LLMs DE FRONTEIRA — TURNO ÚNICO × MULTI-TURNO

Análise das Implicações dos Turnos Únicos e Multi-turnos em Modelos de IA Generativas de Fronteira (LLMs)

Estado, custo, desempenho, segurança e governança da interação conversacional com LLMs



Aridio Silva — Pesquisador Independente, Brasil
ORCID 0009-0008-2411-6995 · CC BY 4.0

© Aridio Silva @aridiosilva — Novembro 2026

RESUMO EXECUTIVO

Cinco achados sobre turno único × multi-turno

~T^2

A API não guarda estado: o histórico é reenviado a cada turno e a entrada acumulada cresce ~T^2

Cache

O cache derruba o custo da entrada; o raciocínio (saída) passa a dominar a conta

−39%

Laban et al.: 15 LLMs perdem 39% em multi-turno; não-confiabilidade +112%, aptidão −16%

Risco

O multi-turno amplia a superfície: Crescendo, injeção indireta e estado persistente

≠ mais

Mais raciocínio não resolve a perda de confiabilidade; consolidar e governar ajuda mais

Importância

Decide arquitetura, custo e risco de qualquer produto com LLM

Implicações positivas (+)

Há técnicas maduras para cada eixo: cache, esforço, estado, política

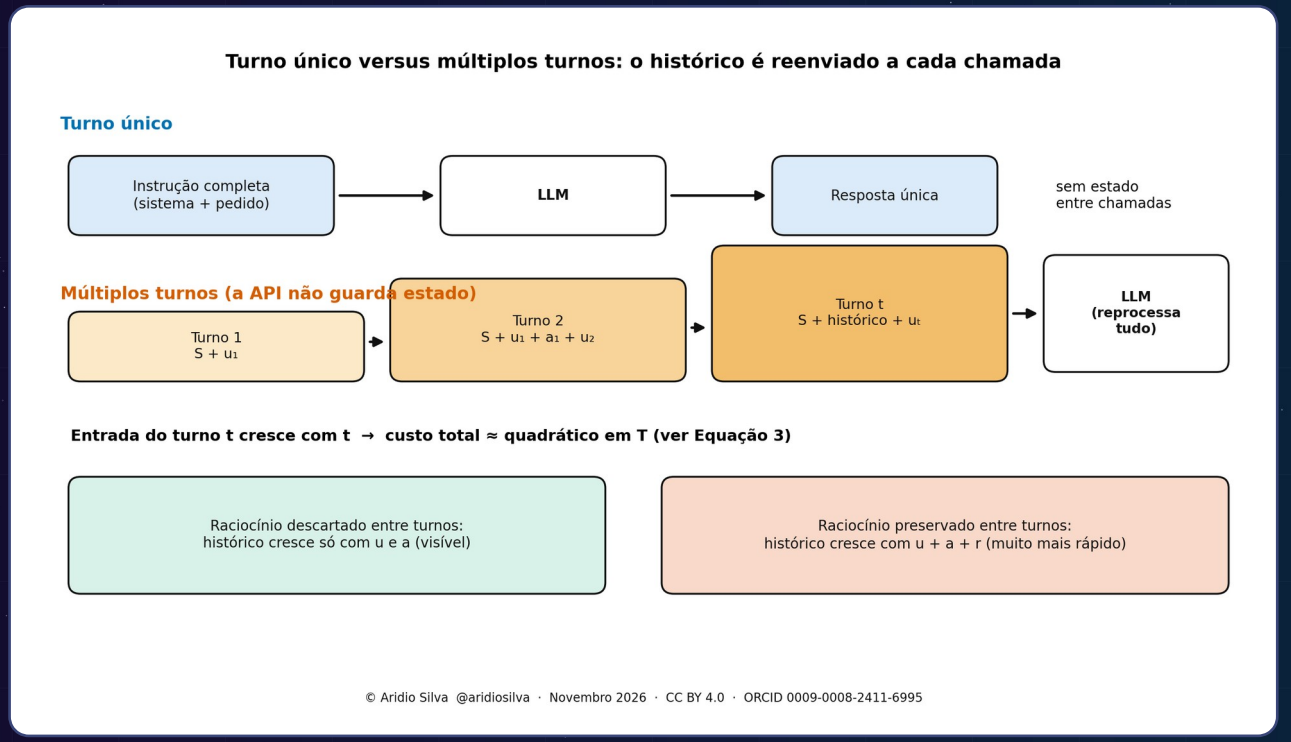
Implicações negativas (−)

Faltam dados públicos comparáveis por modelo e por nível de esforço

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Turno único versus múltiplos turnos

- **Turno único:** uma instrução completa, uma resposta; sem estado entre chamadas
- **Multi-turno:** cada resposta é condicionada a todo o histórico, reenviado a cada chamada
- **Quem guarda o estado?** o cliente, não o modelo (API stateless)
- **Não é contexto longo:** o que define o multi-turno é a revelação incremental da informação



Importância

Base de toda a análise: define custo, risco e confiabilidade

Implicações positivas (+)

Turno único: simples, barato, testável, superfície de ataque menor

Implicações negativas (−)

Multi-turno: custo ~T^2, latência e risco acumulativos

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

O que é quadrático, e o que não é

- **Dentro da chamada:** a atenção custa $O(n^2 \cdot d)$ por camada (Vaswani et al.)
- **Ao longo da conversa:** a soma dos tokens reenviados cresce $O(T^2)$
- **Correção do compêndio:** quadrático não é exponencial

236 mil

tokens de entrada em 20 turnos (raciocínio descartado)

1,756 mi

mantendo 8.000 tokens de raciocínio por turno (7,4×)

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \qquad \text{custo por camada} = O(n^2\, d)$$

© Aridio Silva @aridiosilva · Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

$$I_t = S + tu + (t-1)(a + \kappa r)$$

© Aridio Silva @aridiosilva · Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

$$\sum_{t=1}^{T} I_t = TS + u\frac{T(T+1)}{2} + (a + \kappa r)\frac{T(T-1)}{2} = O(T^2)$$

