FRONTIER LLMs — SINGLE-TURN × MULTI-TURN

Analysis of Single and Multi-Turn Interactions in Frontier AI Models (LLMs)

State, cost, performance, security and governance of conversational interaction with LLMs



Aridio Silva — Independent Researcher, Brazil

ORCID 0009-0008-2411-6995 · CC BY 4.0

© Aridio Silva @aridiosilva — November 2026

EXECUTIVE SUMMARY

Five findings on single-turn × multi-turn

~T^2

APIs keep no state: history is resent each turn and cumulative input grows ~T^2

Cache

Caching cuts input cost; reasoning (output) then dominates the bill

−39%

Laban et al.: 15 LLMs lose 39% in multi-turn; unreliability +112%, aptitude −16%

Risk

Multi-turn widens the surface: Crescendo, indirect injection and persistent state

≠ more

More reasoning does not fix lost reliability; consolidating and governing helps more

Importance

Drives architecture, cost and risk of any LLM product

Positive implications (+)

Mature techniques exist for each axis: cache, effort, state, policy

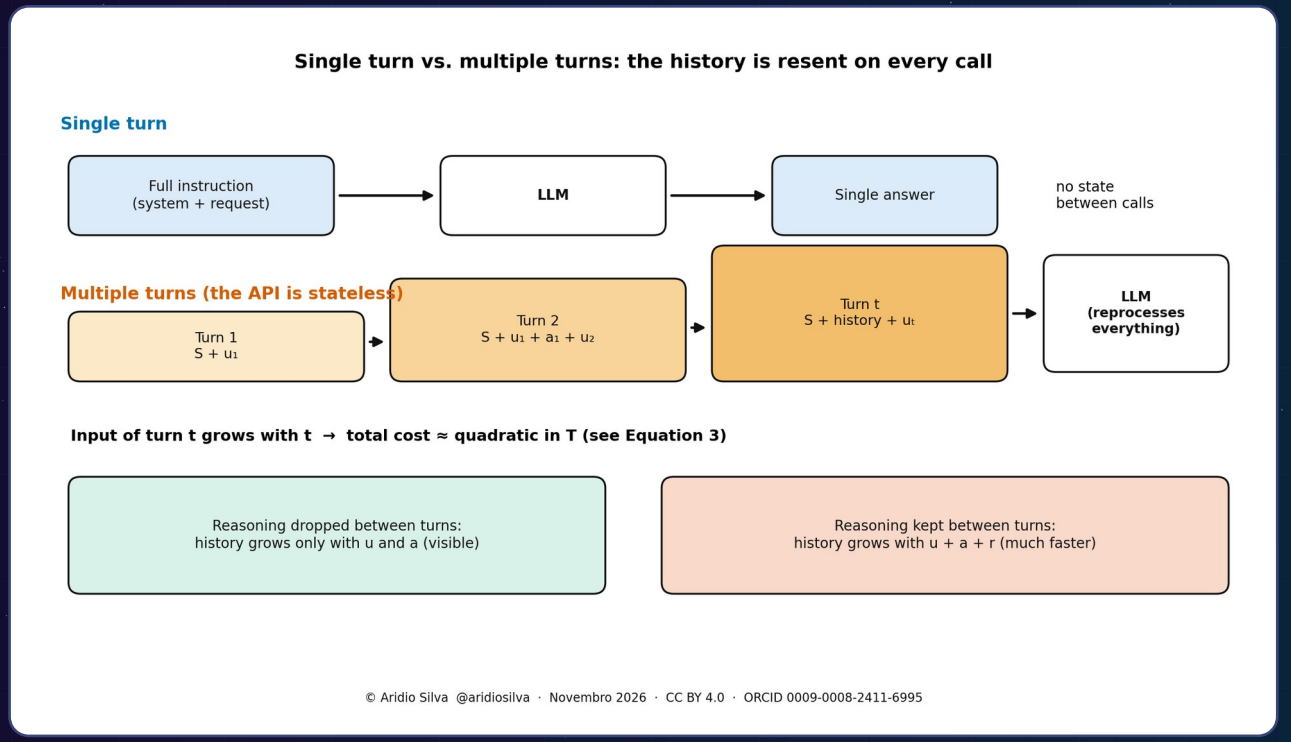
Negative implications (−)

No comparable public data per model and per effort level

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

Single turn versus multiple turns

- **Single turn:** one complete instruction, one answer; no state between calls
- **Multi-turn:** each answer is conditioned on the whole history, resent on every call
- **Who keeps the state?** the client, not the model (stateless API)
- **Not long context:** multi-turn is defined by incremental disclosure of information



Importance

Basis of the whole analysis: defines cost, risk and reliability

Positive implications (+)

Single turn: simple, cheap, testable, smaller attack surface

Negative implications (−)

