

● GOVERNED AUDITABLE TRUST ENGINEERING

IA de Engenharia Auditável

O que mudou na inteligência artificial em 2026, por que a ontologia voltou ao centro, como um squad de agentes trabalha e por que modelos e infraestrutura decidem de quem é o conhecimento.

● APRESENTAÇÃO

A engenharia de megaprojetos não sofre de falta de técnica. Sofre de entropia informacional.

Dados fragmentados, sem padrão e verificados à mão se degradam a cada fronteira entre organizações, disciplinas e sistemas. O resultado aparece no custo e no prazo de quase todo grande empreendimento de energia, óleo e gás e infraestrutura.

Este e-book acompanha a palestra **IA de Engenharia Auditável** e organiza, em quatro capítulos, o raciocínio que a sustenta: o salto recente da IA, a ontologia como endereço do conhecimento, o squad de agentes que trabalha sobre ela e a camada de modelos e hardware que determina quem fica com o que foi aprendido.

A pergunta de fundo é uma só: como acumular competência técnica de engenharia ao longo de muitos projetos, como faz um bom engenheiro, sem guardar os dados confidenciais de nenhum cliente.

CUSTO E PRAZO

8,5%

dos 16.000 projetos analisados atingem custo e prazo ao mesmo tempo.

Flyvbjerg e Gardner (2023)

MEGAPROJETOS

9 em 10

estouraram o custo, com desvios de até 50% em termos reais.

Flyvbjerg (2014)

A ENTROPIA INFORMACIONAL DE UM MEGAPROJETO



Cada fronteira entre organizações e sistemas é um ponto de reformatação, de verificação manual e de perda de rastro.

Esquema ilustrativo: pontos verdes representam informação ainda verificável na origem.

SUMÁRIO

01 · p. 3

Contexto da IA

Março a setembro de 2026: modelos, hardware, investimento, matemática, disputas e agentes.

02 · p. 11

Ontologia

De Aristóteles aos grafos de conhecimento que ancoram a IA na engenharia.

03 · p. 15

Squad de agentes

O produto reflete o time: julgamento, gates, aprendizado e custo de execução.

04 · p. 19

Modelos e infraestrutura

Propriedade intelectual, soberania, independência e execução local.

01

● CONTEXTO DA IA

Seis meses que redesenharam o mapa

De março a setembro de 2026, a IA deixou de ser uma corrida de modelos e virou uma disputa de infraestrutura, de propriedade intelectual e de tráfego. Os números abaixo resumem o período.

LANÇAMENTOS NOTÁVEIS

48

modelos relevantes em sete meses, 19 deles com pesos abertos.

CAPEX 2026

~835 bi US\$

guidance somado de Amazon, Alphabet, Microsoft, Meta e Oracle.

TRÁFEGO WEB

57,5%

das requisições HTTP vêm de bots, contra 42,5% de humanos (Cloudflare).

1.1 · MODELOS

Um lançamento relevante a cada quatro dias

Entre março e setembro de 2026, os laboratórios lançaram modelos em ritmo quase semanal. A seleção abaixo reúne 48 lançamentos notáveis: 29 fechados, de fronteira, e 19 com pesos abertos. Agregadores que contam variantes e modelos de imagem, vídeo e voz passam de 130 entradas no mesmo período.

● Fechado ● Pesos abertos

MAR 2026	ABR	MAI	JUN	JUL	AGO	SET
7 lançamentos	12 lançamentos	4 lançamentos	7 lançamentos	6 lançamentos	7 lançamentos	5 lançamentos
GPT-5.4 OpenAI	Gemma 4 Google	Gemini 3.5 Flash Google	MiniMax M3 MiniMax	GPT-5.6 OpenAI	Qwen3.8-Max Alibaba	Claude Fable 5.1 Anthropic
Gemini 3.1 Flash-Lite Google	Claude Mythos Preview Anthropic	Qwen3.7 Max Alibaba	Nemotron 3 Ultra NVIDIA	Grok 4.5 xAI	Muse Spark 1.2 Meta	Gemini 3.8 Flash Google
Phi-4-reasoning-vision Microsoft	GLM-5.1 Zai	Mistral Medium 3.5 Mistral	MAI-Thinking-1 Microsoft	Muse Spark 1.1 Meta	Grok 4.6 xAI	Muse Spark 1.3 Meta
Grok 4.20 xAI	Muse Spark Meta	Claude Opus 4.8 Anthropic	Gemma 4 12B Google	Kimi K3 Moonshot	Muse Glimmer 30B Meta	GPT-6 Astra OpenAI
Nemotron 3 Super NVIDIA	Claude Opus 4.7 Anthropic		Claude Fable 5 Anthropic	Gemini 3.6 Flash Google	Gemini 3.7 Flash Google	Grok 4.7 xAI
Mistral Small 4 Mistral	Qwen3.6 27B e 35B Alibaba		GLM-5.2 Zai	Claude Opus 5 Anthropic	Qwen3.8-27B Alibaba	
GPT-5.4 mini e nano OpenAI	Kimi K2.6 Moonshot		Claude Sonnet 5 Anthropic		GLM-5.3 Zai	
	GPT-5.5 OpenAI					
	DeepSeek V4 DeepSeek					
	Nemotron 3 Nano Omni NVIDIA					
	Qwen3.6 Max Alibaba					
	Grok 4.3 xAI					

Quem chegou ao topo

OpenAI lançou GPT-5.4, 5.5, 5.6 e o GPT-6 Astra em 3 de setembro. A Anthropic passou por Opus 4.7, 4.8, Fable 5 e Opus 5. O Google empilhou versões Flash, da 3.5 à 3.8.

A distância encurtou

Segundo a Epoch AI, desde janeiro de 2026 os melhores modelos abertos ficam cerca de **4 meses** atrás dos fechados. DeepSeek V4, Kimi K3 e Qwen3.8 puxaram essa curva.

Novos entrantes

A Meta estreou a família Muse, da Superintelligence Labs. A Microsoft apresentou modelos próprios, MAI-Thinking-1 e MAI-Code-1. A Amazon reorganizou sua pesquisa de modelos.

Fontes: notas de lançamento das empresas; LLM Gateway, linha do tempo 2026; Epoch AI, "Open vs. closed gap" (29/05/2026). Seleção e classificação COODEX.

1.2 · MODELOS LOCAIS E HARDWARE

A inteligência cabe na mesa. E o hardware passou a valorizar.