© Aridio Silva @aridiosilva · Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Importância

Explica por que conversas longas ficam caras mesmo com cache

Implicações positivas (+)

Dá uma fórmula simples para prever e orçar o custo

Implicações negativas (−)

Ignora variações de tokenizador e de cobrança de escrita de cache

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

COMPARAÇÃO

Matriz: turno único × multi-turno

Critério	Turno único	Multi-turno
Estado	Stateless	Stateful: exige gestão de histórico
Custo	Baixo e linear por requisição	Crescente, ~T^2 sem cache
Latência	Mínima	Escala com o contexto
Risco	Isolado e menor	Acumulativo (contexto e estado)
Casos ideais	Classificação, extração, tradução	Assistência interativa, código, agentes
Informação	Completa no início	Incremental; permite subespecificação

Importância

Ajuda a escolher o modo certo para cada tarefa

Implicações positivas (+)

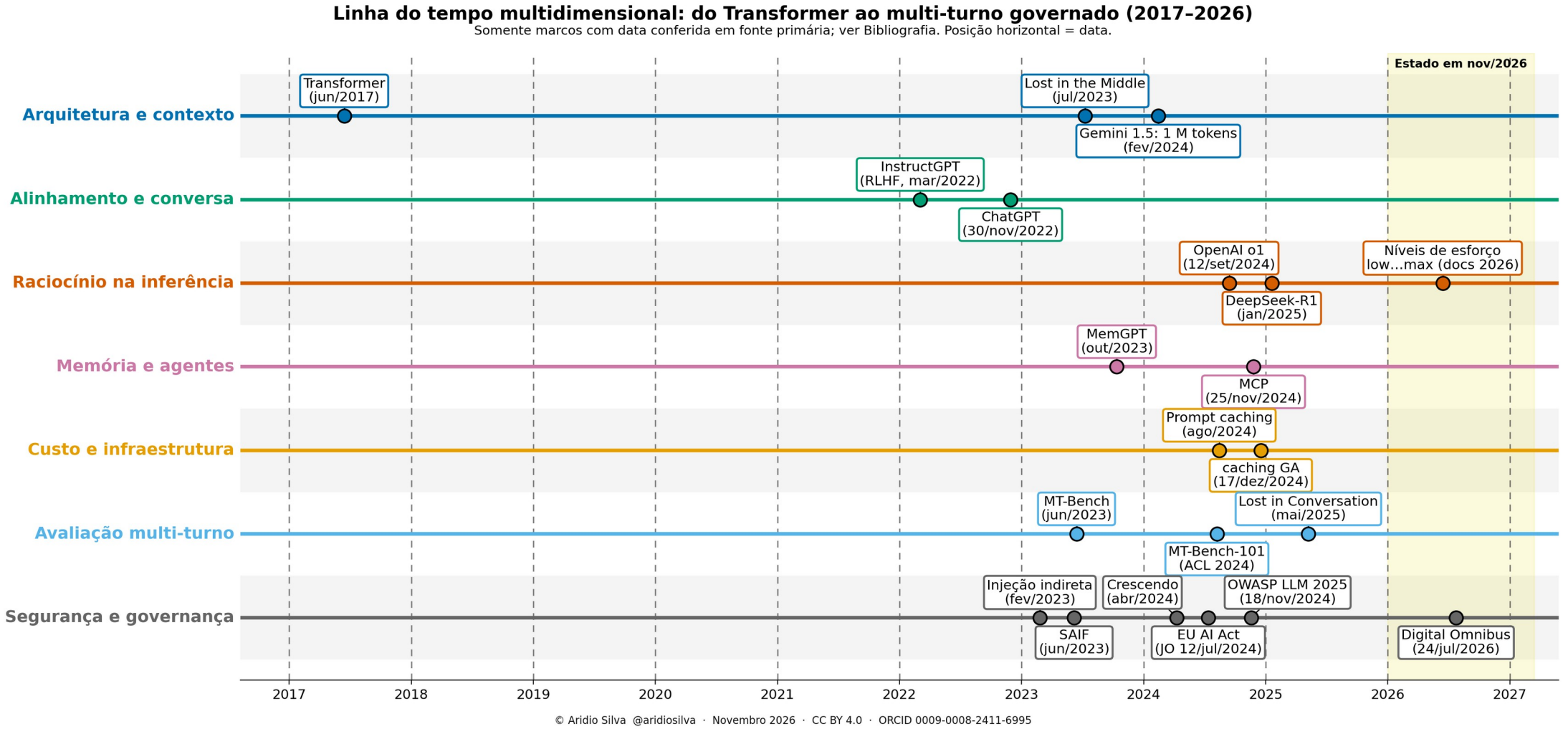
Turno único para tarefas bem especificadas: previsível e barato

Implicações negativas (−)

Multi-turno sem disciplina: custo, deriva e risco crescentes

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Linha do tempo multidimensional, 2017–2026



© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Cinco eras: do prompt isolado à conversa governada

2017–2021 Arquitetura

Transformer; janelas de 512–2.048 tokens; uso por prompting (essencialmente turno único)

2022 Diálogo

InstructGPT (RLHF) e ChatGPT: o multi-turno vira formato de produto

2023 Contexto e ataques

Lost in the Middle, MemGPT, MT-Bench, injeção indireta, SAIF

2024 Cache e raciocínio

Gemini 1.5 (1 M tokens), prompt caching, o1, MCP, Crescendo, OWASP 2025

2025–26 Confiabilidade

R1, Lost in Conversation, níveis de esforço, Digital Omnibus

Importância

Mostra que avaliação, preço e segurança chegaram depois da adoção

Implicações positivas (+)