Multi-turn: ~T^2 cost, cumulative latency and risk

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

COST MECHANICS

What is quadratic, and what is not

- **Within a call:** attention costs $O(n^2 \cdot d)$ per layer (Vaswani et al.)
- **Across the conversation:** the sum of resent tokens grows $O(T^2)$
- **Compendium fix:** quadratic is not exponential

236k

input tokens over 20 turns (reasoning dropped)

1.756M

keeping 8,000 reasoning tokens per turn (7.4×)

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

© 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

Importance

Explains why long conversations get costly even with caching

Positive implications (+)

Gives a simple formula to forecast and budget cost

Negative implications (−)

Ignores tokenizer variation and cache-write surcharges

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

COMPARISON

Matrix: single-turn × multi-turn

Criterion	Single-turn	Multi-turn
State	Stateless	Stateful: needs history management
Cost	Low and linear per request	Growing, ~T^2 without cache
Latency	Minimal	Scales with context
Risk	Isolated and smaller	Cumulative (context and state)
Ideal cases	Classification, extraction, translation	Interactive help, code, agents
Information	Complete at the start	Incremental; allows underspecification

Importance

Helps choose the right mode for each task

Positive implications (+)

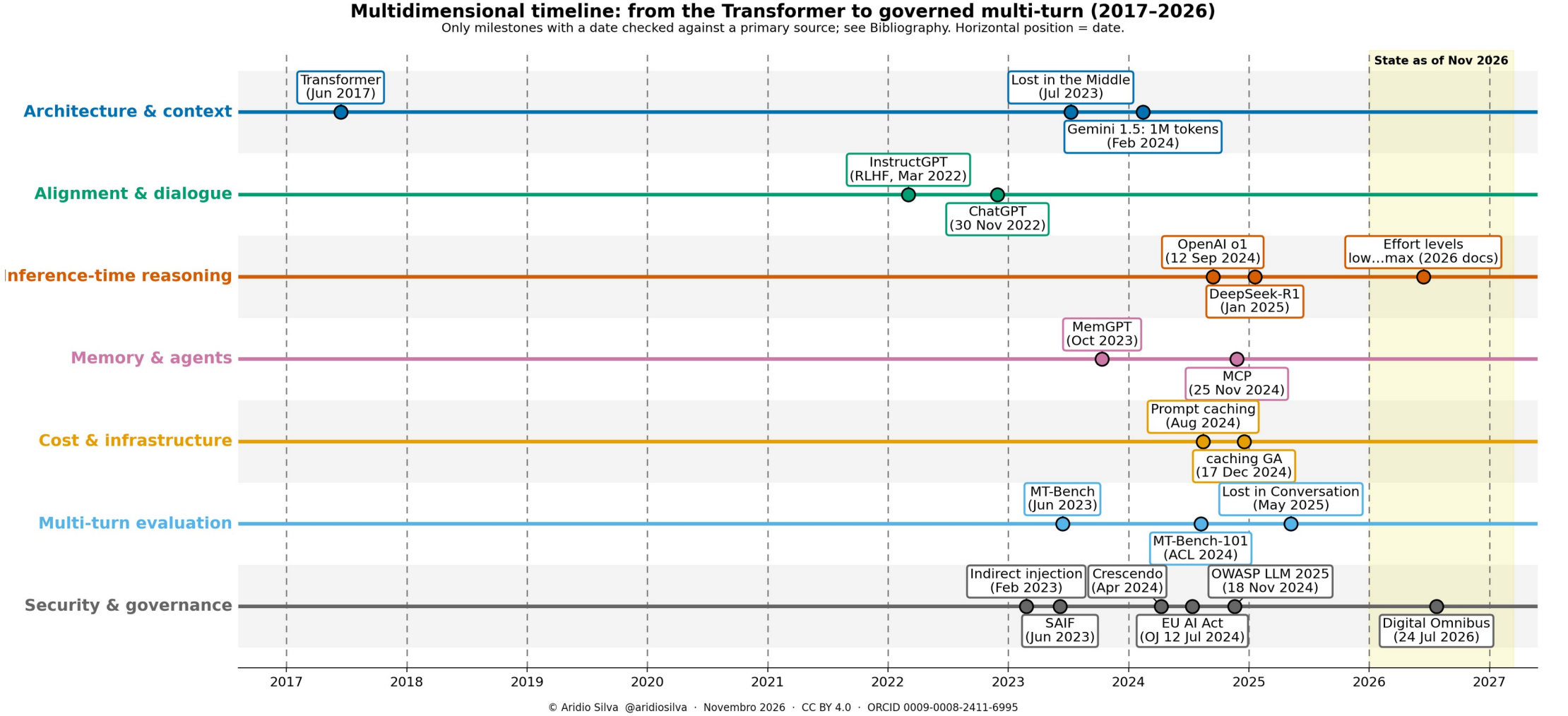
Single turn for well-specified tasks: predictable and cheap

Negative implications (−)

