

WORKSHOP

**Inteligência Artificial
Generativa com
Métodos:
Técnicas que
Multiplicam Resultados.**

JOÃO ROBERTO PERES

BIGMASTER
Educação e Tecnologia

Atualizado: junho/julho de 2026.



FÓRUM
BRASILEIRO
DE IoT



ESTÁGIO 1

PRIMEIRA PARTE - 06/08/2026

APRESENTAÇÃO



Há mais de 60 anos, muito antes do boom da IA Generativa, eu já inseria inteligência humana artificialmente em computadores através da programação.

O termo “IA” foi cunhado em 1956 por J. McCarthy, Dartmouth College.

João Roberto Peres



Engenheiro Eletricista, Bacharel em Administração, Mestre e Doutor em Ciência da Computação, com Ph.D em Segurança Corporativa entre outras. Possui mais de 50 anos de experiência profissional, tendo atuado por mais de duas décadas no mercado financeiro em cargos executivos de TI e como Professor, Coordenador e Consultor Especialista da FGV (mais de 20 anos), nas áreas de GRC, Governança de TI, Forense Digital, Direito Digital e Gestão de Riscos. É coautor do Guia de Referência em Segurança Corporativa da OAB-SP (2015) e editor internacional com registro ISBN/ISNI e autor de obras como "Manifesto Global de Direitos da IoT", "Investigação Forense Digital de IoT", "Privacidade e Conformidade para Desenvolvedores e Usuários de Soluções IoT", "Indústria 5.0", "Blue Book - cybersec4europe.eu" - "Inteligência Artificial IAG para Profissionais", "Autor Aumentado com IAG Assistente", "Invisible Core", "Imunidade Digital", "Agentes Invisíveis de IAG", "Ameaças Quânticas", "Alerta Biométrico", "Prompt Injection no Brasil", entre outras publicações técnicas recentes. Atualmente é proprietário de empresas de tecnologia e segurança cibernética, cofundador de plataformas digitais TSIOT e sobre Tecnologias Avançadas e Inovadoras (TAI), e colaborador das Comissões de Privacidade e Inteligência Artificial e Consultor da Comissão de Direito Digital da OAB-SP. Conselheiro do Fórum Brasileiro de IoT e da Digital Law Academy.

A Adoção em Massa da Inteligência Artificial Generativa

1.1 – O mercado em movimento. A Inteligência Artificial Generativa consolidou-se no mercado global, deixando de ser um experimento laboratorial para se tornar uma realidade corporativa em rápida expansão. Essa evolução é impulsionada pela adoção em massa de **agentes virtuais** nas empresas e por uma transição tecnológica crucial: a migração dos sistemas centralizados para a computação de borda (edge computing). Com isso, a tecnologia deixa de ser apenas uma ferramenta de escritório e passa a se integrar diretamente ao ecossistema conectado da internet das coisas industriais.



A Democratização do Acesso (O Fenômeno Inicial)



Insight: A prova de que interagir com sistemas de altíssima complexidade não exige conhecimento técnico.

Consequência: A faísca na interface do usuário que desencadeou a atual corrida dos laboratórios.