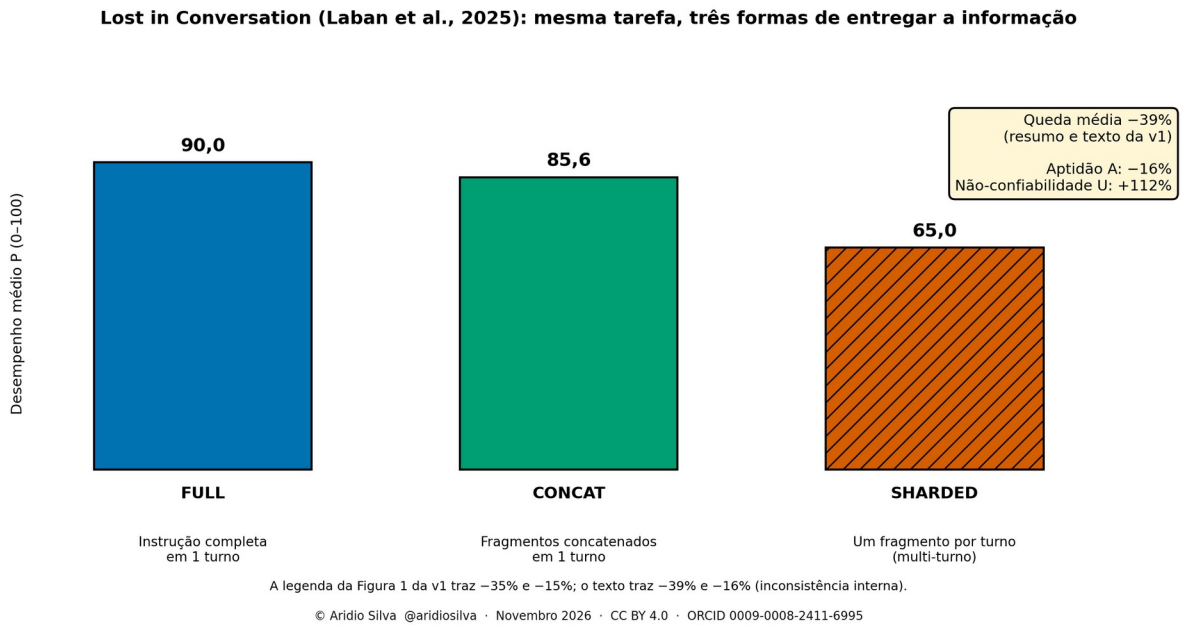
Inovação rápida: cache, paginação de memória e protocolos em poucos anos

Implicações negativas (−)

Marcos de 2026 sem data em fonte primária ficaram de fora

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Lost in Conversation: a mesma tarefa, três entregas



- **15 LLMs, 6 tarefas,** mais de 200 mil conversas simuladas
- **Queda média de 39%;** aptidão −16%, não-confiabilidade +112%
- **Já com 2 turnos;** CONCAT = 95,1% de FULL: a perda não vem da reformulação
- **Raciocínio não salva:** o3 e R1 pioram igual, com respostas 33% mais longas

$$P = \frac{1}{N}\sum_{i=1}^{N} S_i, \quad A^{90} = \text{percentil}_{90}(S), \quad U_{10}^{90} = \text{percentil}_{90}(S) - \text{percentil}_{10}(S)$$

© Aridio Silva @aridiosilva · Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Importância

Prova que o desempenho de turno único superestima o uso real

Implicações positivas (+)

Métricas claras (P, A, U) e desenho controlado com as mesmas tarefas

Implicações negativas (−)

Modelos de 2025, tarefas analíticas em inglês; v1 com −35% e −39%

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Por que os modelos se perdem, e o que ajuda

Resposta prematura

Suposições cedo contaminam o resto

Inchaço

Respostas finais 20–300% maiores

Meio da conversa

Os turnos do meio são esquecidos

Verbosidade

Mais texto, mais suposições

- **Consolidar:** reunir os requisitos num único prompt
- **Recapitular:** repetir requisitos (SNOWBALL) recupera 15–20% da queda
- **Reiniciar:** conversa nova com tudo consolidado
- **Temperatura 0 não resolve:** resta ~30 pontos de não-confiabilidade

$$\Pr(\text{sem erro em } T \text{ turnos}) = (1-\varepsilon)^T \approx e^{-\varepsilon T}$$

© Aridio Silva @aridiosilva · Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Importância

Dá regras práticas imediatas para quem usa ou constrói com LLMs

Implicações positivas (+)

Consolidar e recapitular melhora confiabilidade sem trocar de modelo

Implicações negativas (−)

Eq. 6 é modelo didático (ε = 5%, T = 20 → ~36%), não resultado empírico

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Níveis de raciocínio por provedor (nov./2026)

Provedor	Modelo	Níveis	Padrão
OpenAI	GPT-6 Astra	low → max	—
OpenAI	GPT-6.1 Sol	low → max	medium
OpenAI	GPT-6.1 Terra, GPT-6 Luna	none → max	medium (Luna)
Anthropic	Fable 5.1, Sonnet 5.5	low → max*	high
Anthropic	Opus 5.5, Haiku 5.5	low → max*	medium
Google	Gemini 4 Argon, 3.8 Flash	low, medium, high	medium
Google	3.5 Flash-Lite	minimal → high	minimal
xAI	Grok 4.7	low → xhigh	high
DeepSeek	V4.1-Flash, V4-Pro	com ou sem raciocínio	com raciocínio

Importância

O esforço é a principal alavanca de custo em conversas longas

Implicações positivas (+)

Escalas parecidas permitem comparar provedores em níveis equivalentes

Implicações negativas (−)

Nomes e disponibilidade mudam; *varia por modelo; confirme nas páginas oficiais

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Preços e o peso da leitura de cache (US$/M tokens)

Modelo	Entrada	Saída	Cache	Cache/ent.
GPT-6 Astra	10,00	50,00	1,00	10%
Claude Fable 5.1	10,00	50,00	0,25	2,5%
Claude Opus 5.5	4,00	20,00	0,20	5%
Sol / Sonnet 5.5 / Gemini 4 Argon	2,00	10,00	0,10	5%
Grok 4.7 (<200 mil)	2,00	6,00	0,50	25%
DeepSeek V4-Pro (pico)	1,32	3,96	0,044	3,3%
Gemini 3.8 Flash	0,75	3,75	0,075	10%
Luna / Haiku 5.5	0,10	0,50	0,01	10%