Modelos que rodam em um notebook ou em uma estação de trabalho já entregam o que era fronteira há menos de um ano. Ao mesmo tempo, a demanda por computação inverteu uma regra antiga: equipamento usado não está perdendo valor.

RODANDO LOCALMENTE

Alibaba · abr/2026

Qwen3.6-27B

77,2% no SWE-bench Verified, acima do Claude Sonnet 4 de maio de 2025 (72,7%). Licença Apache 2.0.

Google · abr/2026

Gemma 4

As versões E2B e E4B rodam offline em celulares e placas embarcadas. Licença Apache 2.0.

Meta · ago/2026

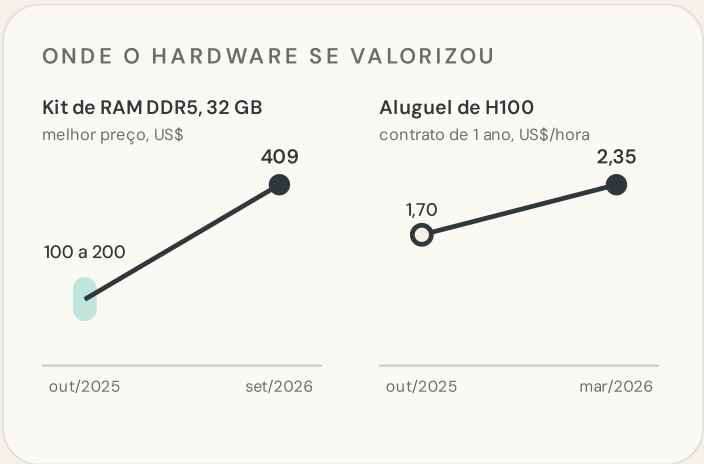
Muse Glimmer 30B

Em 4 bits ocupa de 17 a 20 GB e roda em uma GPU de 24 GB ou em um Mac M5 Max.

Apple · mar/2026

MacBook Pro M5 Max

Até 128 GB de memória unificada a 614 GB/s, em um notebook.



É a primeira vez?

Não no sentido literal: houve o superciclo de DRAM em 2017 e 2018 e a escassez de GPUs de 2020 e 2021. O inédito é ver GPUs de data center com anos de uso subindo de preço, e uma alta de memória puxada pela IA, e não pelo consumo. A Amazon atribuiu parte do aumento do seu capex de 2026 ao custo maior de memória.

DRAM, 1T26

+93 a 98%

preço de contrato sobre o trimestre anterior.

ALUGUEL H100

+40%

contrato de 1 ano: US\$ 1,70 para US\$ 2,35 por hora.

REVENDA B200

158%

do preço de lançamento, valor residual reportado.

MICRON

+229%

na bolsa em 2026, até julho, com a corrida por memória.

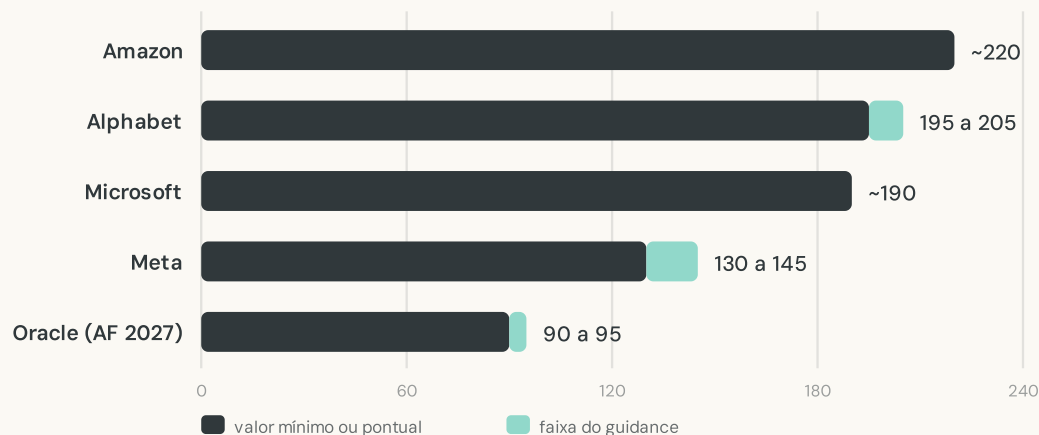
Fontes: The Decoder (25/04/2026); Google, blog Gemma 4; InfoQ (ago/2026); Apple Newsroom (03/03/2026); TrendForce (01/06/2026); SemiAnalysis; Silicon Data via Tech Insider (set/2026); Tom's Hardware, RAM Price Index (14/09/2026); Yahoo Finance (08/07/2026).

1.3 · INVESTIMENTO EM INFRAESTRUTURA

O maior projeto de infraestrutura já empreendido

A expressão é do Gartner. Em 2026, só cinco empresas planejam investir perto de US\$ 835 bilhões em data centers, chips e energia. Em 2022, a Alphabet sozinha investia US\$ 31 bilhões.

CAPEX PREVISTO PARA 2026, EM BILHÕES DE US\$



Guidance após os resultados do 2º trimestre de 2026. Faixas indicam o intervalo divulgado. Oracle: ano fiscal 2027, valor bruto. Microsoft: ano-calendário, antes do ajuste contábil de leasing.

GASTO MUNDIAL COM IA

2,7 tri US\$

em 2026, alta de 49,5%. Deste total, US\$ 1,48 tri é infraestrutura (Gartner).

STARGATE

0,3 de 9 GW

operando em Abilene, dos mais de 9 GW planejados até 2029 (Epoch AI, abr/2026).

TOKENS NO GOOGLE

3,2 quatrilhões/mês

processados por mês, 7 vezes o volume de um ano antes (Google I/O, mai/2026).

Por que isso importa para quem opera ativos. Energia virou o insumo crítico da IA. Os mesmos setores que a COODEX atende, energia, óleo e gás e infraestrutura pesada, são os que vão construir e alimentar essa capacidade. A IA é, ao mesmo tempo, ferramenta e cliente da engenharia.

1.4 · MATEMÁTICA

Quando a IA começou a resolver problemas em aberto

Em 2025, a IA ainda disputava medalhas em olimpíadas. Em 2026, passou a atacar problemas que matemáticos tentavam resolver havia décadas, e a entregar provas verificáveis por máquina, na linguagem Lean.