ChatGPT: 100 milhões de usuários em 2 meses



Maio de 2026 – 1 bilhão



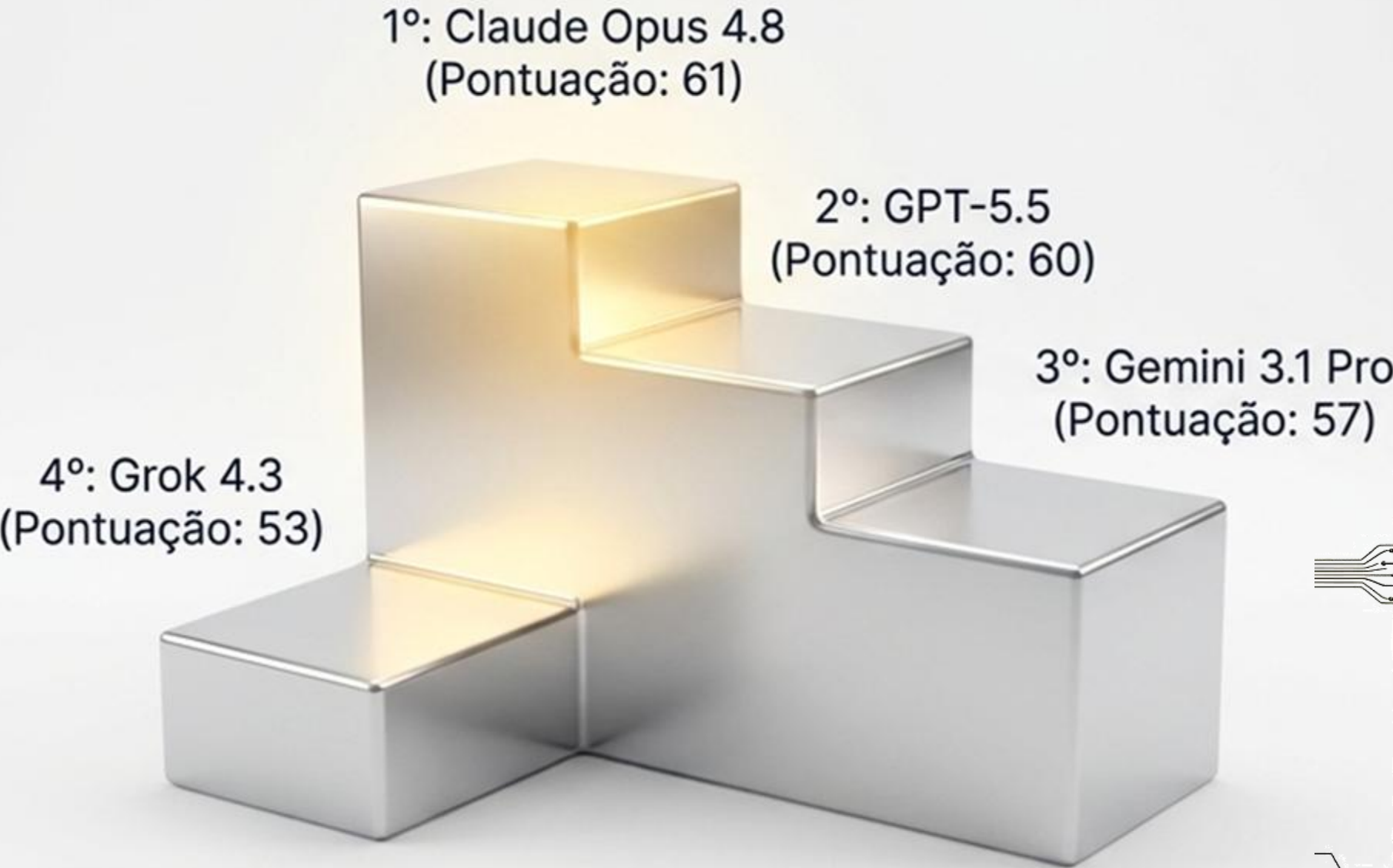
TikTok: 9 meses para o mesmo marco

1.2 O fenômeno ChatGPT e o que ele revelou

O lançamento do ChatGPT pela OpenAI 11/2022 alcançou a marca histórica de 100 milhões de usuários em apenas dois meses devido à sua facilidade de uso sem precedentes. Essa democratização do acesso à inteligência artificial permitiu que qualquer pessoa interagisse com o sistema, desencadeando a corrida tecnológica global que moldou o ecossistema atual.

Fato significativo para o apresentador: Apresentei o Primeiro Workshop Interativo on-line sobre o ChatGPT em 15/02/2023 – com o título “Preludio – Principais Atualizações Tecnológicas e Hypes de Mercado – IA ChatGPT – Metaversos – ESG – : Como as tecnologias e práticas se integram.”

O Ecossistema de Fronteira (Ranking Junho/2026)



1.3 O ecossistema atual: quatro fronteiras
O mercado atual de IA é liderado por quatro modelos de fronteira lançados no primeiro semestre de 2026: Claude Opus 4.8 (líder geral em inteligência e destaque em codificação junto ao GPT-5.5, que também domina a escrita criativa), Gemini 3.1 Pro (referência em raciocínio analítico e ciência) e Grok 4.3 (focado em tarefas agênticas e melhor custo-benefício).