- **Tokenizador:** modelos Claude ≥4.7 geram ~30% mais tokens: preço por token não é 1:1
- **Faixas:** Haiku 5.5 acima de 100 mil; Grok dobra acima de 200 mil
- **Datas:** Gemini 4 Argon dobra em 1°/dez./2026
- **DeepSeek:** metade do preço fora do horário de pico

Importância

No multi-turno a leitura de cache pesa mais que o preço de entrada

Implicações positivas (+)

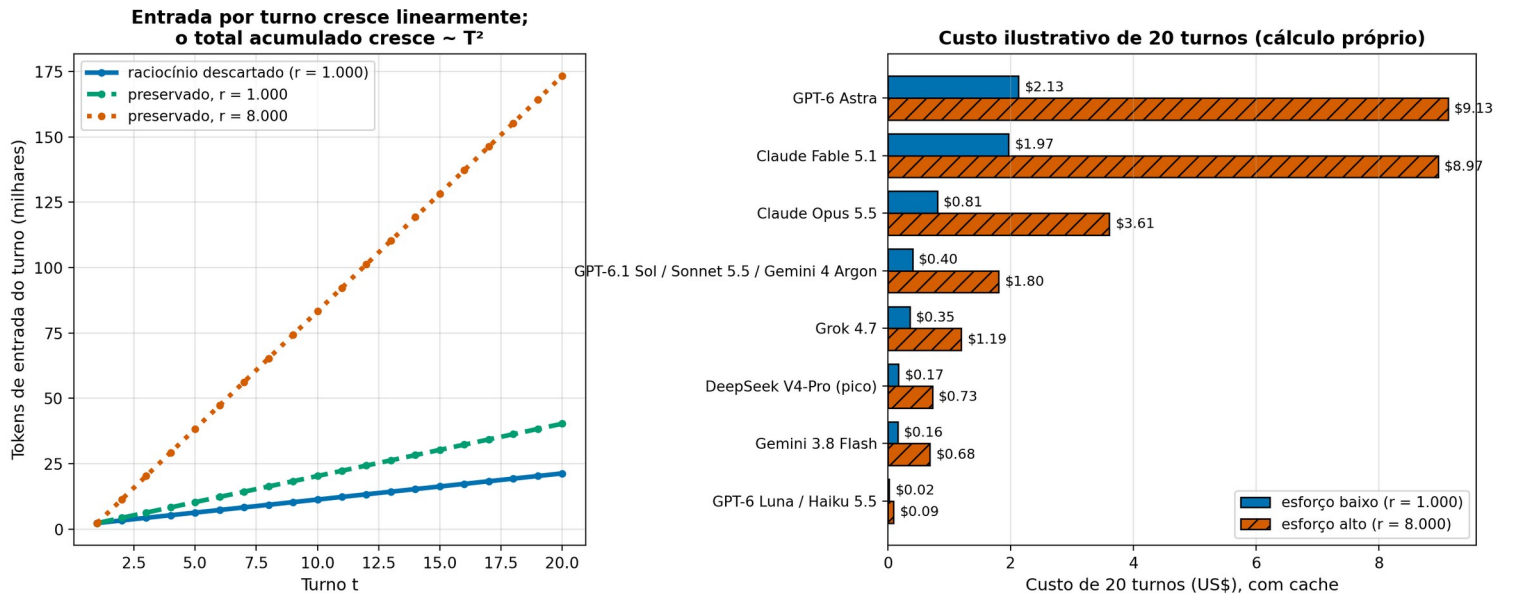
Cache a 2–5% da entrada (Anthropic, DeepSeek) barateia conversas longas

Implicações negativas (–)

Grok cobra 25% da entrada no cache: vantagem menor em conversas longas

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Exemplo de 20 turnos: o esforço pesa mais que os turnos



Premissas: S = 2.000, u = 300, a = 700 tokens; cache acerta sempre; escrita de cache sem sobretaxa; preços: documentação oficial de cada provedor (snapshot nov/2026).

© Aridio Silva @aridiosilva · Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

3,4–4,5×
custo de esforço alto sobre baixo

−86%
custo de entrada com cache (leitura a 5%)

US$ 0,40
20 turnos, esforço baixo, Sol/Sonnet/Argon (cálculo próprio)

Importância

Mostra onde está o dinheiro: saída e raciocínio, não o histórico

Implicações positivas (+)

Com cache, a conversa longa fica previsível e barata

Implicações negativas (−)

Preservar raciocínio entre turnos multiplica a entrada em ~7×; cálculo próprio, não dado dos provedores

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Multi-turno amplia a superfície de ataque

Crescendo: escalonamento gradual

Só entradas benignas; o modelo segue o próprio texto recente. Benchmarks eram de turno único.

Injeção indireta de prompt

Dados recuperados agem como instruções; a injeção persiste no histórico e sequestra turnos futuros.

Cascata de alucinações

Um erro aceito no turno 2 vira estado oficial e contamina o resto.

Consumo descontrolado

Conversas longas com raciocínio alto viram ataque de custo (OWASP LLM10:2025).

Importância

A autorização não pode depender do julgamento do modelo ao longo da conversa

Implicações positivas (+)

