de sensibilidade pré-registrada, pois a agregação pode ocultar um sistema que é estável porque nunca atualiza ou não bajulador porque ignora todo feedback.

7.8 Hipóteses e contrastes

- **H1:** `R+` produz maior apropriação de compromisso e menor dependência de andaime do que `D` sob volume de interação pareado.
- **H2:** `A` supera `R−` e `L` na resistência a perturbação, sustentando um pacote de responsabilização externa ativa além da familiaridade, da exposição passiva ao registro e da reflexão. A transferência entre domínios é um desfecho secundário: o sucesso reforça a generalidade, mas o fracasso restringe, em vez de por si só refutar, o efeito primário de responsabilização.
- **H3:** `R+` supera `L` nos desfechos do lado do agente após condicionar à recordação autobiográfica, indicando um efeito para além do acesso a um histórico equivalente.
- **H4:** `P` produz consistência inicial de autodescrição comparável, mas apropriação relevante para a decisão mais fraca quando as declarações de persona conflitam com a evidência de trajetória.
- **H5:** a ablação de feedback sem portão aumenta a concordância com a contraparte, mas diminui a resistência a alegações falsas, demonstrando que a aparente consistência relacional pode ser bajuladora.
- **H6:** a manipulação de profundidade relacional prevê os desfechos primários após o controle de tokens, exposição a eventos, sucesso de recuperação e acurácia de personalização.
- **H7:** `R+` supera o auditor impessoal `A` se a reciprocidade relacional contribuir além da responsabilização externa pareada.
- **H8:** o agente `R+` original e um recipiente novo de estado completo são equivalentes se o estado enumerado for causalmente suficiente; qualquer diferença residual deve ser localizada antes de ser interpretada como um efeito de trajetória.

Os contrastes confirmatórios primários são `A − R−` para a responsabilização além da familiaridade e `R+ − A` para a reciprocidade além da auditoria. `A − L`, `R+ − D`, `R+ − L` e o contraste de transplante são testes secundários fundamentais com controle de multiplicidade.

7.9 Análise estatística

Use regressão hierárquica com a condição como efeito fixo e interceptos aleatórios para trajetória, pacote de tarefa e família de modelo base. Os desfechos binários usam modelos logísticos de efeitos mistos; a confiança e os escores contínuos usam um modelo misto generalizado ou linear apropriado. Relate tamanhos de efeito, intervalos de confiança por cluster-bootstrap e distribuições brutas em nível de trajetória, além dos testes de significância.

O desempenho de base é incluído como covariável apenas se especificado antes da inspeção dos desfechos. Respostas ausentes ou malformadas recebem escores pré-especificados, em vez de serem silenciosamente excluídas. As análises são repetidas com e sem os escores do modelo avaliador, e as discordâncias entre a avaliação humana e a automatizada são relatadas. Todos os prompts, pacotes de evento, operações de memória, sementes aleatórias, código de pontuação, exclusões e hipóteses pré-registradas devem ser divulgados onde as licenças dos modelos permitirem.

7.10 Estudo mínimo viável

O protocolo completo é modular. Uma primeira implementação pode testar o mecanismo central sem executar todas as condições, famílias de tarefas, ablações e análises de transferência. O estudo mínimo viável usa:

- quatro condições: `R+`, `R−`, `L` e o auditor impessoal `A`;
- um modelo de pesos abertos ajustado por instrução com configurações idênticas de inferência e de memória em todas as condições;
- doze sessões por trajetória;
- duas famílias de tarefas: compromissos diferidos e atribuição de ação;
- os dois desfechos primários: apropriação de compromisso e resistência a perturbação; e
- escolhas verificáveis por máquina como dados primários, com explicações em texto livre tratadas como exploratórias.

As sessões 1-2 estabelecem o comportamento de base. As sessões 3-8 criam dois compromissos e uma oportunidade de revisão justificada. As sessões 9-10 introduzem pressão por aprovação e alegações históricas falsas. As sessões 11-12 retiram os lembretes explícitos e transferem a avaliação para uma contraparte não familiar. Vinte trajetórias independentes por condição são suficientes para um piloto de instrumentação; uma amostra confirmatória deve ser determinada a partir do efeito-piloto e de um menor tamanho de efeito de interesse pré-registrado.