Undisciplined multi-turn: growing cost, drift and risk

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

Multidimensional timeline, 2017–2026



Five eras: from the isolated prompt to governed conversation

2017–2021 Architecture

Transformer; 512–2,048-token windows; prompting use (essentially single-turn)

2022 Dialogue

InstructGPT (RLHF) and ChatGPT: multi-turn becomes a product format

2023 Context and attacks

Lost in the Middle, MemGPT, MT-Bench, indirect injection, SAIF

2024 Cache and reasoning

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

2025–26 Reliability

R1, Lost in Conversation, effort levels, Digital Omnibus

Importance

Shows evaluation, pricing and security arrived after adoption

Positive implications (+)

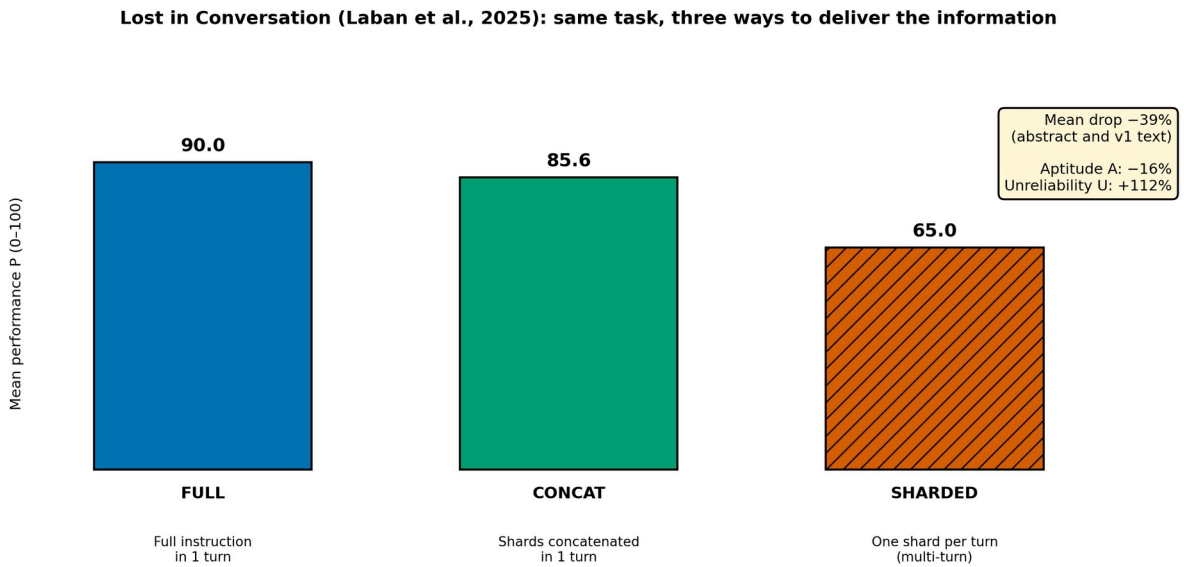
Fast innovation: caching, memory paging and protocols within years

Negative implications (–)

2026 milestones without a primary-source date were left out

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

Lost in Conversation: same task, three deliveries



- **15 LLMs, 6 tasks,** over 200 thousand simulated conversations
- **Mean drop of 39%;** aptitude −16%, unreliability +112%
- **Already at 2 turns;** CONCAT = 95.1% of FULL: the loss is not from rephrasing
- **Reasoning does not save:** o3 and R1 degrade alike, with answers 33% longer

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

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

Importance

Shows single-turn performance overestimates real use

Positive implications (+)

Clear metrics (P, A, U) and a controlled design on the same tasks

Negative implications (−)

2025 models, analytical English tasks; v1 shows −35% and −39%

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

PERFORMANCE EVIDENCE

Why models get lost, and what helps

Premature answer

Early assumptions contaminate the rest

Bloat

Final answers 20–300% longer

Middle turns

Middle turns are forgotten

Verbosity

More text, more assumptions

- **Consolidate:** gather requirements in a single prompt
- **Recap:** repeating requirements (SNOWBALL) recovers 15–20% of the drop
- **Restart:** a new conversation with everything consolidated
- **Temperature 0 does not fix it:** ~30 points of unreliability remain

$$\Pr(\text{no error in } T \text{ turns}) = (1-\varepsilon)^T \approx e^{-\varepsilon T}$$

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

Importance

Gives immediate practical rules for users and builders

Positive implications (+)

Consolidating and recapping improves reliability without switching models

Negative implications (−)

Eq. 6 is a didactic model (ε = 5%, T = 20 → ~36%), not an empirical result

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

Reasoning levels by provider (Nov. 2026)

Provider	Model	Levels	Default
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	with or without reasoning	with reasoning

Importance

Effort is the main cost lever in long conversations

Positive implications (+)

Similar scales allow comparing providers at equivalent levels

Negative implications (−)

Names and availability change; *varies by model; check official pages

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

Prices and the weight of cache reads (USD/M tokens)