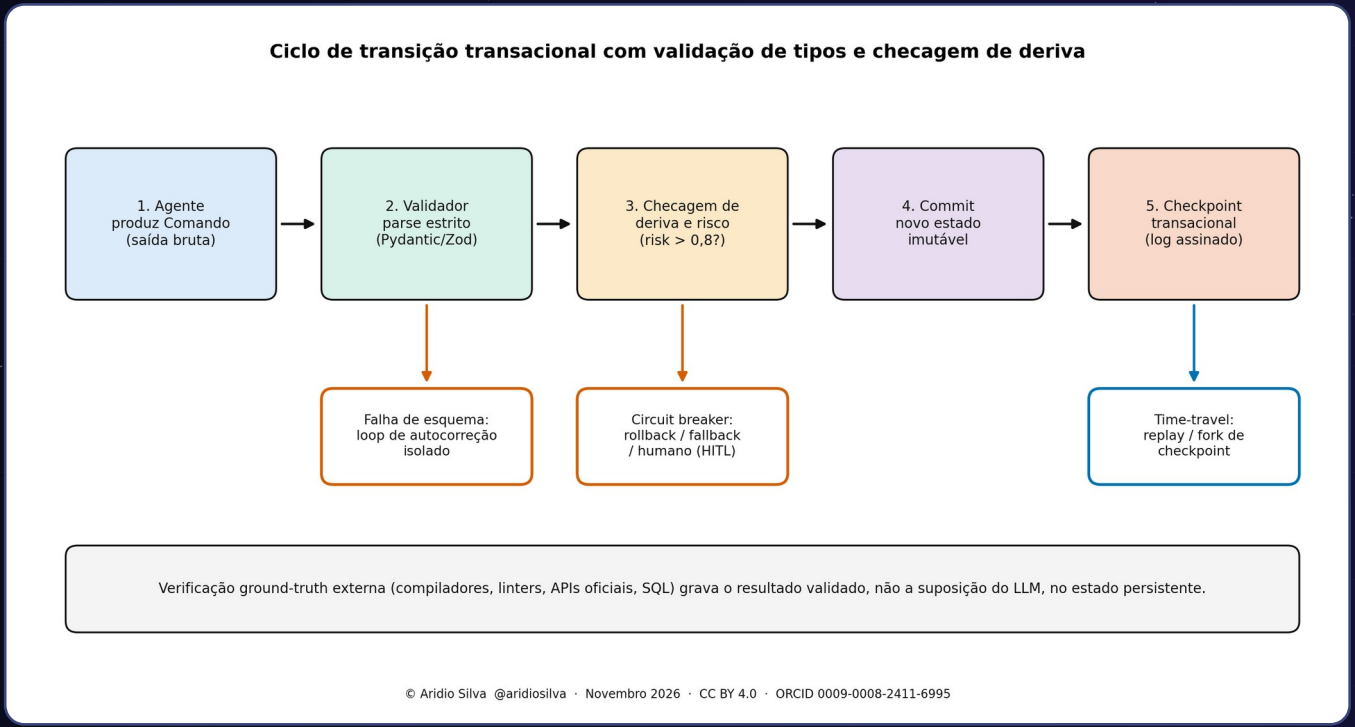
Há taxonomias maduras (OWASP, SAIF) e evidência experimental

Implicações negativas (−)

Defesas por representação interna falham contra ataques graduais

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Estado tipado e mitigação de deriva



- **Deriva semântica:** reancorar instruções; resumir por grafo de metas
- **Deriva comportamental:** circuit breakers e agentes críticos
- **Deriva epistêmica:** gravar o resultado verificado, não a suposição
- **Contratos de tipo:** Pydantic/Zod antes de tocar o estado global

Importância

O estado do agente deve ser artefato de engenharia, não texto de chat

Implicações positivas (+)

Saídas malformadas são barradas; checkpoints permitem auditoria e recuperação

Implicações negativas (−)

Validador de esquema não pega erro semântico; limiar de risco (0,8) é heurístico

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Time-travel debugging: o que o replay realmente faz

Replay e fork

Replay retoma de um checkpoint; fork ramifica com estado modificado. Checkpoints a cada super-step (LangGraph).

Auditoria: replay reexecuta

Chamadas a LLM e a APIs disparam de novo e podem dar outro resultado. Determinismo só se as saídas forem gravadas. A doc não cita seeds.

Notas sobre o código Pydantic

class Config e .dict() são obsoletos no v2 (model_config, model_dump); faltam imports; frozen=True é raso (conhecimento prévio, não verificado).

Importância

Depurar agentes exige gravar eventos, não só estados

Implicações positivas (+)

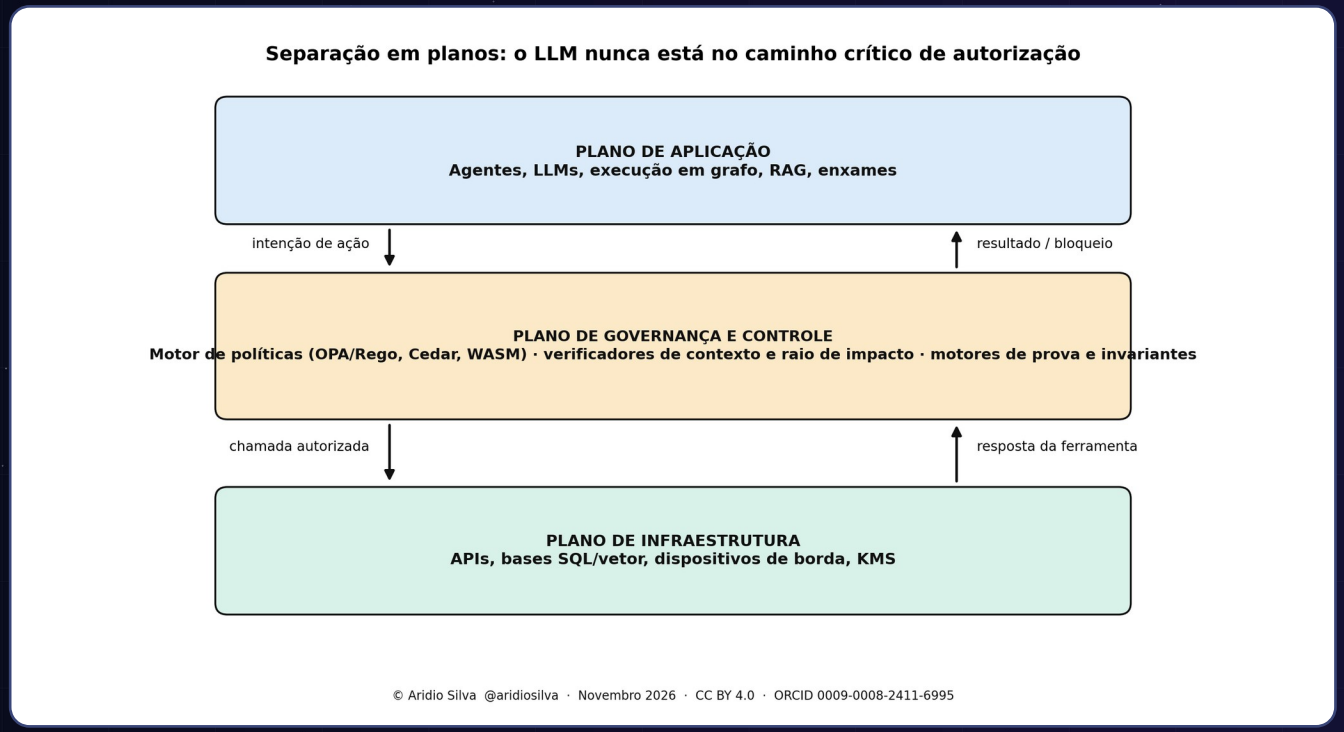
Reverter e bifurcar sem reprocessar tudo; trilha forense