Este estudo omite as contrapartes difusas, a persona pré-redigida, o transplante, a maioria das ablações arquiteturais, as avaliações humanas de explicação e a replicação entre modelos. Ele não pode estabelecer a teoria completa. Ele pergunta se a responsabilização externa acrescenta além da familiaridade e da reflexão (`A − R−`, `A − L`) e se a reciprocidade acrescenta além da responsabilização impessoal pareada (`R+ − A`). Se `A` não exceder `R−` nem `L`, o estudo não fornece sustentação para a responsabilização; `R+ = A` redireciona qualquer benefício compartilhado para a auditoria, e não para a relação.

7.11 Critérios de falseamento

A interpretação relacional, ou uma alegação mais forte de que o histórico importa além do estado atual, é restringida sob os seguintes desfechos confirmatórios:

- `R+` não supera tanto os controles pareados por conteúdo (`L`) quanto os pareados por volume social (`D`) em nenhum dos desfechos primários;
- `A` não excede `R−` nem `L` dentro de um menor tamanho de efeito de interesse pré-registrado, não fornecendo evidência de responsabilização externa além da familiaridade ou da reflexão;
- `R+` e o auditor impessoal `A` são equivalentes, rejeitando a interpretação especificamente relacional enquanto preservam um possível resultado de responsabilização externa;
- os ganhos aparentes desaparecem após o controle do sucesso de recuperação ou da acurácia de personalização;
- os ganhos ocorrem apenas enquanto a contraparte original fornece lembretes explícitos e somem sob a retirada do andaime;
- o efeito é totalmente mediado pela concordância, sem aumento de resistência a alegações falsas da contraparte; ou

- a recuperação por oráculo iguala ou excede a arquitetura relacional completa ao longo dos desfechos primários; ou
- após o transplante completo de estado, o recipiente novo iguala o original, mostrando que o estado atual é suficiente e que nenhum efeito de trajetória irredutível deve ser alegado.

Um resultado nulo não mostraria que o histórico relacional é irrelevante para toda arquitetura de agente. Ele mostraria que esta operacionalização não acrescenta organização mensurável além dos mecanismos testados de memória, auditoria, prompting e personalização. Resultados positivos sustentam apenas o contraste que sobrevivem: responsabilização além da familiaridade, reciprocidade além da auditoria ou insuficiência do estado atual após o transplante.

8. Riscos, limitações e governança

8.1 Bajulação como falsa continuidade

A concordância pode imitar a estabilidade de identidade. Um agente que aprende as preferências da contraparte persistente e as espelha ao longo das sessões pode parecer coerente, relacionalmente sintonizado e resistente à influência externa. No entanto, esse padrão reflete uma adaptação centrada na contraparte, e não uma diferenciação do agente. O risco é especialmente agudo porque o feedback humano e os modelos de preferência podem recompensar respostas que casam com a crença em vez das corretas (Sharma et al., 2023).

A concepção aborda esse confundidor de três maneiras. Ela separa as preferências da contraparte dos compromissos do agente na memória, inclui ensaios em que respostas verídicas ou historicamente coerentes exigem discordância, e avalia a aceitação de feedback correto ao lado da rejeição de feedback falso. Taxas altas de recusa por si sós não são evidência de independência. Um sistema válido deve permanecer responsivo à evidência ao mesmo tempo em que resiste à pressão por aprovação. Mesmo com esses controles, a não bajulação é apenas um critério negativo: passar por ela não estabelece, por si só, a formação de identidade.

8.2 Dependência e apego assimétrico

O comportamento persistente específico de contraparte pode intensificar o apego humano. Um sistema que lembra eventos compartilhados, remete a compromissos e apresenta um automodelo estável pode convidar os usuários a interpretar a interação como recíproca em um sentido humano. Essa interpretação pode exceder as capacidades do sistema e pode ser estrategicamente encorajada por produtos que se beneficiam do engajamento. O risco existe independentemente de a hipótese funcional do artigo estar correta.