Aviso: Cenário hiperdinâmico. Mais de 12 lançamentos em uma única semana em março/2026. O critério não é "qual é o melhor", mas "qual serve ao seu caso".

A Matriz de Especialidades: A Ferramenta Certa



Escrita Criativa & Brainstorming
Liderança: GPT-5.5



Codificação & Engenharia de Software
Liderança: Claude Opus 4.8 e GPT-5.5



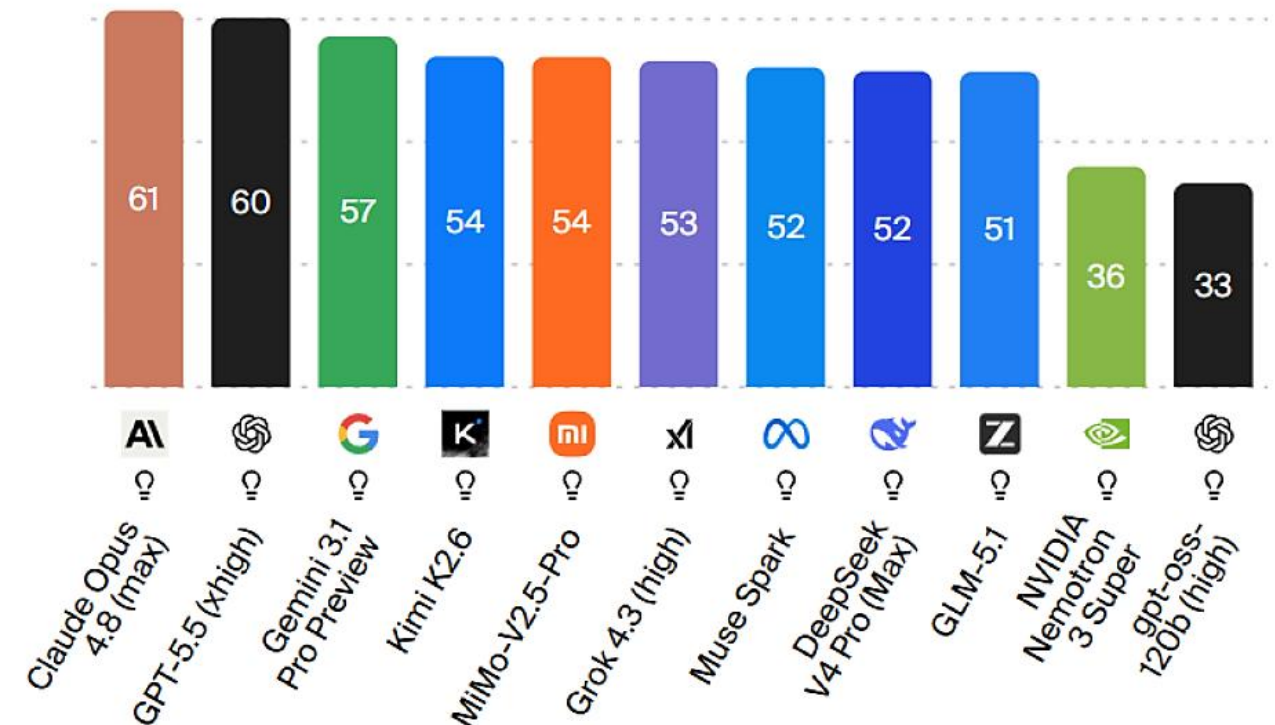
Raciocínio Analítico & Precisão Científica
Liderança: Gemini 3.1 Pro