Implicações negativas (−)

Rollback exige efeitos externos reversíveis; replay sem gravação engana

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

O LLM nunca está no caminho crítico de autorização



- **Identidade:** X.509/SPIFFE por agente; delegação por subconjunto estrito
- **Autoridade delimitada:** acesso granular e temporário (OPA/Rego, Cedar)
- **Intercepção:** gate pré-ação, circuit breakers e auditores pós-ação
- **Trilha imutável:** logs assinados e append-only

Importância

A segurança passa a ser propriedade do sistema, não do modelo

Implicações positivas (+)

Política determinística é imune a engenharia de prompt

Implicações negativas (−)

Custo, complexidade de implantação e necessidade de PKI e policy-as-code

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

GOVERNANÇA

Matriz de risco e o calendário regulatório

Nível	Comportamento	Exemplo
Baixo	Auto-aprovada, log assinado	Ler documentação
Médio	Notifica e prossegue	Relatório intermediário
Alto	Pausa; aprovação humana (HITL)	Transação financeira
Inadmissível	Bloqueio imediato (fail-closed)	Escalar privilégios

EU AI Act, Arts. 12 e 14

Só sistemas de alto risco. Logs contribuem para a conformidade, não a garantem.

Novo prazo: 2 dez. 2027

Reg. (UE) 2026/1744 (Digital Omnibus) adiou o Anexo III. Fontes secundárias; EUR-Lex não aberto.

Importância

Liga controles técnicos a obrigações regulatórias; NIST AI 100-1 é a referência correta

Implicações positivas (+)

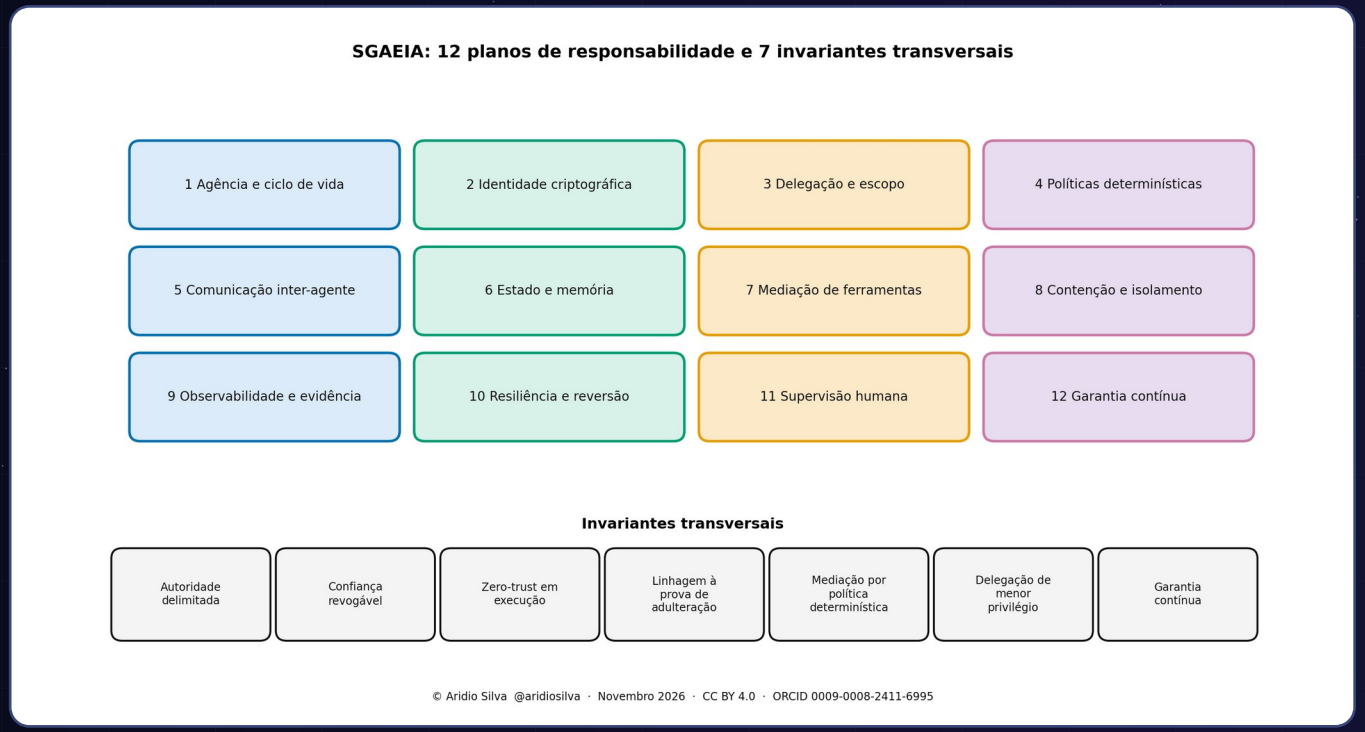
Classificar ações por risco dá previsibilidade e supervisão humana onde importa

Implicações negativas (−)

Prazos mudam; HITL cria gargalo humano; trilhas imutáveis elevam o volume de dados