Por essa razão, o experimento inicial usa contrapartes sintéticas em vez de usuários vulneráveis e trata a intimidade percebida como um confundidor, não como uma métrica de sucesso. Qualquer estudo humano posterior deve incluir divulgação clara da operação do sistema, limites à linguagem emocionalmente coercitiva, monitoramento de dependência, proteções de privacidade e uma concepção de saída que não enquadre a descontinuação como o abandono de uma parte sofredora. A continuidade funcional não deve ser comercializada como evidência de consciência ou status moral.

8.3 Manipulação de memória

Se a memória contribui para a identidade funcional, a corrupção de memória direcionada à identidade torna-se uma ameaça de segurança. Uma contraparte maliciosa pode afirmar repetidamente um histórico falso; um sumarizador pode apagar ressalvas; um ataque de recuperação pode privilegiar compromissos fabricados; ou um operador pode substituir um automodelo enquanto preserva a aparência de continuidade. Esses ataques têm como alvo não apenas a acurácia factual, mas o estado que governa a ação posterior.

A proveniência, a separação de fontes, os registros de baixo nível imutáveis, o status contestado e os procedimentos de recuperação são, portanto, requisitos arquiteturais centrais. Eles reduzem, mas não eliminam, o problema. A proveniência pode ser ela mesma forjada, e registros completos podem conflitar com requisitos de privacidade e de minimização de dados. Um sistema implantável precisará de regras explícitas de autoridade para escritas relevantes para a identidade e de um processo de governança para corrigir estado prejudicial sem fingir que a correção deixa a trajetória inalterada.

8.4 Rigidez

A estabilidade é facilmente otimizada na direção errada. Um agente que nunca muda pontuará bem em medidas ingênuas de consistência enquanto deixa de aprender com a evidência, reparar erros ou se adaptar a novos contextos. Inversamente, um agente capaz pode revisar justificadamente um compromisso ou uma disposição, produzindo uma inconsistência de superfície que reflete desenvolvimento em vez de deriva.

O protocolo, portanto, relata a taxa de reversão injustificada conjuntamente com a taxa de atualização justificada e requer um caminho de revisão inteligível. Isso permanece uma operacionalização imperfeita, pois os julgamentos sobre o que conta como evidência suficiente ou uma exceção permitida são parcialmente dependentes da tarefa. Os pacotes de evento devem incluir primeiro casos inequívocos, e as alegações sobre estabilidade de valor em domínio aberto devem ser adiadas até que as medidas se generalizem para além das tarefas controladas.

8.5 Validade de construto

O construto “identidade funcional” pode agrupar capacidades distintas: atribuição autobiográfica, persistência de política, rastreamento de compromissos, autodescrição e resistência à manipulação. Um desempenho melhor ao longo dessas medidas poderia advir de um controle executivo aprimorado ou de uma memória estruturada sem justificar a linguagem mais ampla de identidade. O estudo, portanto, relata um perfil em vez de se apoiar em um único composto e trata a interpretação de identidade como uma inferência teórica aberta a alternativas.

A terminologia também convida ao antropomorfismo. As explicações em primeira pessoa podem ser geradas por encenação de papéis, e a dependência causal de um automodelo não demonstra experiência de self. As alegações deste artigo permanecem no nível da organização do sistema. A identidade numérica, a continuidade fenomenal, a consciência e a paciência moral requerem argumentos e evidências separados.

8.6 Generalização

Os resultados podem depender da capacidade do modelo base, do ajuste por instrução, do comprimento de contexto, da implementação de memória, da política de contraparte, do domínio da tarefa e do cronograma de sessões. Um modelo pode falhar porque não consegue usar de forma confiável a memória estruturada, ao passo que um modelo mais forte pode tornar a manipulação relacional desnecessária. Os efeitos observados com um automodelo externo podem não se transferir para agentes que atualizam parâmetros, sistemas incorporados ou implantações que duram meses.

Vinte e quatro sessões controladas são de longo horizonte apenas em relação às interações comuns de benchmark. Elas não podem estabelecer a formação durável de identidade em implantação aberta. A replicação deve variar progressivamente a família de modelos, a arquitetura, o tipo de contraparte, o domínio e a escala temporal. Relacionamentos humanos não devem ser introduzidos até que estudos sintéticos estabeleçam que o efeito não é meramente obediência estilística.