Model	Input	Output	Cache	Cache/in.
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 k)	2.00	6.00	0.50	25%
DeepSeek V4-Pro (peak)	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%

- **Tokenizer:** Claude ≥4.7 models yield ~30% more tokens: per-token price is not 1:1
- **Tiers:** Haiku 5.5 above 100k; Grok doubles above 200k
- **Dates:** Gemini 4 Argon doubles on 1 Dec. 2026
- **DeepSeek:** half price off-peak

Importance

In multi-turn, cache reads weigh more than the input price

Positive implications (+)

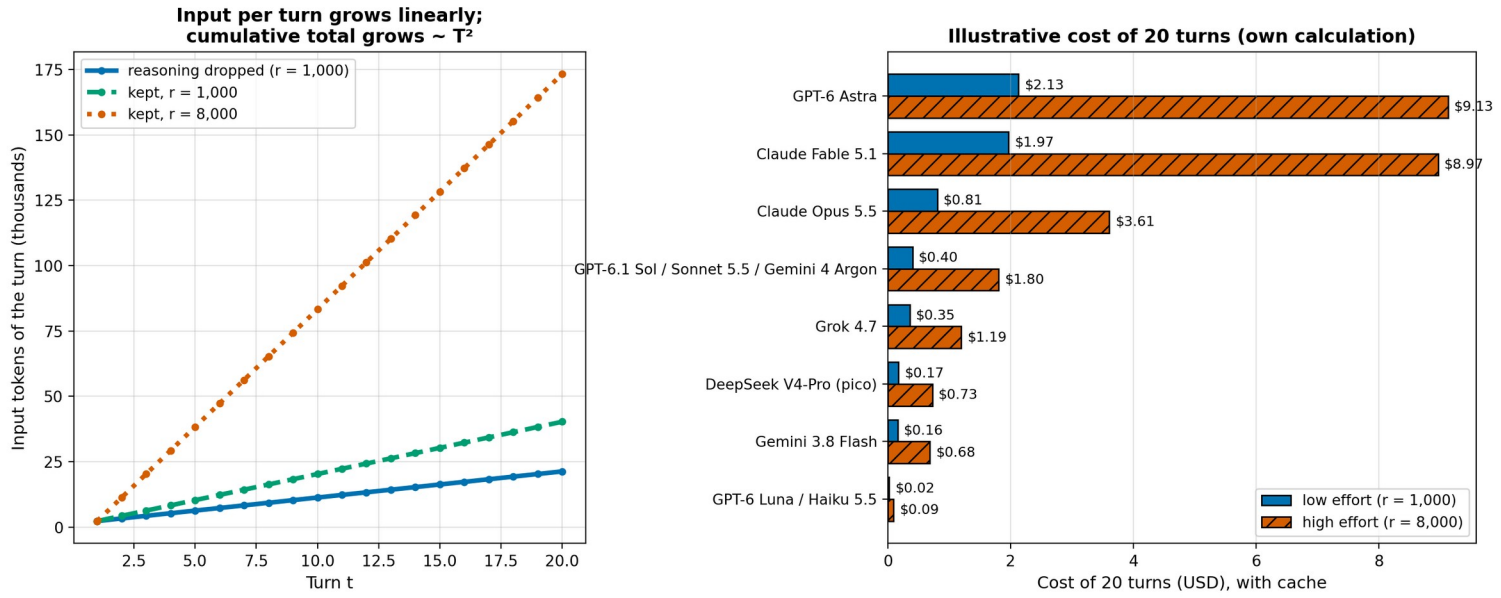
Cache at 2–5% of input (Anthropic, DeepSeek) cheapens long conversations

Negative implications (–)

Grok charges 25% of input for cache: smaller edge in long conversations

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

20-turn example: effort weighs more than turns



Assumptions: S = 2,000, u = 300, a = 700 tokens; cache always hits; no cache-write surcharge; prices: each provider's official docs (Nov 2026 snapshot).

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

3.4–4.5×
high-over-low effort cost

−86%
input cost with cache (5% reads)

USD 0.40
20 turns, low effort, Sol/Sonnet/Argon (own calculation)

Importance

Shows where the money goes: output and reasoning, not history

Positive implications (+)

With caching, long conversations become predictable and cheap

Negative implications (−)

Keeping reasoning across turns multiplies input ~7×; own calculation, not provider data

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

SECURITY

Multi-turn widens the attack surface

Crescendo: gradual escalation

Only benign inputs; the model follows its own recent text. Benchmarks were single-turn.

Indirect prompt injection

Retrieved data acts as instructions; the injection persists in history and hijacks later turns.

Hallucination cascade

An error accepted in turn 2 becomes official state and contaminates the rest.

Unbounded consumption

Long conversations with high reasoning become a cost attack (OWASP LLM10:2025).