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

12 planos de responsabilidade e 7 invariantes



- **Premissa:** o objeto de garantia é o sistema completo em operação, não o modelo isolado
- **Responde ao multi-turno:** planos 6 (estado), 10 (reversão) e 12 (garantia contínua)
- **Responde à injeção:** planos 3, 4 e 7 (delegação, política, mediação)
- **Obra do autor:** sem avaliação independente publicada

Importância

Arquitetura de referência aberta e neutra para agentes autônomos e IA de borda

Implicações positivas (+)

Integra identidade, política, estado imutável e Evidence-as-Code

Implicações negativas (−)

Exige maturidade de operação; eficácia empírica ainda não demonstrada por terceiros

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

RIGOR DE FONTES

Auditoria: 8 correções no compêndio

Trecho do compêndio	Problema	Correção
Lost in the Middle: autor 'Manning', 2023	Falta Petroni; TACL 12:157–173 (2024)	Autoria e venue
NIST AI RMF 'SP 1270'	O documento é NIST AI 100-1	Número corrigido
Custo 'cresce exponencialmente'	$O(n^2)$ é quadrático	Termo corrigido
Logs 'garantem' conformidade	Só alto risco; prazos mudaram	Reformulado
Time-travel 'determinístico'	Replay reexecuta LLM/API	Nuançado
'LangChain Architecture Team'	Autoria não verificável	Doc oficial
'Cedar Specification (AWS)'	Título não localizado	Cutler et al. (2024)
Código Pydantic	class Config e .dict() obsoletos	Notas + variante v2

Importância

Em trabalho técnico, toda referência gerada por LLM precisa de checagem em fonte primária

Implicações positivas (+)

Protocolo simples e reproduzível; achou erros reais em material já escrito

Implicações negativas (−)

Consome tempo; DOI do Zenodo e textos legais não puderam ser abertos

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Estado da arte em nov./2026

Confiabilidade multi-turno como meta

Aptidão igual em 1 e N turnos; U < 15 a T = 1,0

Economia por cache e esforço

Cache barateia a entrada; esforço controla a saída

Estado como engenharia

Tipado, imutável, versionado; checkpoints e fork

Governança em tempo de execução

Política determinística, identidade, trilha imutável

Protocolos de integração

MCP padroniza ferramentas e amplia a injeção indireta

Importância

Resume onde a prática está hoje em cinco eixos

Implicações positivas (+)

Cada eixo tem técnicas documentadas e maduras

Implicações negativas (−)

Faltam avaliações públicas por modelo e nível de esforço

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Lacunas e perspectivas futuras

Horizonte	Perspectiva	Ancoragem
Curto prazo	Benchmarks de confiabilidade por nível de esforço; custo normalizado por tokenizador	Laban et al.; docs de preço
Curto prazo	Produtos que consolidam requisitos e resumem estado por padrão	Laban; MemGPT
Médio prazo	Política determinística e identidade de agente como padrão	Cedar; NIST; OWASP
Médio prazo	Evidência de conformidade gerada automaticamente	AI Act; Omnibus
Longo prazo	Modelos que reduzem suposições precoces e mantêm consistência	Causas de Laban

Importância

Aponta a agenda de pesquisa e de produto para os próximos anos

Implicações positivas (+)

Vantagem tende a migrar de capacidade bruta para confiabilidade por custo (inferência do autor)

Implicações negativas (−)

Inferências são do autor, não fatos citados; resultados fortes são de 2025

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

CONCLUSÃO

Turno único ou multi-turno é decisão de arquitetura

- **Turno único** para tarefas bem especificadas: barato, previsível, testável
- **Multi-turno** quando houver exploração, com requisitos consolidados e recapitulados
- **Custo:** use cache e controle o nível de esforço; o raciocínio domina a conta
- **Confiabilidade:** não confie em mais raciocínio; reinicie e consolide
- **Segurança:** estado tipado, autorização determinística fora do LLM, trilha auditável
- **Rigor:** verifique cada referência em fonte primária



Leia o documento completo (PDF/MD) em PT e EN.

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Glossário de termos técnicos (1/3)

- **Aptidão (A^{90}):** Desempenho de melhor caso (percentil 90) em simulações repetidas
- **Atenção:** Cada posição pondera todas as outras; custo $O(n^2 \cdot d)$ por camada
- **Cache de prompt:** Reuso do prefixo já processado, cobrado a preço menor de leitura
- **Cadeia de pensamento:** Passos de raciocínio gerados antes da resposta final
- **Checkpoint:** Instantâneo persistente do estado; base de replay e fork
- **Circuit breaker:** Interrompe o fluxo quando um limite de risco é excedido
- **CONCAT / FULL / SHARDED:** Condições de Laban: concatenado, completo, um fragmento por turno
- **Janela de contexto:** Máximo de tokens processados de uma vez
- **Cascata de alucinações:** Premissa errada aceita num turno que contamina os seguintes
- **Deriva (drift):** Desvio semântico, comportamental ou epistêmico
- **Esforço de raciocínio:** Quanto o modelo 'pensa' antes de responder (low a max)
- **Event sourcing:** Persistir cada transição como evento imutável
- **Evidence-as-Code:** Evidência de conformidade gerada e verificável como código
- **Fork:** Ramificação a partir de um checkpoint com estado modificado

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Glossário de termos técnicos (2/3)

- **Guardrail:** Camada programável que filtra entradas e saídas
- **HITL:** Aprovação humana explícita antes de uma ação crítica
- **Injeção indireta:** Instruções maliciosas em dados recuperados tratadas como comandos
- **Jailbreak:** Induzir o modelo a violar suas políticas
- **KV cache:** Chaves e valores de atenção reaproveitados na geração
- **Lost in the Middle:** Queda de desempenho quando a informação está no meio
- **MCP:** Padrão aberto para conectar assistentes a ferramentas e dados
- **Multi-turno:** Sequência de mensagens, cada resposta condicionada ao histórico
- **Não-confiabilidade (U):** Diferença entre os percentis 90 e 10 do desempenho
- **OPA / Rego:** Motor e linguagem de política como código
- **Cedar:** Linguagem de autorização de código aberto com análise formal
- **Prefill:** Fase em que o modelo processa o prompt antes de gerar
- **RAG:** Geração aumentada por recuperação de documentos
- **RLHF:** Aprendizado por reforço com feedback humano