- mai/2026

Conjectura da distância unitária de Erdős, de 1946

Um modelo interno da OpenAI encontrou um contraexemplo e refutou a conjectura. O matemático Tim Gowers avaliou que nenhuma prova anterior feita por IA tinha chegado perto do padrão de uma revista de topo.
- mai/2026

AlphaProof Nexus, Google DeepMind

Gemini 3.1 Pro combinado com o provador Lean 4 resolveu 9 de 353 problemas de Erdős em aberto e 44 conjecturas da enciclopédia de sequências OEIS, segundo preprint no arXiv.
- jul/2026

Olimpíada Internacional de Matemática, Xangai

Primeiras notas perfeitas, 42 de 42, com correção oficial: Huawei Celia e RedNote dots-note-3.0. Em 2025, o melhor resultado de IA tinha sido 35 de 42.
- ago/2026

Dez avanços, OpenAI

Uma versão interna do GPT-6 Astra resolveu 10 problemas em aberto, com certificados em Lean, a um custo de cerca de US\$ 2.000 em tokens. Entre eles, os problemas de Erdős 146, 180 e 183.
- set/2026

Navier-Stokes, um dos Problemas do Milênio

A OpenAI afirma ter provado a formação de singularidade em 3D, com prova verificada em Lean, usando cerca de 10.000 agentes por 88 horas. O Clay Institute ainda lista o problema como aberto e fará uma avaliação sem pressa. Parte dos matemáticos questiona se a versão provada é a que importa.

565 / 652

problemas de Erdős resolvidos e em aberto na base mantida por Thomas Bloom (ago/2026). A velocidade com que a IA passou a fechá-los levou a Quanta Magazine a perguntar por que esses problemas lendários estão caindo agora.

A lição para a engenharia não é que a máquina ficou "mais inteligente". É que resultados passaram a vir acompanhados de uma verificação formal independente. É exatamente o que se espera de um parecer técnico: não basta acertar, é preciso provar como se chegou lá.

1.5 · PROPRIEDADE INTELECTUAL

O conhecimento virou o ativo em disputa

Em 2026, os tribunais passaram a decidir o que é de quem na IA: livros usados em treino, músicas, reportagens e até o conhecimento de engenheiros que trocam de empresa.

Apple v. OpenAI

AUDIÊNCIA EM 01/10/2026

Ação ajuizada em julho de 2026 no Distrito Norte da Califórnia por segredo industrial de hardware. A Apple acusa a OpenAI, a io Products, empresa de Jony Ive comprada em 2025, e dois ex-funcionários de levar conhecimento confidencial. Segundo a ação, mais de 400 ex-funcionários da Apple foram para a OpenAI. A OpenAI pediu o arquivamento e chama o caso de "bagunça criada pela própria Apple".

CASO	O QUE ESTÁ EM JOGO	SITUAÇÃO EM SET/2026
Bartz v. Anthropic	Livros usados em treino sem licença	Acordo de US\$ 1,5 bi aprovado em 22/07, cerca de US\$ 3.000 por obra. O maior acordo de copyright da história dos EUA.
NYT v. OpenAI e Microsoft	Reportagens usadas em treino	Argumentos apresentados em 08/09. O Departamento de Justiça se manifestou a favor da OpenAI.
Musk v. OpenAI	Mudança da OpenAI para fins lucrativos	Júri rejeitou as acusações por prescrição em 18/05. Musk promete recorrer.
xAI v. OpenAI	Segredo industrial	Arquivada em caráter definitivo em 15/06.
Gravadoras v. Anthropic	Letras e obras musicais	Nova ação de Sony e Warner em 29/08, somada à de Concord e UMG.
Getty v. Stability AI	Imagens usadas em treino	Recurso autorizado no Reino Unido. Nos EUA, a maior parte das acusações segue.

Se o acervo técnico é o ativo principal de uma empresa de engenharia, a pergunta certa não é qual modelo usar. É onde esse conhecimento fica quando o contrato termina.

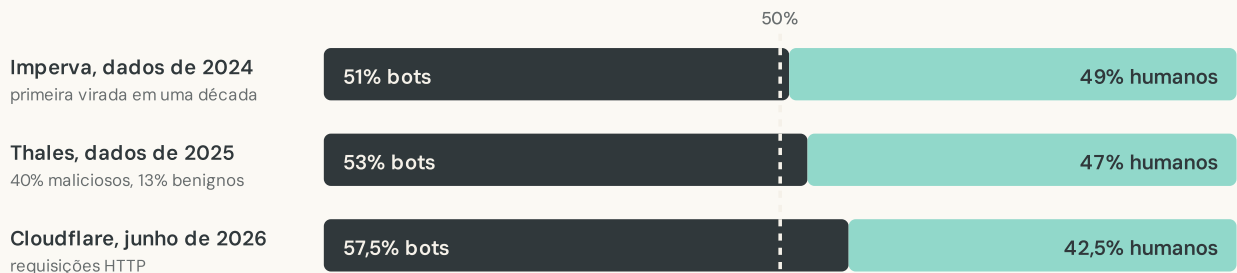
Fontes: Jurist (jul/2026); Axios (01/09 e 08/09/2026); OpenAI, "Apple is getting this wrong"; NPR (18/05/2026); Al Jazeera (15/06/2026); TechCrunch (29/08/2026); IPKat (07/01/2026); Loeb & Loeb (abr/2026). Informação jornalística, não constitui opinião jurídica.

1.6 · AGENTES NA INTERNET

A internet agora tem mais máquinas do que pessoas

Em junho de 2026, o CEO da Cloudflare afirmou que os bots passaram o tráfego humano pela primeira vez na história da internet. Ele esperava que isso só acontecesse em 2027.

PARTICIPAÇÃO NO TRÁFEGO, AUTOMATIZADO X HUMANO



Cada empresa mede de um jeito (requisições, sessões, base de clientes), então os números mostram a tendência, não uma série única.

IA AGÊNTICA

+7.851%

de crescimento no tráfego de agentes de IA em um ano, contra +3,1% do tráfego humano (HUMAN Security, abr/2026).

BOTS MOVIDOS A IA

12,5x

mais requisições bloqueadas por dia, de 2 milhões para 25 milhões (Thales, abr/2026).

Outros marcos do período

Tarefas longas. Pelo METR, o Claude Mythos Preview conclui com 50% de sucesso tarefas que levam mais de 16 horas de um especialista humano.

Receita. A Anthropic chegou a US\$ 65 bi de receita anualizada no fim de julho. A OpenAI declarou US\$ 40 bi.

Adoção. O ChatGPT tinha 900 milhões de usuários semanais em fevereiro. O app Gemini passou de 900 milhões de usuários mensais.