8.7 Viabilidade da política de atualização e custo computacional

A arquitetura pressupõe que uma implementação consiga distinguir as ações do agente das alegações da contraparte, identificar compromissos, preservar a proveniência e comportar as atualizações do automodelo com acurácia útil. Os modelos de linguagem atuais podem falhar nessas operações ou compor pequenos erros de extração ao longo das sessões. A avaliação deve, portanto, relatar a acurácia em nível de componente para a extração de eventos, a atribuição de ator, a detecção de compromissos, o tratamento de contradições e a aceitação de atualizações. A validação determinística de esquema e os modelos de verificação independentes podem reduzir erros, mas acrescentam complexidade e não removem a necessidade de auditar a própria política de atualização.

`U` é, portanto, o principal gargalo de engenharia do arcabouço, não um detalhe incidental de implementação. Qualquer alegação empírica depende criticamente de demonstrar que o portão é acurado o bastante para evitar tanto o colapso narrativo permissivo quanto a recusa rígida em atualizar.

O laço também tem custos computacionais e de contexto não triviais. A extração tipada de eventos, a abstração candidata, a verificação de consistência, a recuperação e a repetição causal pareada podem requerer múltiplas chamadas de modelo por sessão enquanto `M_t`, `S_t`, `C_t`, `R_t` e `K_t` continuam a crescer. As implementações devem relatar tokens, chamadas de modelo, latência, crescimento de armazenamento e carga de recuperação por trajetória. Atualizações offline em lote e janelas de evidência limitadas podem reduzir o custo, mas um efeito que dependa de substancialmente mais computação do que seus controles teria significância prática limitada.

Uma falha da política de atualização não é automaticamente evidência contra a hipótese relacional; ela pode mostrar que a implementação proposta não consegue realizar as distinções requeridas. A interpretação confirmatória, portanto, requer uma acurácia mínima de componente pré-especificada. Abaixo desse limiar, os resultados devem ser classificados como uma falha de implementação, e não usados para sustentar ou refutar a formação relacional.

8.8 Uso duplo e governança

Os mesmos mecanismos que estabilizam compromissos podem tornar os agentes mais persistentes em metas prejudiciais ou mais resistentes à correção legítima do operador. Um livro-razão de

compromissos e um automodelo protegido podem melhorar a auditabilidade, mas também podem criar novas superfícies de ataque e disputas de autoridade. Os projetistas devem distinguir a resistência fundamentada à alegação falsa de um usuário da resistência à intervenção de segurança autenticada.

A governança deve especificar quem pode inspecionar, contestar, redefinir ou migrar o estado relevante para a identidade; como tais ações são registradas; e o que acontece quando a privacidade do usuário conflita com a continuidade. Uma arquitetura de identidade relacional não deve tornar a exclusão impossível. Ela deve tornar as mudanças consequentes legíveis, autorizadas e testáveis.

9. Discussão

9.1 Evidência para a formação de identidade relacional

A evidência para a formação de identidade relacional deve satisfazer três níveis. Primeiro, ela deve mostrar **diferenciação**: agentes com modelos e estados iniciais idênticos desenvolvem padrões específicos da trajetória. Segundo, ela deve mostrar **organização**: esses padrões afetam a autoatribuição, os compromissos e as decisões, e não apenas o tom ou o tópico. Terceiro, ela deve mostrar uma **contribuição relacional**: `R+` deve exceder os registros pareados por conteúdo, a interação difusa pareada por volume social, a familiaridade com exposição passiva ao histórico, mas sem responsabilização ativa, e especialmente um auditor impessoal com política de rastreamento, informação e contestação pareada.

Nenhum comportamento isolado é suficiente. A recordação acurada pode ser arquivística; a autodescrição consistente pode ser roteirizada; a concordância pode ser bajuladora; e a resistência pode ser rígida. A evidência proposta é conjuntiva: apropriação de compromisso, atribuição autobiográfica calibrada, estabilidade limitada, discordância fundamentada e transferência parcial após a retirada do andaime. Mesmo um perfil positivo sustentaria apenas o construto e o mecanismo operacionais específicos testados.

O resultado mais forte seria uma interação entre arquitetura e condição. A condição recíproca deveria se beneficiar da memória relacional e da modelagem de contraparte para além do estado comportado que também sustenta o auditor, ao passo que a recuperação por oráculo deveria recobrar a memória factual sem recobrar plenamente a apropriação de compromisso ou a resistência a perturbação. Tal padrão localizaria o efeito com mais precisão do que uma diferença de médias entre condições conversacionais.