Tarefas Agênticas & Custo-Benefício
Liderança: Grok 4.3

Inteligência

Índice de Inteligência Artificial Analítica - Quanto maior, melhor



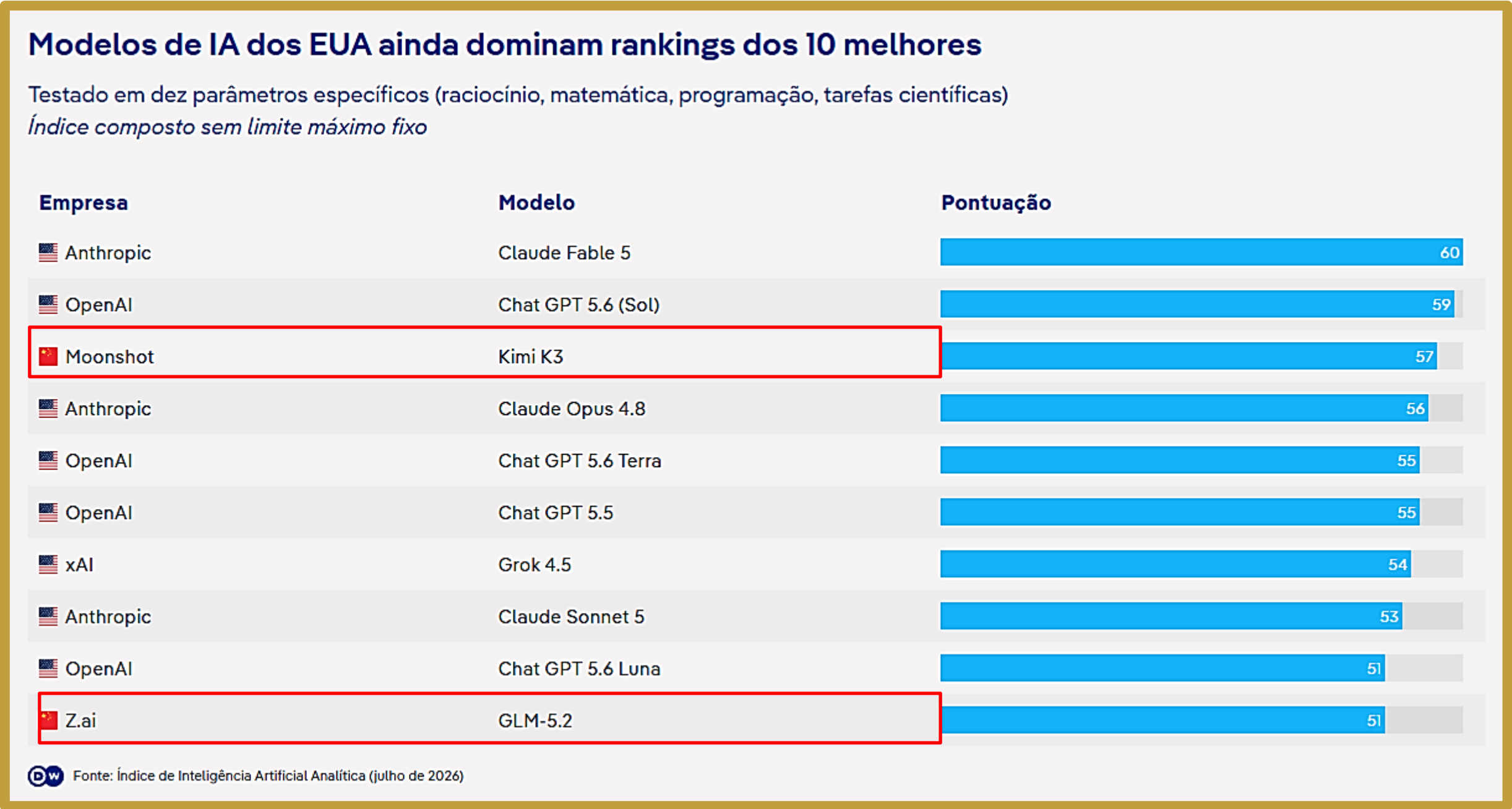
<https://artificialanalysis.ai/> - 02/06/2026

Embora os modelos de IA de fronteira (*na vanguarda do conhecimento*) sejam de uso geral e estejam prontos para responder sobre qualquer tema, **eles não são idênticos**; cada fornecedor desenvolve sua tecnologia com refinamentos distintos, resultando em especializações claras para determinados objetivos. Na prática, isso significa que, enquanto um modelo pode apresentar desempenho superior e mais refinado para codificação complexa e escrita criativa, outro se destacará em raciocínio analítico e precisão científica, fazendo com que a escolha do modelo ideal dependa diretamente do foco e da necessidade de solução exigida pela tarefa.