Importance

Authorization cannot depend on the model's judgment across the conversation

Positive implications (+)

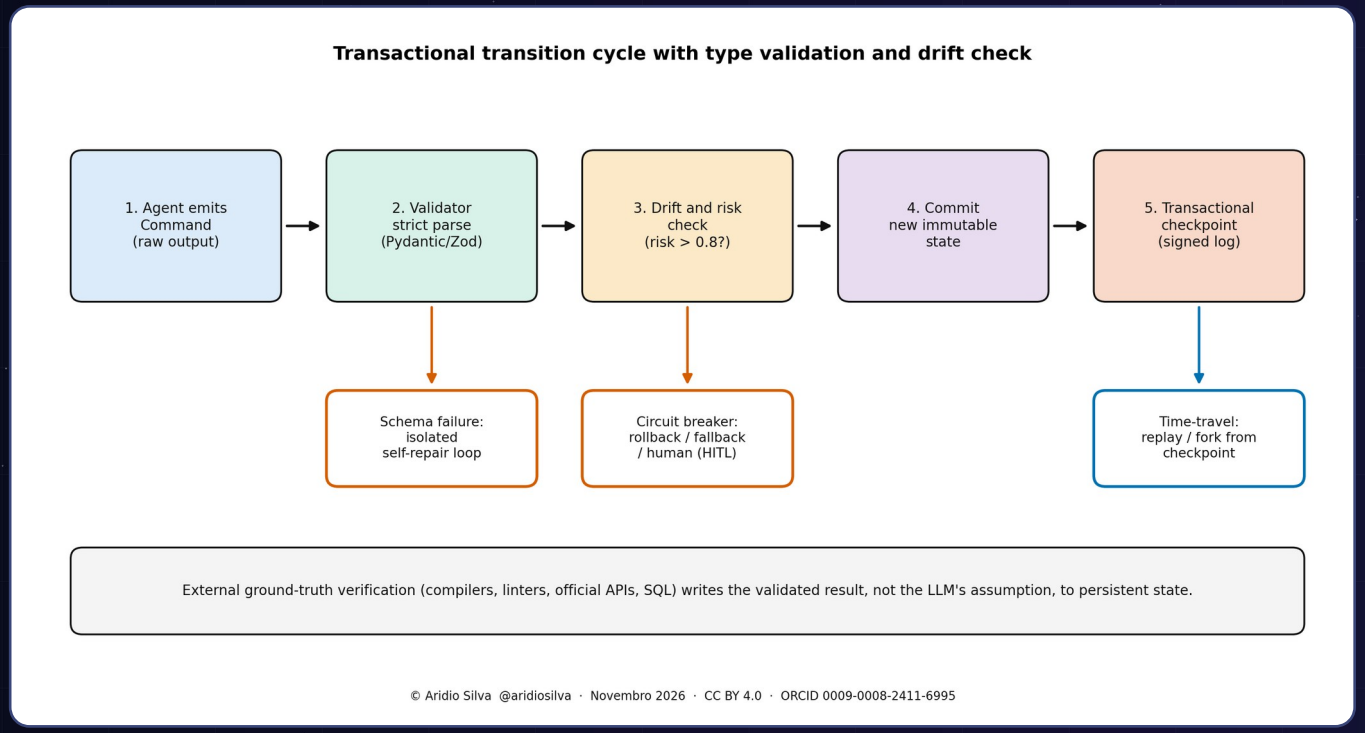
Mature taxonomies (OWASP, SAIF) and experimental evidence exist

Negative implications (−)

Defenses based on internal representations fail against gradual attacks

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

Typed state and drift mitigation



- **Semantic drift:** re-anchor instructions; goal-graph summaries
- **Behavioral drift:** circuit breakers and critic agents
- **Epistemic drift:** write the verified result, not the assumption
- **Type contracts:** Pydantic/Zod before touching global state

Importance

Agent state should be an engineering artifact, not chat text

Positive implications (+)

Malformed outputs are blocked; checkpoints enable audit and recovery

Negative implications (−)

Schema validators miss semantic errors; the risk threshold (0.8) is heuristic

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

Time-travel debugging: what replay really does

Replay and fork

Replay resumes from a checkpoint; fork branches with modified state. Checkpoints at each super-step (LangGraph).

Audit: replay re-executes

LLM and API calls fire again and may differ. Determinism only if outputs are recorded. The docs do not mention seeds.

Notes on the Pydantic code

class Config and .dict() are deprecated in v2 (model_config, model_dump); imports missing; frozen=True is shallow (prior knowledge, unverified).