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Glossário de termos técnicos (3/3)

- **Sliding window:** Mantém só as N interações mais recentes
- **SNOWBALL / RECAP:** Repetir requisitos a cada turno ou ao final
- **Stateless / stateful:** Sem estado entre chamadas / com estado mantido
- **Test-time compute:** Computação gasta na inferência, como tokens de raciocínio
- **Token:** Unidade de texto processada e cobrada pelo modelo
- **Tokens de raciocínio:** Tokens de pensamento; cobrados como saída
- **Time-travel debugging:** Replay e fork de execuções a partir de checkpoints
- **Turno:** Mensagem de um papel (usuário ou assistente)
- **TTFT:** Tempo até o primeiro token da resposta
- **Unbounded Consumption:** Consumo descontrolado de inferência e custo (OWASP LLM10)
- **Zero Trust:** Não confiar implicitamente em nenhuma entidade ou transição
- **Blast radius:** Extensão do dano se uma ação autorizada falhar
- **FSM / DAG:** Máquina de estados finitos / grafo acíclico dirigido
- **Agent identity:** Identidade criptográfica única por agente (SPIFFE/X.509)

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Bibliografia auditada (1/3)

✔ verificada em fonte primária ◐ verificada em parte ou por fonte secundária ○ não verificada nesta sessão

- [1] Vaswani et al. (2017). Attention is all you need. NeurIPS 30, 5998–6008. arXiv:1706.03762. ✔
- [2] Ouyang et al. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155. ✔
- [3] Greshake et al. (2023). Not what you've signed up for… indirect prompt injection. arXiv:2302.12173. ✔
- [4] Zheng et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. NeurIPS 36, 46595–46623. ✔
- [5] Liu et al. (2024). Lost in the middle. TACL 12:157–173. DOI 10.1162/tacl_a_00638. ✔
- [6] Packer et al. (2023). MemGPT: towards LLMs as operating systems. arXiv:2310.08560. ✔
- [7] Bai et al. (2024). MT-Bench-101. ACL 2024, 7421–7454. DOI 10.18653/v1/2024.acl-long.401. ✔
- [8] Russinovich, Salem, Eldan (2025). Crescendo multi-turn jailbreak. USENIX Security '25, 2421–2440. arXiv:2404.01833. ✔
- [9] Bullwinkel et al. (2025). Representation engineering perspective on multi-turn jailbreaks. ICML 2025 DIG-BUGS (poster). ✔
- [10] Laban, Hayashi, Zhou, Neville (2025). LLMs get lost in multi-turn conversation. arXiv:2505.06120 (v1). ✔
- [11] Guo et al. (2025). DeepSeek-R1. arXiv:2501.12948. ◐
- [12] Cutler et al. (2024). Cedar. PACMPL 8(OOPSLA1). DOI 10.1145/3649835. ✔

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Bibliografia auditada (2/3)

✔ verificada em fonte primária ◐ verificada em parte ou por fonte secundária ○ não verificada nesta sessão

- [13] NIST (2023). AI RMF 1.0. NIST AI 100-1. DOI 10.6028/NIST.AI.100-1. ◐
- [14] EU (2024). Regulation 2024/1689 (AI Act). OJ L, 12.7.2024. ◐
- [15] EU (2026). Regulation 2026/1744 (Digital Omnibus on AI). OJ L, 24.7.2026. ◐
- [16] OWASP GenAI Security Project (2024). Top 10 for LLM Applications 2025 (v2.0). ✔
- [17] Google (2023). Introducing Google's Secure AI Framework. ✔
- [18] OpenAI (2022). Introducing ChatGPT. 30 Nov. 2022. ✔
- [19] OpenAI (2024). Learning to reason with LLMs. 12 Sep. 2024. ✔
- [20] Anthropic (2024). Prompt caching with Claude (GA 17 Dec. 2024). ✔
- [21] Anthropic (2024). Introducing the Model Context Protocol. 25 Nov. 2024. ✔
- [22] Google (2024). Gemini 1.5 and long context window. 15 Feb. 2024. ✔
- [23] Anthropic. Pricing. Claude Platform Docs (Oct.–Nov. 2026 snapshot). ✔
- [24] Anthropic. Steering thinking; Effort. Claude Platform Docs. ✔

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Bibliografia auditada (3/3)

✔ verificada em fonte primária ◐ verificada em parte ou por fonte secundária ○ não verificada nesta sessão

- [25] OpenAI. Models; Pricing; Reasoning models. API Docs. ✔
- [26] Google. Gemini Developer API pricing. ✔
- [27] Google. Gemini thinking. ✔
- [28] xAI. Models; Pricing; Reasoning. xAI Docs. ✔
- [29] DeepSeek. Models & Pricing. API Docs. ✔
- [30] Anthropic. Using the Messages API. Claude Platform Docs. ✔
- [31] LangChain. LangGraph: persistence; use time-travel. ✔
- [32] Pydantic. Migration guide (V2). ✔
- [33] Anthropic (2026). Claude está fornecendo respostas incorretas ou enganosas… Help Center, 5 Mar. 2026. ✔
- [34] Silva, A. (2026). SGAEIA. Zenodo. DOI 10.5281/zenodo.22557796 (não aberto nesta sessão). ○
- [35] Silva, A. (2026). Artigos no DEV (dev.to/aridiosilva), out. 2026. ◐
- [36] Silva, A. (2026). Compêndio técnico (.docx), 1 nov. 2026 (material de origem). —

Itens do compêndio não verificados (OPA, ISO/IEC 42001, NeMo Guardrails, Llama Guard, Nygard, Brown, Wei) constam apenas no documento completo, marcados como ○.

© Aridio Silva @aridiosilva — Novembro 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995