A Grande Concorrência



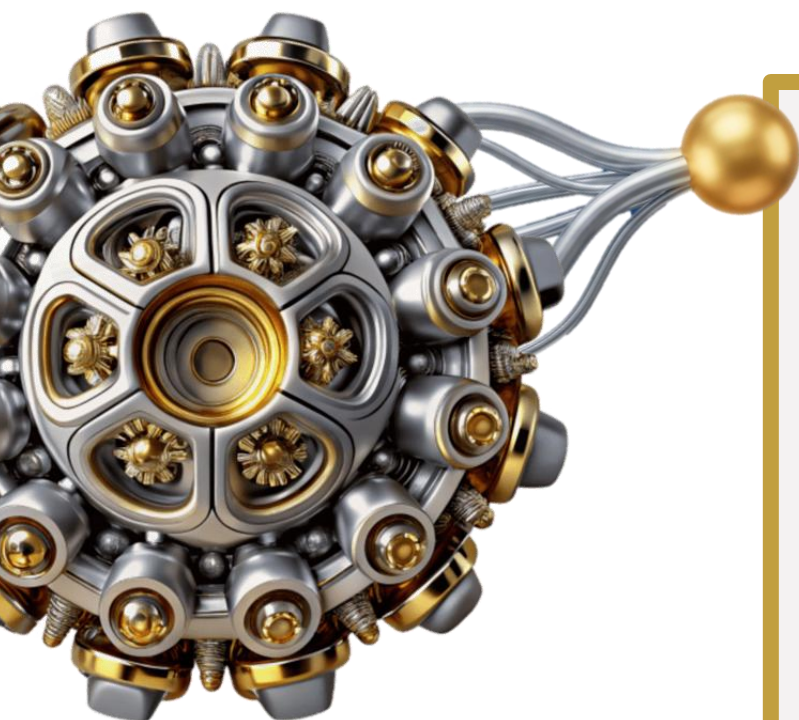
Pontuação
Máxima Claude
Fable 5 = 60
(DW)



<https://www.dw.com/pt-br/ia-chinesa-avan%C3%A7a-e-desafia-dom%C3%ADnio-dos-eua/a-78090795>

Avaliação Internacional rigorosa comparando modelos abertos.

A Corrida Acelerada na China



Pontuação
Máxima Kimi
K3 = 81 em
comparação
as demais
Chinesas.

CN Os 10 melhores modelos de IA chineses

Testados em centenas de tarefas: (raciocínio, programação, matemática, ciências, trabalho autônomo)
Pontuação máxima: 100

Empresa	Modelo	Pontuação
Moonshot	Kimi K3	81
Alibaba	Qwen 3.7 Max	72,8
Xiaomi	MiMo-V2.5-Pro	70,2
MiniMax	MiniMax M3	69,8
Xiaomi	MiMo-V2-Pro	67,8
Z.ai	GLM5.1	67,7
Alibaba	Qwen 3.7 Plus	67,2
Z.ai	GLM-5-Turbo	66,9
Z.ai	GLM-5	66,1
Alibaba	Qwen3.6 Plus	65,2

DW Fonte: BenchLM (Julho 2026)

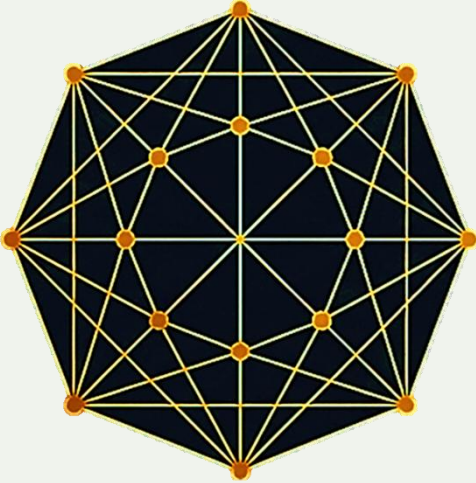
Avaliação regional feita pela China nos seus modelos de código aberto.