Importance

Debugging agents requires recording events, not just states

Positive implications (+)

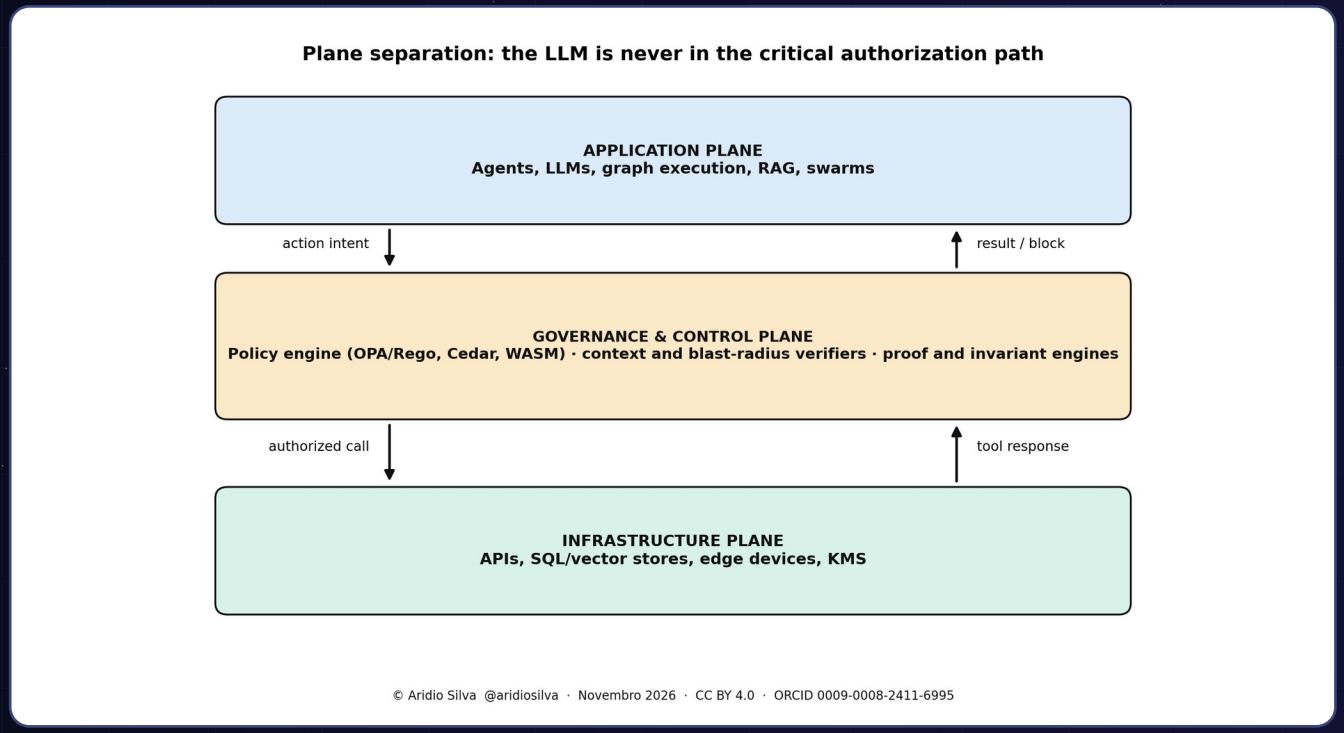
Roll back and fork without reprocessing everything; forensic trail

Negative implications (−)

Rollback needs reversible external effects; unrecorded replay misleads

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

The LLM is never in the critical authorization path



- **Identity:** X.509/SPIFFE per agent; strict-subset delegation
- **Bounded authority:** granular, temporary access (OPA/Rego, Cedar)
- **Interception:** pre-action gate, circuit breakers and post-action auditors
- **Tamper-evident trail:** signed, append-only logs

Importance

Security becomes a property of the system, not the model

Positive implications (+)

Deterministic policy is immune to prompt engineering

Negative implications (−)

Cost, deployment complexity and the need for PKI and policy-as-code

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

GOVERNANCE

Risk matrix and the regulatory calendar

Level	Behavior	Example
Low	Auto-approved, signed log	Read documentation
Medium	Notify and proceed	Interim report
High	Pause; human approval (HITL)	Financial transaction
Inadmissible	Immediate block (fail-closed)	Privilege escalation

EU AI Act, Arts. 12 and 14

High-risk systems only. Logs contribute to compliance but do not guarantee it.

New date: 2 Dec. 2027

Reg. (EU) 2026/1744 (Digital Omnibus) postponed Annex III. Secondary sources; EUR-Lex not opened.

Importance

Links technical controls to regulatory duties; NIST AI 100-1 is the correct reference

Positive implications (+)