● SÍNTESE DO CAPÍTULO

Inteligência bruta ficou abundante. Confiança, não.

Os modelos ficaram mais capazes, mais baratos de rodar localmente e mais numerosos. O gargalo mudou de lugar. Numa indústria onde um erro custa uma parada de planta, o que falta é saber de onde veio cada resposta, quem a verificou e com base em qual documento.

01

Capacidade virou commodity

Com 48 lançamentos em sete meses e modelos abertos a poucos meses da fronteira, escolher um modelo deixou de ser vantagem competitiva.

02

Computação virou ativo escasso

Memória e GPU se valorizam. Quem controla sua própria infraestrutura controla custo, prazo e confidencialidade.

03

Conhecimento virou disputa

Os tribunais estão definindo a fronteira entre usar e se apropriar de conhecimento. Governança deixou de ser opcional.

04

Agentes viraram maioria

Se máquinas já produzem mais tráfego que pessoas, a pergunta é como auditar o trabalho delas. Os próximos capítulos respondem a isso.

A pergunta deixou de ser o que a IA consegue fazer.
Passou a ser o que dá para provar que ela fez.



02

● ONTOLOGIA

Dar endereço ao conhecimento

Antes de ser um termo de ciência da computação, ontologia foi a pergunta mais antiga da filosofia: o que existe, e como as coisas que existem se relacionam.

Rafael Sanzio, *A Escola de Atenas*, 1509 a 1511, afresco, Stanza della Segnatura, Palácio Apostólico, Vaticano. Domínio público. No centro, Platão aponta para o alto e Aristóteles estende a mão para o mundo sensível.

2.1 · ORIGEM DA PALAVRA

O estudo do ser

ὄν, ὄντος

ón, óntos

ser, ente, aquilo que é

+

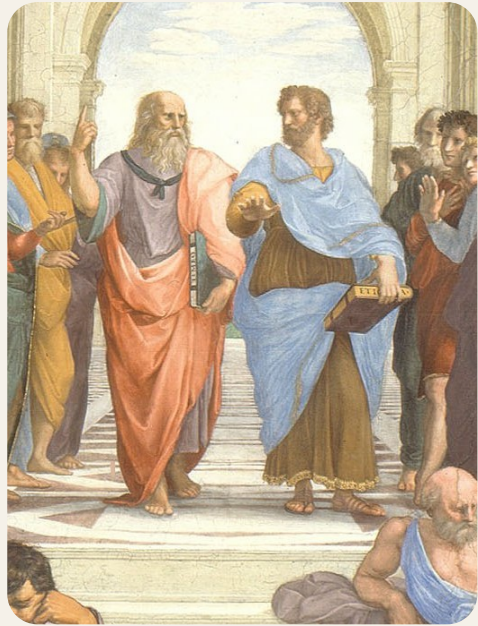
λόγος

lógos

palavra, razão, estudo

Aristóteles não usou a palavra "ontologia". Ela só apareceu no século XVII. Mas foi ele quem definiu o campo: na *Metafísica*, descreve uma ciência que estuda "o ser enquanto ser" e, nas *Categorias*, propõe um jeito de classificar tudo o que pode ser dito sobre uma coisa.

É, na prática, a primeira ontologia formal do Ocidente: um vocabulário fechado de tipos e relações para descrever o mundo sem ambiguidade. Vinte e três séculos depois, é o mesmo problema que enfrenta quem precisa fazer uma máquina entender uma folha de dados de equipamento.



Detalhe: Platão segura o *Timeu*; Aristóteles, a *Ética*. A mão espalmada para o chão costuma ser lida como a defesa do conhecimento a partir das coisas concretas.

AS DEZ CATEGORIAS DE ARISTÓTELES, APLICADAS A UMA VÁLVULA

01 Substância a válvula XV-1203, de bloqueio	02 Quantidade diâmetro nominal de 6 pol.	03 Qualidade aço inox, classe de pressão 300	04 Relação instalada na linha 6"-P-1001	05 Lugar unidade 200, área de processo
06 Tempo revisão C da folha de dados	07 Posição montada na horizontal	08 Posse tem atuador e certificado de teste	09 Ação bloqueia o fluxo	10 Paixão é acionada pelo sistema de controle

Exemplo ilustrativo. TAG e dados fictícios.

DA FILOSOFIA À COMPUTAÇÃO



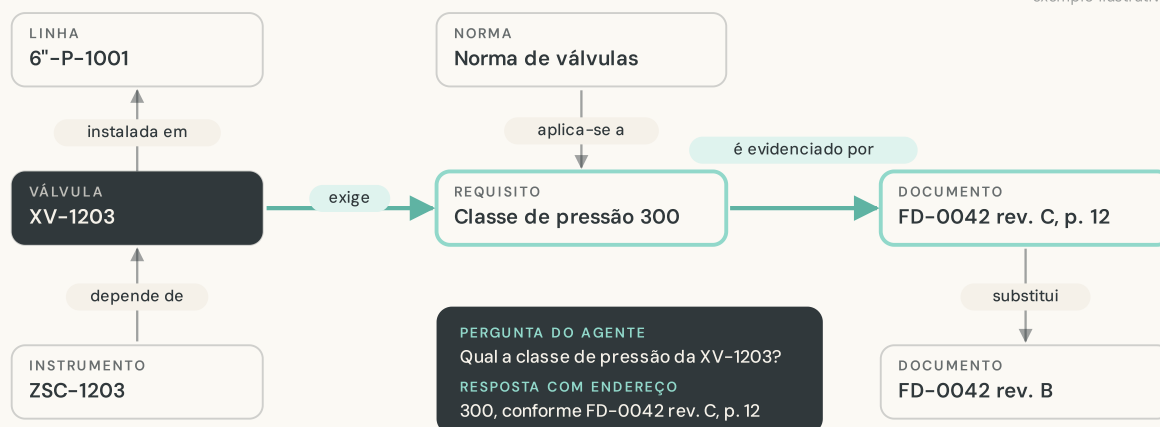
2.2 · APLICAÇÃO EM IA

Um modelo de linguagem sabe falar. A ontologia diz sobre o quê.

Na definição clássica de Tom Gruber, de 1993, uma ontologia é "uma especificação explícita de uma conceitualização". Em engenharia, é o mapa que diz o que é um equipamento, uma linha, um instrumento, um requisito, e como cada um se liga aos documentos que o descrevem.