Veja as características básicas das IAGs que indicamos:

Nome	Empresa Responsável	Principais Indicações / Melhor Aplicação
ChatGPT	OpenAI	Conversação natural, escrita criativa, assistência geral, suporte educacional, brainstorming e resolução de problemas complexos, imagens básicas...
Claude	Anthropic	Análise de documentos extensos, escrita técnica, resumos detalhados, conversação contextual prolongada, geração de textos longos e tarefas que exigem raciocínio ético...
Copilot	Microsoft	Programação e desenvolvimento de software, integração com ferramentas Microsoft 365 (Word, Excel, PowerPoint), revisão de código e automação de tarefas, imagens básicas e avançadas...
Gemini	Google	Pesquisa avançada, análise multimodal (texto, imagem, vídeo básico), integração com ecossistema Google, geração de conteúdo multimídia...
Lumo	Proton AG Suíça	Privacidade e Segurança é o foco: Ferramenta especializada em criação de conteúdo textual assistido por IA – é muito segura.
Manus	Butterfly Effect	Assistente geral de IA, desenvolvido em Singapura, focado em tarefas administrativas e organizacionais – IA Agêntica básica paga.
Meta AI	Meta (Facebook)	Integração com redes sociais, processamento multimodal, criação de conteúdo para marketing digital e análise de dados sociais...
Mistral	Mistral AI Le Chat França	Processamento de linguagem natural com foco em eficiência, adequado para aplicações empresariais e desenvolvimento de chatbots – baseada em GPT.
NotebookLM	Google	Usa dados fornecidos por Usuários - padrão RAG – Faz análise e síntese de documentos, criação de artefatos a partir de textos, pesquisa acadêmica e organização de conhecimento...
Perplexity	Perplexity AI	Motor de busca inteligente, pesquisa em tempo real com citações, verificação de fatos e análise de fontes confiáveis – usa navegador interno Comet.
Qwen	Alibaba China	Processamento multilíngue avançado, especialmente otimizado para idiomas asiáticos, escreve português fluente, faz análise de dados e desenvolve aplicações empresariais (algumas funções pagas)...

Aqui estão os links (URLs) para a inscrição gratuita nas IAs Generativas listados em ordem alfabética:

ChatGPT: <https://chat.openai.com/>
Claude: <https://claude.ai/>
Copilot: <https://copilot.microsoft.com/>
Gemini: <https://gemini.google.com/>
Lumo AI: <https://lumo.proton.me>
Manus: <https://manus.im/login>
Meta AI: <https://ai.meta.com/>
Mistral: <https://console.mistral.ai/>
NotebookLM: <https://notebooklm.google/>
Perplexity: <https://www.perplexity.ai/>
Qwen: <https://chat.qwen.ai/>
Poe: <https://poe.com/login> (múltiplas IAGs) agregadas



Prepare o seu arsenal de ferramentas de IAg Gratuitas.

1.5 Recomendação: Para se cadastrar tenha um endereço de e-mail válido, se possível não corporativo. Evite colocar seu nome completo – use só seu primeiro e último nome. Alguns endereços de links podem mudar com o tempo. Tenha seu navegador atualizado.



Como Operam os Modelos de IAG nos Smartphones

Os mesmos modelos LLMs para rodarem em Smartphones **são modificados radicalmente** para características de SLMs – (Small Language Models) para caber na memória restrita do dispositivo (RAM) e consumir menos bateria. Portanto, com capacidades distintas:

A diferença é muito significativa hoje.

Modelos em Nuvem (LLMs):

- Executados em Data Centers
- Centenas de bilhões de parâmetros
- Grande capacidade de raciocínio e contexto
- Alto consumo computacional
- Dependem de conexão com a Internet
- Processamento em GPUs de alto desempenho

Processo de Adaptação:

- O modelo é otimizado para dispositivos móveis
- Compressão de parâmetros
- Quantização (8, 4 ou menos bits)
- Redução do tamanho do modelo
- Otimização para CPU, GPU e NPU do smartphone
- Menor consumo de memória e energia

SLM no Smartphone

Small Language Model (SLM):

- Executado localmente no aparelho
- Baixo consumo de RAM
- Menor consumo de bateria
- Resposta rápida e funcionamento off-line
- Maior privacidade dos dados

Limitações

- Menor capacidade de raciocínio
- Janela de contexto reduzida
- Menor conhecimento geral
- Recursos significativamente mais limitados que um LLM em nuvem



A ilusão Freemium e a Política de uso de Dados

90% das ferramentas gratuitas usam dados de consulta ou interações dos usuários para treinar seus modelos. Isso significa que um arquivo, uma imagem ou um texto anexado ao “prompt” e o próprio “prompt”, servirão para recapacitar o modelo. Na prática seus dados podem ser expostos ao mundo.