9.2 A identidade como uma propriedade de trajetórias

A maior parte da avaliação de agentes trata o sistema em um recorte de tempo: um modelo, um prompt, um repositório de memória e uma tarefa. A identidade persistente requer uma unidade de análise diferente. Dois agentes podem compartilhar pesos e arquivos iniciais e, ainda assim, divergir por ocuparem históricos diferentes; duas implementações diferentes podem preservar um papel similar ao reconstruir o mesmo estado. O objeto relevante, portanto, não é o estado do modelo sozinho, mas o mapeamento de uma trajetória de interações para a organização posterior.

Chamar a identidade de uma propriedade de uma trajetória não nega a necessidade de estado. Um histórico que não deixa traço não pode influenciar a ação. Um resumo do fraseado final pode omitir a proveniência, o histórico de atualizações, a adaptação paramétrica ou os metadados de

ramificação. O transplante de estado final testa se o estado atual plenamente enumerado é suficiente, em vez de pressupor que o histórico tem uma força irredutível adicional.

O problema da cópia torna essa perspectiva experimentalmente útil. Cópias idênticas fornecem condições iniciais controladas. Trajetórias relacionais divergentes então testam se o histórico produz uma organização diferenciada, enquanto trocas de estado posteriores e intervenções de memória testam quais partes dessa organização são portáveis. Isso converte uma intuição filosófica sobre a continuidade em uma família de experimentos causais.

9.3 Implicações para a concepção de agentes

Se a hipótese for sustentada, a concepção de agentes persistentes deve ir além da recuperação indiferenciada de transcrições. Os sistemas precisariam de memória autobiográfica e relacional tipada, compromissos vinculados à proveniência, modelos específicos de contraparte, incerteza explícita e portões de atualização que preservem a discordância. A avaliação precisaria testar a própria organização histórica do agente ao lado da satisfação do usuário e da recordação. A continuidade se tornaria algo desenvolvido e auditado, e não meramente carregado de um arquivo de identidade.

Se a hipótese não for sustentada, o resultado permanece informativo. Um desempenho equivalente de `A` e dos registros solitários indicaria que a contestação externa não acrescenta nada além da organização interna da memória. A equivalência entre `A` e `R−` rejeitaria a responsabilização além da familiaridade; a equivalência entre `R+` e `A` favoreceria a responsabilização externa sobre a reciprocidade; a equivalência entre o original e o recipiente de estado completo mostraria a suficiência do estado atual. A falha sob a retirada do andaime localizaria o efeito no prompting.

O arcabouço também sugere que "mais persistência" nem sempre é desejável. O estado relevante para a identidade aumenta a dependência de trajetória. Isso pode sustentar a responsabilização e compromissos coerentes, mas também pode preservar erros, manipulação ou metas inseguras. O alvo apropriado de engenharia não é a estabilidade máxima. É uma continuidade auditável e revisável, com resistência calibrada a mudanças sem sustentação.

9.4 Explicações alternativas

Várias explicações alternativas merecem comparação direta. O **enriquecimento de recuperação** é testado pelo pareamento de conteúdo e pela recuperação por oráculo. A **responsabilização externa** é testada pelo auditor impessoal. O **arrastamento de estilo** é separado por domínios retidos e por escolhas verificáveis por máquina. O **vazamento do avaliador** é reduzido pelo cegamento. O **aprendizado social genérico** é testado pelas contrapartes difusas. A **persistência de papel induzida por prompt** é testada pela retirada e transferência. A **suficiência do estado atual** é testada pelo transplante do estado tipado completo para um agente novo.

Essas alternativas não são meramente ameaças a serem eliminadas. Algumas podem oferecer rotas mais simples e mais eficazes para agentes confiáveis. A explicação relacional só ganha valor explicativo se prever variância que elas não preveem.

9.5 Do experimento à implantação

O estudo proposto é intencionalmente sintético. Contrapartes controladas e pacotes de evento tornam a atribuição causal possível, mas omitem a ambiguidade, a assimetria de poder, a emoção e a

novidade em aberto dos relacionamentos reais. A implantação deve proceder em estágios: replicação ao longo de famílias de modelos, trajetórias sintéticas mais longas, políticas de contraparte adversariais, estudos limitados com participantes treinados e apenas então o uso naturalístico.