COMO UMA AFIRMAÇÃO TÉCNICA GANHA ENDEREÇO

exemplo ilustrativo



Semântica

Objetos, atributos e relações do domínio: equipamentos, linhas, instrumentos, documentos e requisitos.

Cinética

Fontes, versões, referências e linhagem. Cada afirmação sabe de qual documento, versão e página veio.

Dinâmica

Ações auditáveis, regras e permissões. O que foi consultado, decidido e por quem fica registrado.

RELAÇÕES QUE IMPORTAM

EXIGE APLICA-SE A DEPENDE DE

É EXCEÇÃO DE SUBSTITUI

É EVIDENCIADO POR

PADRÕES ABERTOS

Interoperável com **ISO 15926**, **CFIHOS** e **DEXPI**, sem substituir os documentos normativos originais.

Sem ontologia, um modelo responde com base no que parece provável. Com ontologia, responde com base no que está registrado, e mostra onde. Na prática, a inconsistência passa a ser capturada na entrada do documento, e não na auditoria.

● 2.3 · A ONTOLOGIA DA COODEX

41 classes, 210 atributos e 490 relações de engenharia



Classes, em amarelo

Os tipos do domínio: equipamento, linha, válvula, instrumento, cabo, elemento civil, item de material, documento.

Arestas

Relações tipadas entre classes. É por elas que os agentes navegam para cruzar uma TAG entre documentos e sistemas.

Atributos, em azul

As propriedades de cada classe. A estrutura é o que se acumula de projeto em projeto; o documento do cliente não.

Visualização da ontologia de engenharia da COODEX, quadro capturado da renderização 3D exibida na palestra.



03

● SQUAD DE AGENTES

Um time que trabalha sobre a ontologia, não sobre arquivos soltos

Cada agente tem disciplina, ferramentas e permissões próprias. Trocar um agente ou um modelo não altera a estrutura do trabalho.

3.1 · O PRODUTO REFLETE O TIME

A Lei de Conway vale para agentes

Em 1968, o cientista da computação Melvin Conway observou que organizações que projetam sistemas acabam produzindo desenhos que copiam suas próprias estruturas de comunicação. Um time dividido em quatro grupos entrega um sistema em quatro partes.

Com agentes de IA, a regra fica ainda mais visível. Um único agente generalista produz respostas generalistas. Um squad organizado como uma engenharia de verdade, com disciplinas, interfaces e coordenação, produz um trabalho que respeita essas mesmas fronteiras.

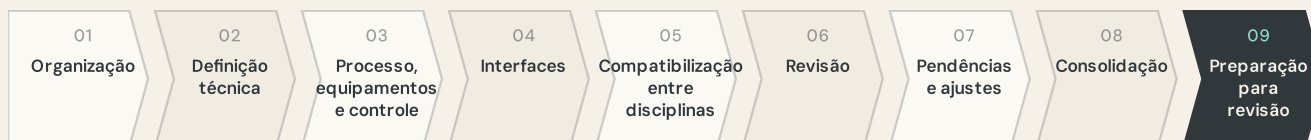
"Organizações que projetam sistemas são obrigadas a produzir desenhos que são cópias de suas estruturas de comunicação."

Melvin E. Conway, "How Do Committees Invent?", Datamation, 1968. Tradução livre.

COMO O SQUAD ESPELHA UMA ENGENHARIA DE EPC



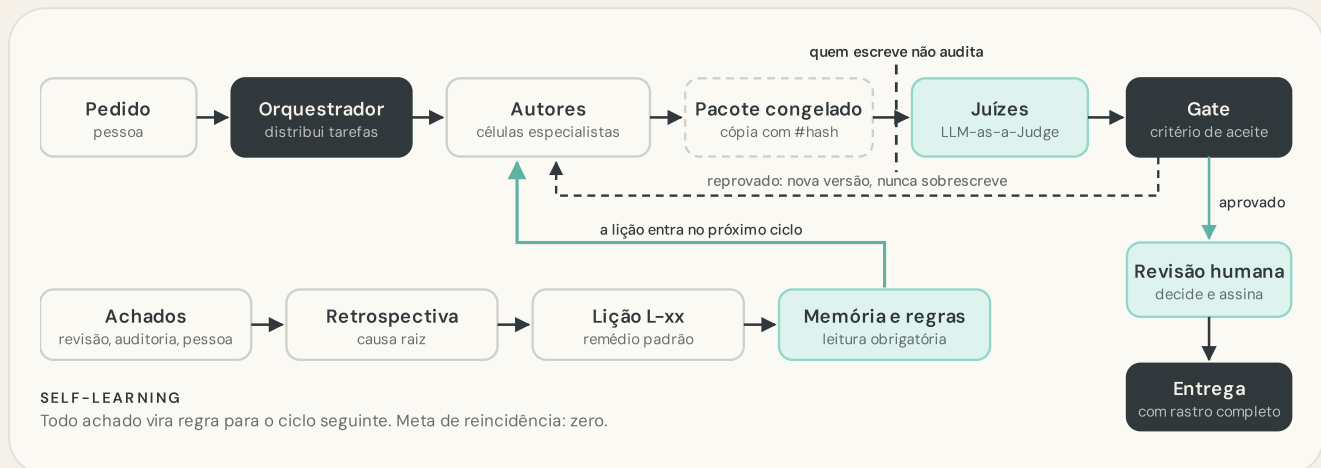
O FLUXO QUE O SQUAD PERCORRE, DO PEDIDO À REVISÃO HUMANA



3.2 · MECANISMOS DE CONFIANÇA

Quem escreve não audita

Um squad confiável não depende de um modelo que nunca erra. Depende de uma arquitetura em que todo erro é visto por outro agente, barrado por um gate e transformado em regra para o próximo ciclo.



LLM-as-a-Judge

Um modelo avalia a saída de outro contra critérios explícitos, com nota e justificativa. O juiz lê, não edita.

Gates de aprovação

Pontos de controle com critério de aceite. O trabalho só avança quando passa. Sem evidência, não há aprovação.

Separação de funções

Autoria, revisão e auditoria em agentes distintos, com permissões de escrita diferentes. Autor nunca audita o próprio trabalho.

Pacote congelado

O que vai para julgamento é uma cópia imutável, identificada por hash. Nenhum arquivo é sobrescrito: cada mudança gera nova versão.

Abstenção calibrada

Sem evidência suficiente, o sistema diz que não sabe, pede esclarecimento ou encaminha para revisão humana.