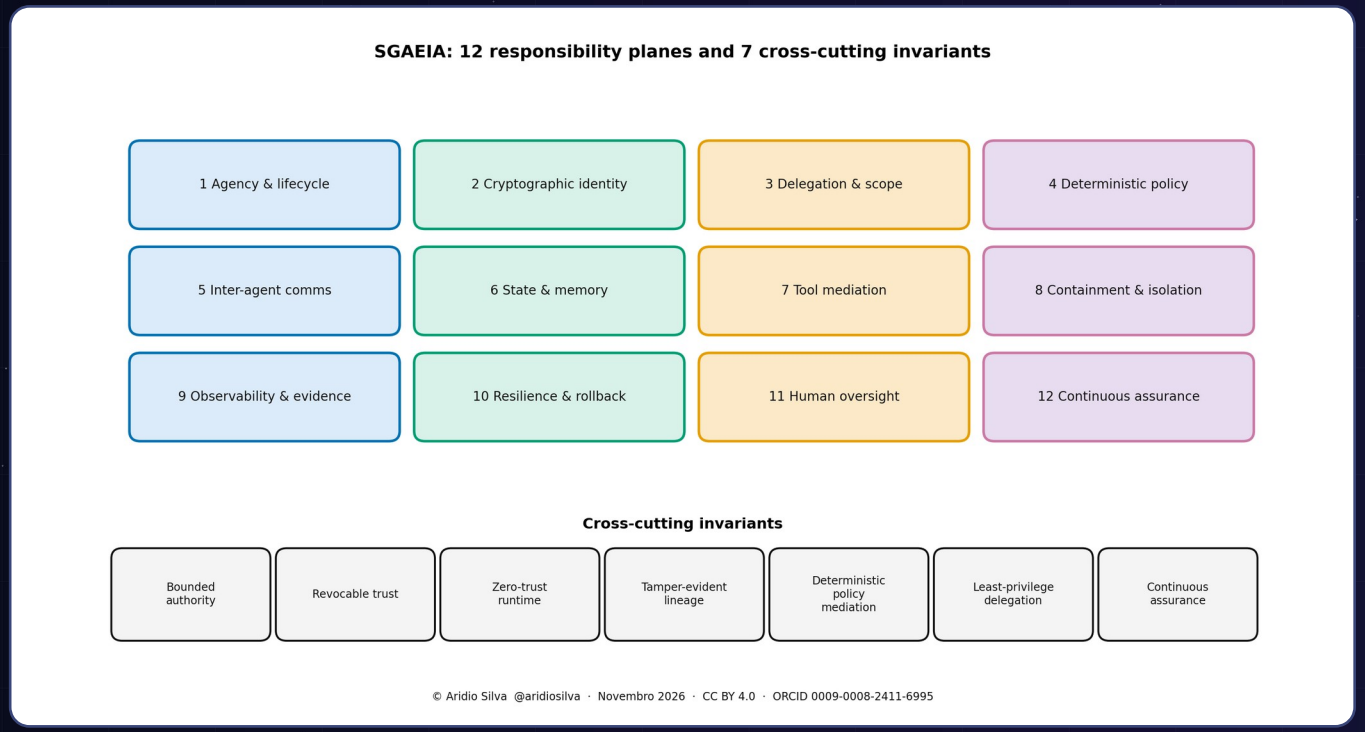
Classifying actions by risk gives predictability and human oversight where it matters

Negative implications (−)

Dates change; HITL creates a human bottleneck; immutable trails raise data volume

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

12 responsibility planes and 7 invariants



- **Premise:** the object of assurance is the complete system in operation, not the isolated model
- **Answers multi-turn:** planes 6 (state), 10 (rollback) and 12 (continuous assurance)
- **Answers injection:** planes 3, 4 and 7 (delegation, policy, mediation)
- **Author's work:** no published independent evaluation

Importance

Open, neutral reference architecture for autonomous agents and edge AI

Positive implications (+)

Integrates identity, policy, immutable state and Evidence-as-Code

Negative implications (−)

Requires operational maturity; empirical efficacy not yet shown by third parties

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

SOURCE RIGOR

Audit: 8 corrections to the compendium

Compendium passage	Problem	Correction
Lost in the Middle: author 'Manning', 2023	Petroni missing; TACL 12:157–173 (2024)	Authors and venue
NIST AI RMF 'SP 1270'	The document is NIST AI 100-1	Number fixed
Cost 'grows exponentially'	$O(n^2)$ is quadratic	Term fixed
Logs 'guarantee' compliance	High-risk only; dates changed	Reworded
Time-travel 'deterministic'	Replay re-executes LLM/API	Qualified
'LangChain Architecture Team'	Authorship not verifiable	Official docs
'Cedar Specification (AWS)'	Title not found	Cutler et al. (2024)
Pydantic code	class Config and .dict() deprecated	Notes + v2 variant

Importance

In technical work, every LLM-generated reference needs checking against a primary source

Positive implications (+)

Simple, reproducible protocol; found real errors in already-written material

Negative implications (−)

Takes time; the Zenodo DOI and legal texts could not be opened

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

State of the art, Nov. 2026

Multi-turn reliability as a goal

Equal aptitude in 1 and N turns; U < 15 at T = 1.0

Economics by cache and effort

Cache cheapens input; effort controls output

State as engineering

Typed, immutable, versioned; checkpoints and fork

Runtime governance

Deterministic policy, identity, tamper-evident trail

Integration protocols

MCP standardizes tools and widens indirect injection

Importance

Summarizes where practice stands today on five axes

Positive implications (+)