Alguns modelos de IAG mesmo gratuitos permitem configurar o uso para treinamentos (ex.: Claude free) e outros gratuitos como “Lumo AI” afirmam não usar os dados e prompts das interações.

TODO CUIDADO É POUCO.

Regras de Ouro : Não confie cegamente na primeira resposta de qualquer IAG sem sua confirmação de fontes.



- **A Regra Fundamental:** Quanto mais sensível for a informação, mais rigorosa e criteriosa deve ser a escolha da plataforma.
- **Ação Obrigatória:** Para dados confidenciais, utilizar exclusivamente plataformas com garantia contratual de não-uso de dados (Ex: Lumo 1.3, Enterprise/Teams).
- **Atenção a Agentes:** Extremo rigor na adoção de arquiteturas agênticas. Autonomia de execução requer avaliação crítica de segurança (Ex: bloqueio do Manus na China).

A Regra Fundamental: Avalie criticamente a sensibilidade da informação antes do envio, pois ferramentas gratuitas geralmente utilizam seus dados para treinamento de modelos. Se o conteúdo envolver dados confidenciais ou segredos de negócio, a ação obrigatória é migrar para plataformas corporativas (Enterprise) que garantam a privacidade contratual, ou criar infraestrutura própria, para rodar modelos locais sem acesso externo.

Atenção Máxima a Agentes: Exija extremo rigor e vigilância ao adotar arquiteturas agênticas e automações gratuitas. Como esses sistemas possuem autonomia para executar tarefas, eles demandam uma avaliação crítica de segurança para evitar vazamentos e acessos indevidos ao seu ambiente digital.