Human-in-the-loop

Agentes preparam e recomendam. Pessoas decidem, assinam e enviam. Nenhum agente age para fora sozinho.

Self-learning

Cada achado vira causa raiz, remédio e lição numerada, de leitura obrigatória no ciclo seguinte. Achado que volta é falha de aprendizado.

Memória por agente

Cada agente registra regras novas, armadilhas e erros ao fim de cada tarefa e relê antes da próxima.

Rastreabilidade

Cada execução registra modelo, agente, ferramentas, fontes, trechos recuperados, decisões e tempo.

RAG ancorado

Busca híbrida com reranking sobre fontes autorizadas. A resposta cita documento, versão e página.

Guardrails e hooks

Regras técnicas que impedem, por exemplo, gravar fora da pasta de cada agente ou enviar algo para fora.

Avaliação contínua

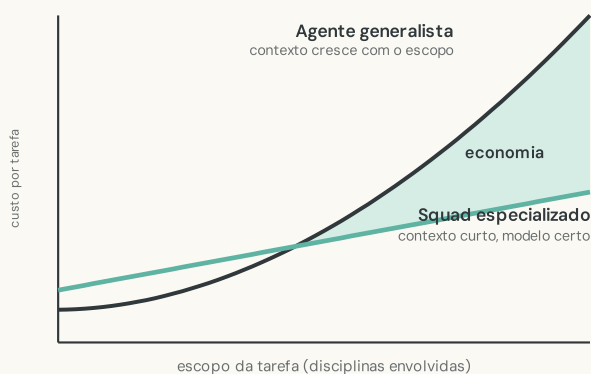
O EngBench-BR, benchmark brasileiro de IA para engenharia conduzido pela CODEX, compara o squad e a linha de base nas mesmas tarefas.

3.3 · TAMANHO DO TIME X CUSTO DE EXECUÇÃO

Mais especialização, menos custo por tarefa

Parece contraintuitivo: um squad com dezenas de agentes custaria mais do que um agente só. Não custa. Em IA, o custo não vem do número de agentes. Vem de quanto contexto cada chamada carrega e de qual modelo precisa ser usado para resolvê-la.

CUSTO POR TAREFA CONFORME O ESCOPO CRESCE



Esquema qualitativo, sem escala numérica.

Contexto curto

O especialista carrega só a sua disciplina e o trecho da ontologia que importa. O generalista carrega tudo, em toda chamada.

Modelo do tamanho certo

Tarefa estreita cabe em modelo menor, muitas vezes local. O modelo grande fica para coordenação e julgamento.

Cache e reuso

Instruções estáveis por agente são reaproveitadas de uma chamada para outra, e o custo marginal cai.

Paralelismo

Células trabalham ao mesmo tempo. O prazo depende da disciplina mais lenta, não da soma de todas.

Erro barato

Um erro barrado no gate custa uma nova versão. O mesmo erro descoberto na obra custa retrabalho.

$$\text{custo} \approx \sum \text{tokens} \times \text{preço} + \text{retrabalho}$$

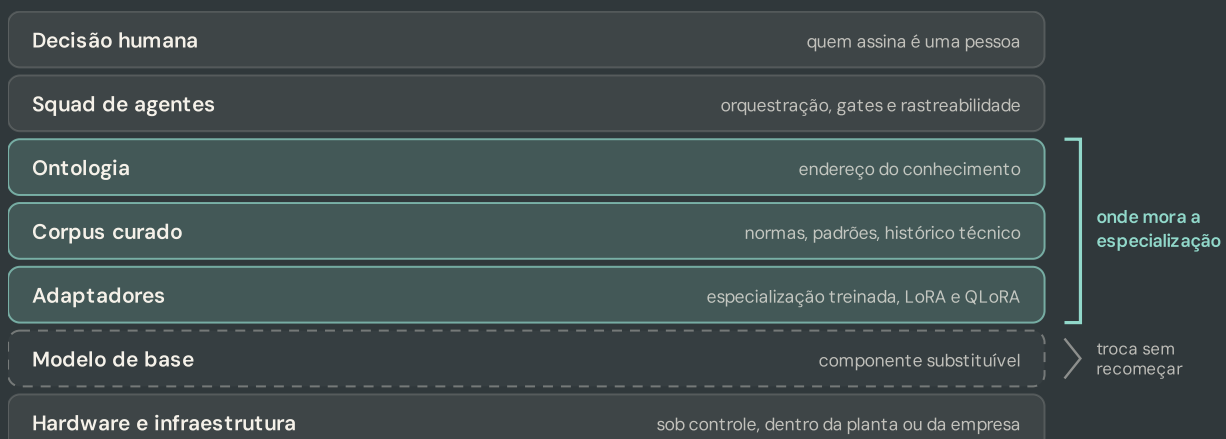
Especializar reduz os três termos ao mesmo tempo: menos tokens por chamada, preço menor por token e menos retrabalho. É a mesma lógica de uma equipe de engenharia bem montada, em que ninguém precisa saber tudo, mas cada um precisa saber muito bem a sua parte.

04

● MODELOS E INFRAESTRUTURA

A camada que decide de quem é o conhecimento

É a camada menos visível da arquitetura. E é a que determina, no fim do contrato, quem fica com o que foi aprendido.



4.1 · QUATRO RAZÕES

Por que a camada de modelos e hardware decide o jogo



Propriedade intelectual

O acervo técnico de uma empresa de engenharia é o seu ativo principal. Quando esse acervo alimenta um modelo de terceiros, parte do valor migra para quem opera o modelo. Manter modelo, adaptação e corpus sob controle próprio é o que preserva a propriedade intelectual do que foi aprendido.



Soberania nacional

Energia, óleo e gás e infraestrutura são setores de continuidade crítica. Sistemas que apoiam decisão técnica nesses setores precisam ser auditáveis e operáveis sob jurisdição nacional. Sem capacidade própria de treinar, avaliar e auditar, a régua técnica passa a ser definida fora do país.



Independência de fornecedores

O modelo de base deve ser um componente substituível, não o alicerce. A especialização real mora na ontologia, no corpus curado e nos adaptadores. Assim, trocar de fornecedor ou adotar um modelo melhor é uma decisão de engenharia, não um recomeço do zero.



Execução local na indústria

Documentos de projeto pertencem ao cliente e muitas vezes não podem transitar por infraestrutura de terceiros. Rodar dentro da planta ou em infraestrutura controlada deixa de ser preferência e vira requisito. Também resolve latência, conectividade instável em campo e cláusulas de confidencialidade.