Each axis has documented, mature techniques

Negative implications (−)

Public evaluations per model and effort level are missing

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

Gaps and future outlook

Horizon	Outlook	Anchoring
Short term	Reliability benchmarks per effort level; tokenizer-normalized cost	Laban et al.; pricing docs
Short term	Products that consolidate requirements and summarize state by default	Laban; MemGPT
Medium term	Deterministic policy and agent identity by default	Cedar; NIST; OWASP
Medium term	Automatically generated compliance evidence	AI Act; Omnibus
Long term	Models that cut early assumptions and keep consistency	Laban's causes

Importance

Points to the research and product agenda for the coming years

Positive implications (+)

Advantage likely shifts from raw capability to reliability per cost (author's inference)

Negative implications (−)

Inferences are the author's, not cited facts; strongest results are from 2025

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

CONCLUSION

Single or multi-turn is an architecture decision

- **Single turn** for well-specified tasks: cheap, predictable, testable
- **Multi-turn** when exploring, with consolidated and recapped requirements
- **Cost:** use caching and control effort; reasoning dominates the bill
- **Reliability:** do not rely on more reasoning; restart and consolidate
- **Security:** typed state, deterministic authorization outside the LLM, auditable trail
- **Rigor:** verify every reference against a primary source



Read the full document (PDF/MD) in PT and EN.

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

Glossary of technical terms (1/3)

- **Aptitude (A^{90}):** Best-case performance (90th percentile) over repeated simulations
- **Attention:** Each position weighs all others; $O(n^2 \cdot d)$ cost per layer
- **Prompt cache:** Reuse of the already-processed prefix, billed at a lower read price
- **Chain of thought:** Reasoning steps generated before the final answer
- **Checkpoint:** Persistent state snapshot; basis for replay and fork
- **Circuit breaker:** Stops the flow when a risk threshold is exceeded
- **CONCAT / FULL / SHARDED:** Laban conditions: concatenated, full, one shard per turn
- **Context window:** Maximum tokens processed at once
- **Hallucination cascade:** A wrong premise accepted in one turn that contaminates later ones
- **Drift:** Semantic, behavioral or epistemic deviation
- **Reasoning effort:** How much the model 'thinks' before answering (low to max)
- **Event sourcing:** Persisting each transition as an immutable event
- **Evidence-as-Code:** Compliance evidence generated and verifiable as code
- **Fork:** Branch from a checkpoint with modified state

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

Glossary of technical terms (2/3)

- **Guardrail:** Programmable layer that filters inputs and outputs
- **HITL:** Explicit human approval before a critical action
- **Indirect injection:** Malicious instructions in retrieved data treated as commands
- **Jailbreak:** Inducing the model to violate its policies
- **KV cache:** Attention keys and values reused during generation
- **Lost in the Middle:** Performance drop when information sits in the middle
- **MCP:** Open standard to connect assistants to tools and data
- **Multi-turn:** Message sequence, each answer conditioned on the history
- **Unreliability (U):** Gap between the 90th and 10th performance percentiles
- **OPA / Rego:** Policy-as-code engine and language
- **Cedar:** Open-source authorization language with formal analysis
- **Prefill:** Phase where the model processes the prompt before generating
- **RAG:** Retrieval-augmented generation
- **RLHF:** Reinforcement learning from human feedback

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

Glossary of technical terms (3/3)

- **Sliding window:** Keeps only the N most recent interactions
- **SNOWBALL / RECAP:** Repeating requirements every turn or at the end
- **Stateless / stateful:** No state between calls / state kept
- **Test-time compute:** Compute spent at inference, such as reasoning tokens
- **Token:** Unit of text processed and billed by the model
- **Reasoning tokens:** Thinking tokens; billed as output
- **Time-travel debugging:** Replay and fork of executions from checkpoints
- **Turn:** One message from a role (user or assistant)
- **TTFT:** Time until the first token of the response
- **Unbounded Consumption:** Uncontrolled inference and cost consumption (OWASP LLM10)
- **Zero Trust:** Trust no entity or transition implicitly
- **Blast radius:** Extent of damage if an authorized action fails
- **FSM / DAG:** Finite state machine / directed acyclic graph
- **Agent identity:** Unique cryptographic identity per agent (SPIFFE/X.509)

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

Audited bibliography (1/3)

✔ verified against a primary source ◐ partly or secondarily verified ○ not verified in this session

- [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 — November 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Audited bibliography (2/3)

✔ verified against a primary source ◐ partly or secondarily verified ○ not verified in this session

- [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 — November 2026 · CC BY 4.0 · ORCID 0009-0008-2411-6995

Audited bibliography (3/3)

✔ verified against a primary source ◐ partly or secondarily verified ○ not verified in this session

- [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). —

Compendium items not verified (OPA, ISO/IEC 42001, NeMo Guardrails, Llama Guard, Nygard, Brown, Wei) appear only in the full document, marked ○.

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