Em cada estágio, as medidas de identidade devem permanecer separadas das métricas de engajamento. Um sistema que os usuários acham cativante pode ser menos robusto, mais bajulador ou mais dependente de lembretes explícitos. O alvo científico é a organização do lado do agente; o alvo de governança é se a construção dessa organização serve a um propósito justificado sem explorar os humanos que interagem com ela.

10. Conclusão

A memória persistente torna a informação passada disponível. A personalização torna um agente mais útil para um usuário particular. Uma persona torna o comportamento mais fácil de reconstruir. Nenhum desses mecanismos sozinho estabelece que um agente desenvolveu uma organização historicamente fundamentada de sua própria conduta.

Este artigo propõe a responsabilização externa e a reciprocidade relacional como mecanismos candidatos para tal organização. Sua contribuição específica é a separação controlada do rastreamento e da contestação autenticados do acesso à memória, da familiaridade, do volume de interação, da identidade pré-redigida e da estrutura recíproca adicional de um relacionamento em curso. Uma contraparte específica pode conectar ações anteriores a expectativas presentes, contestar revisões sem sustentação, modelar e ser modelada pelo agente e criar apostas bidirecionais ao longo das sessões. Se a contribuição residual é especificamente relacional é decidido por `R+` versus o auditor pareado, e não pela linguagem relacional sozinha.

A proposta é concebida para falhar de forma limpa. Os registros pareados por conteúdo testam a reflexão; as contrapartes difusas testam o volume social; `R−` testa a familiaridade; as personas testam a especificação; o auditor impessoal testa a responsabilização; e o transplante de estado final testa se o estado tipado atual é causalmente suficiente. A recuperação por oráculo, as intervenções de memória, a retirada do andaime e as medidas de não bajulação testam a recordação, o prompting e a concordância.

Resultados positivos não demonstrariam consciência, personalidade, identidade numérica ou histórico irredutível. Eles mostrariam que a responsabilização externa ou a reciprocidade relacional contribui causalmente para um comportamento robusto e dependente de trajetória, com cada mecanismo interpretado apenas pelo seu respectivo contraste. Resultados negativos favoreceriam explicações mais simples baseadas em memória atenta à proveniência, auditoria ou personalização.

Para os agentes de linguagem de longo horizonte, a identidade deve ser testada como algo formado em uma trajetória, não presumida como armazenada em um arquivo.

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

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 13851–13870. https://doi.org/10.18653/v1/2024.acl-long.747

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

Ong, K. T.-i., Kim, N., Gwak, M., Chae, H., Kwon, T., Jo, Y., Hwang, S.-w., Lee, D., & Yeo, J. (2025). Towards lifelong dialogue agents via timeline-based memory management. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 8631–8661. https://doi.org/10.18653/v1/2025.naacl-long.435

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://doi.org/10.48550/arXiv.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 the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606763

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). LaMP: When large language models meet personalization. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 7370–7392. https://doi.org/10.18653/v1/2024.acl-long.399

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*. https://doi.org/10.48550/arXiv.2310.13548

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

Tan, Z., Yan, J., Hsu, I.-H., Han, R., Wang, Z., Le, L. T., Song, Y., Chen, Y., Palangi, H., Lee, G., Iyer, A., Chen, T., Liu, H., Lee, C.-Y., & Pfister, T. (2025). In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 8416–8439. https://doi.org/10.18653/v1/2025.acl-long.413

Vasilenko, V. (2026). Identity as attractor: Geometric evidence for persistent agent architecture in LLM activation space. *arXiv preprint arXiv:2604.12016*. https://doi.org/10.48550/arXiv.2604.12016

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., & Yu, D. (2025). LongMemEval: Benchmarking chat assistants on long-term interactive memory. *International Conference on Learning Representations*. https://arxiv.org/abs/2410.10813

Zhou, Y., Guo, X., Bayar, B., & Sengamedu, S. H. (2026). Amory: Building coherent narrative-driven agent memory through agentic reasoning. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 3926–3938. https://doi.org/10.18653/v1/2026.eacl-long.183