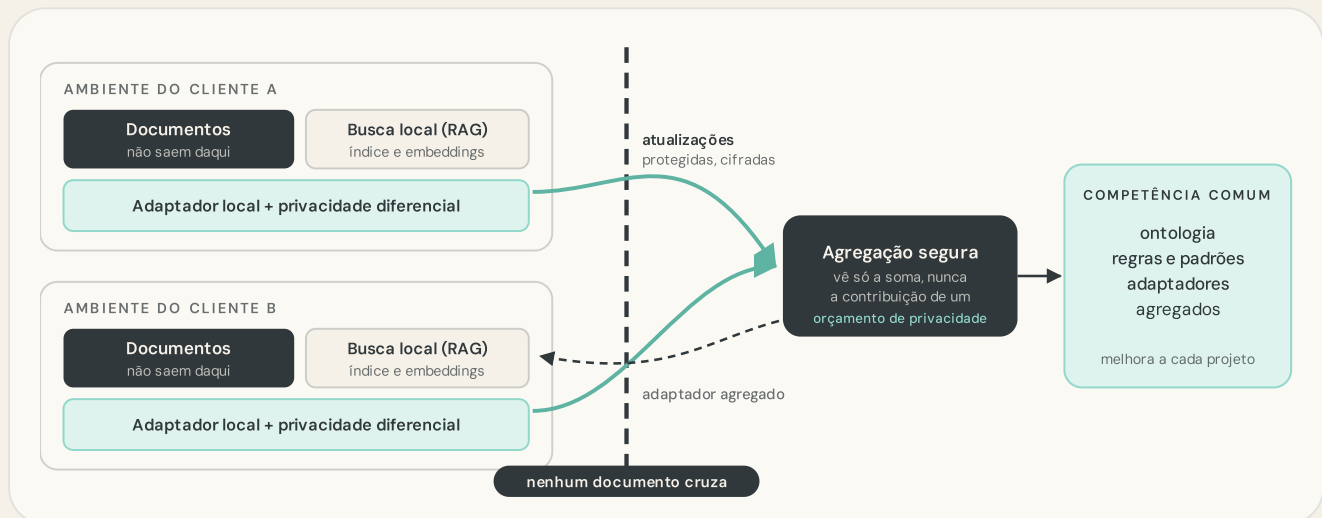
O QUE O CAPÍTULO 1 MUDA NESSA CONTA

FATO DE 2026	CONSEQUÊNCIA PARA QUEM OPERA ATIVOS
Modelos abertos a cerca de 4 meses da fronteira	Dá para ter qualidade de ponta sem enviar documento para fora. A escolha deixou de ser entre qualidade e controle.
Modelos de 27B a 30B rodando em uma GPU de 24 GB ou em um notebook	A execução dentro da planta ou do escritório de projeto é viável com hardware comum.
Memória e GPUs se valorizando	Capacidade própria virou ativo, e não só custo. Depender só de nuvem expõe o projeto à escassez e ao preço de terceiros.
Disputas de PI em série nos tribunais	Saber onde o conhecimento fica e por onde ele passou é, cada vez mais, uma exigência contratual.

4.2 · APRENDER SEM RETER

Competência que se acumula. Documento que não sai.

Um bom engenheiro fica melhor a cada projeto sem levar para casa os documentos do cliente. A mesma lógica pode ser construída em IA: separar o que é conhecimento comum de engenharia do que é dado confidencial de cada cliente.



Fica no cliente

Documentos, índice de busca, embeddings, histórico de consultas e o adaptador treinado localmente.

Pode ser compartilhado

Só atualizações protegidas por privacidade diferencial, cifradas e autenticadas, com orçamento de privacidade contabilizado por cliente.

O que se acumula

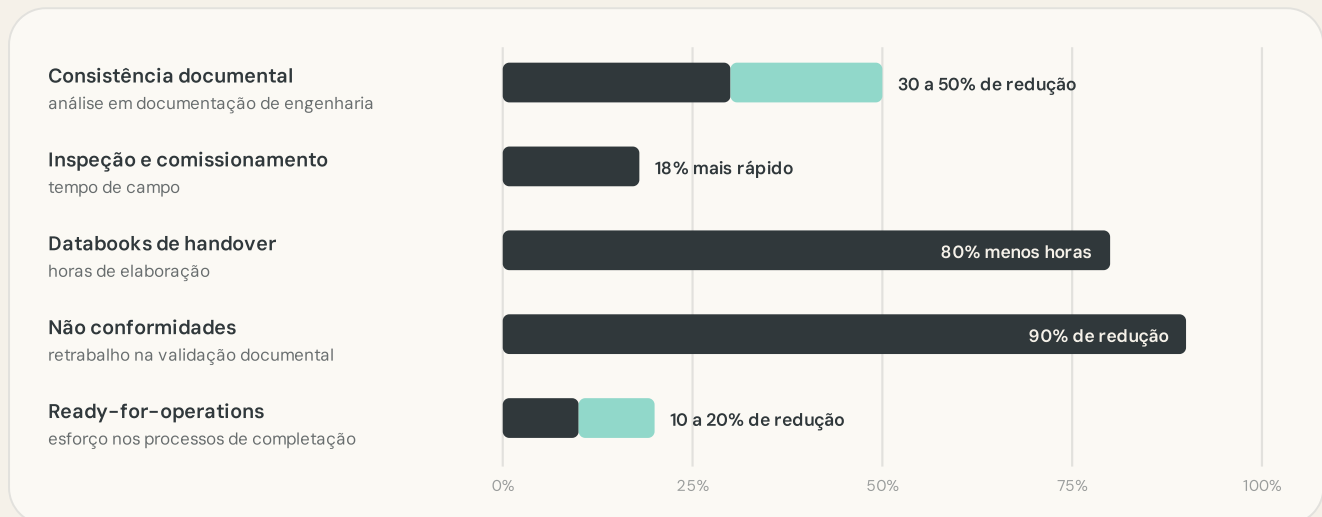
A competência comum: ontologia, regras, padrões e adaptadores agregados. É o que melhora de projeto em projeto.

Auditável por quem está de fora. A prova de que nada ficou retido não pode depender da palavra de quem construiu o sistema. Por isso, a arquitetura prevê ataques controlados ao modelo, pela mesma interface que o usuário acessa, para medir se alguma informação de cliente pode ser extraída.

● RESULTADOS EM CAMPO

O que muda quando a engenharia passa a ser auditável

Comparação entre processos baseline, manuais, e processos assistidos pela plataforma de IA. As faixas indicam a variação medida entre implantações.



Fonte: elaborado pelo autor (2026), a partir de implantações da plataforma.



A COODEX constrói a **MERIS**, plataforma AI-native de inteligência de engenharia, gestão documental e automação de fluxos para EPC, energia, óleo e gás e infraestrutura pesada. Em produção desde outubro de 2024.

200 mil

documentos
processados

80 mil

relatórios gerados

7 mil

análises de
engenharia

56+

agentes
especializados

Para rever

A palestra

Materiais da apresentação em roge.coodex.ai.

Para conhecer

A MERIS

Plataforma, casos e contato em coodex.ai.

● FONTES

Referências consultadas

Dados levantados até 22 de setembro de 2026. Números de desempenho divulgados pelas próprias empresas estão indicados como tal no texto. Links acessados na data de edição.

01 · CONTEXTO DA IA

- OpenAI, notas de lançamento de modelos. help.openai.com/en/articles/9624314
- OpenAI, Introducing GPT-5.5. openai.com/index/introducing-gpt-5-5
- OpenAI, GPT-6 Astra. openai.com/index/gpt-6-astra
- TechCrunch, OpenAI launches GPT-5.6 (09/07/2026) e Astra (03/09/2026)
- Anthropic, Claude Fable 5 e Mythos 5. anthropic.com/news/claude-fable-5-mythos-5
- Axios, Anthropic releases Opus 5 (24/07/2026)
- Google, Gemma 4. blog.google/.../gemma-4
- Google, Gemini 3.5 e 3.8 Flash. blog.google/.../gemini-models
- DeepSeek, V4 release. api-docs.deepseek.com/news/news260424
- TechCrunch, Meta debuts Muse Spark (08/04/2026)
- InfoQ, Meta Muse Glimmer (ago/2026)
- LLM Gateway, linha do tempo 2026. llmgateway.io/timeline/2026
- Epoch AI, Open vs. closed ECI gap (29/05/2026)
- The Decoder, Qwen3.6-27B (25/04/2026)
- Apple Newsroom, MacBook Pro M5 Pro e M5 Max (03/03/2026)
- TrendForce, preços de DRAM e NAND (05/01, 01/06 e 03/07/2026)
- Tom's Hardware, RAM Price Index 2026 (14/09/2026)
- SemiAnalysis, The Great GPU Shortage
- Tech Insider, NVIDIA B200 residual value (set/2026)
- Yahoo Finance, Nvidia stock valuation (08/07/2026)
- Relatórios do 2T26 de Amazon, Alphabet, Microsoft, Meta e Oracle
- TMT Finance, 2026 hyperscaler capex tops US\$700bn (18/08/2026)
- Gartner, AI spending forecast (16/09/2026)
- Epoch AI, OpenAI Stargate: where the US sites stand (17/04/2026)
- Google, Sundar Pichai no I/O 2026
- Scientific American, AI solved an 80-year-old Erdős problem (mai/2026)
- OpenAI, Ten advances in mathematics (01/08/2026)
- SCMP, primeira nota perfeita de IA na IMO (jul/2026)
- Quanta Magazine (03/08 e 08/09/2026); Nature (set/2026); NPR (22/09/2026)
- Jurist, Apple v. OpenAI (jul/2026); Axios (01/09/2026)
- Jurist, acordo Bartz v. Anthropic (jul/2026)
- Axios, NYT v. OpenAI (08/09/2026); NPR, Musk v. OpenAI (18/05/2026)
- Al Jazeera, xAI v. OpenAI (15/06/2026); TechCrunch (29/08/2026)
- IPKat (07/01/2026); Loeb & Loeb, Getty v. Stability (abr/2026)
- Imperva, Bad Bot Report 2025; Thales, Bad Bot Report 2026
- Tom's Hardware, Cloudflare: bots passaram humanos (04/06/2026)
- HUMAN Security, relatório de tráfego (09/04/2026)
- METR, time horizons. metr.org/time-horizons
- Axios, receita de Anthropic e OpenAI (17/08/2026)

02 · ONTOLOGIA

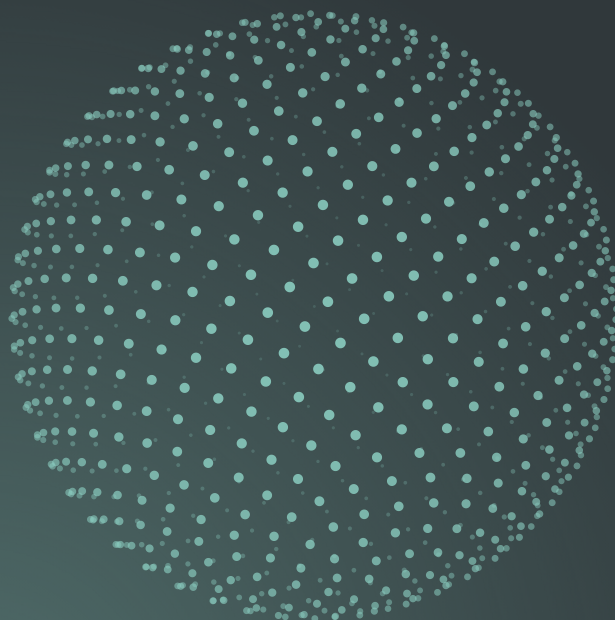
- Aristóteles, Categorias e Metafísica
- Rafael Sanzio, A Escola de Atenas (1509 a 1511), Museu Vaticanos, via Wikimedia Commons, domínio público
- Jacob Lorhard, Ogdos Scholastica (1606)
- Rudolph Goclenius, Lexicon philosophicum (1613)
- Christian Wolff, Philosophia prima sive ontologia (1730)
- T. R. Gruber, A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition (1993)
- ISO 15926; CFIHOS; DEXPI; W3C OWL

03 · SQUAD DE AGENTES

- M. E. Conway, How Do Committees Invent?, Datamation (1968)

APRESENTAÇÃO E RESULTADOS

- B. Flyvbjerg e D. Gardner, How Big Things Get Done (2023)
- B. Flyvbjerg, What You Should Know About Megaprojects and Why, Project Management Journal (2014)
- Indicadores de implantação: elaborado pelo autor (2026)



Como acumular a competência técnica de
engenharia ao longo de muitos projetos, sem
guardar os dados confidenciais de nenhum cliente.

`roge.coodex.ai`