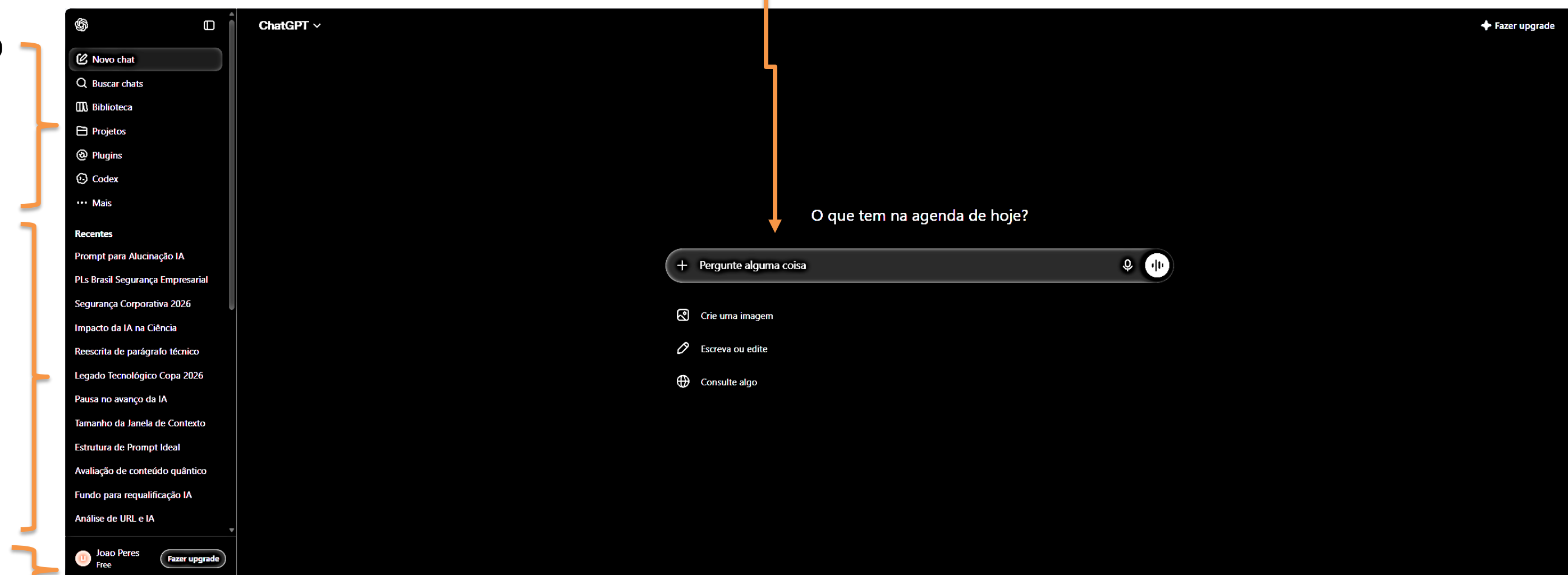
Interface de usuário padrão

Na prática a grande maioria das IAs generativas possuem UI, muito parecidas sempre no idioma em que o usuário fez o cadastro na plataforma. O **Box** central na tela é onde se coloca o “prompt” de comando. As entradas de dados podem ser por texto, arquivos anexados, imagens, arquivos de áudio e vídeo e até entrada e saída por voz.

Área de seleção

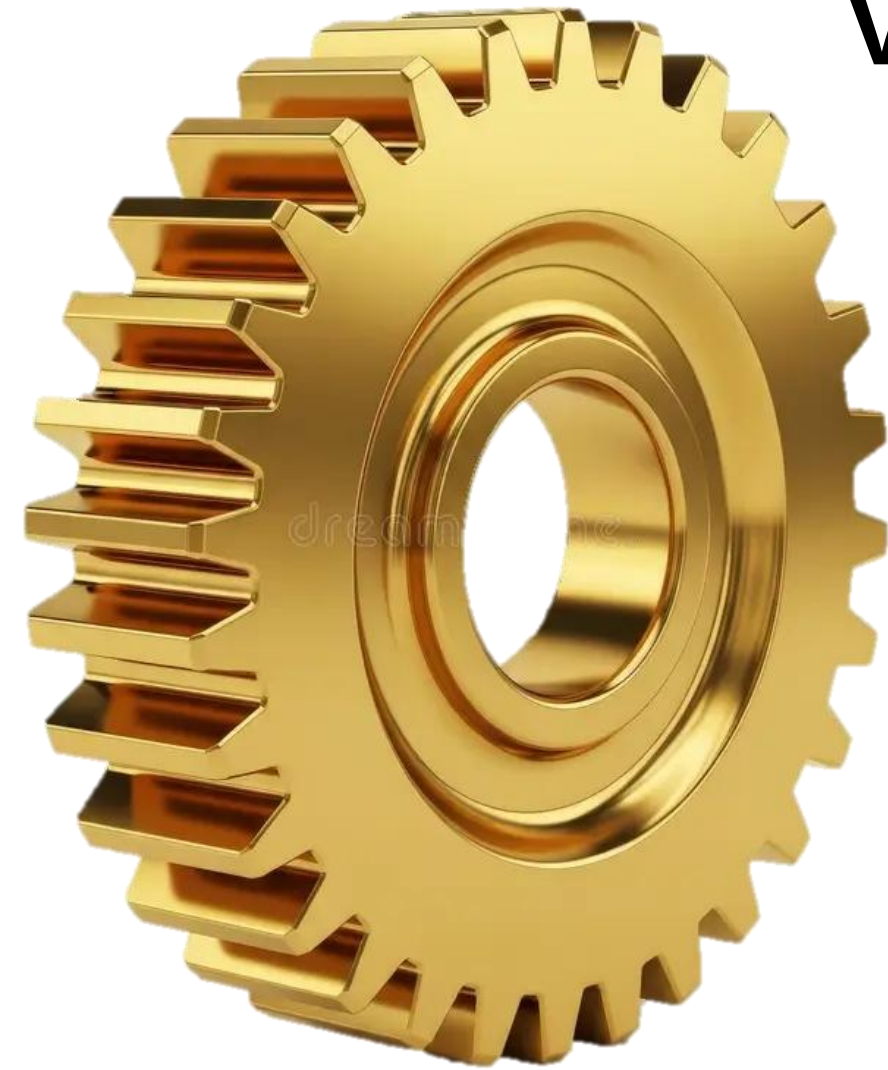
Área de conversas recentes

Área de identificação e configuração



Obs.: Apesar da facilidade e aplicações específicas – não recomendo usar a opção de **prompt por voz** – lembre que a VOZ é um elemento de característica biométrica única de cada pessoa – existe o risco de **deepfake** usando a sua voz caso haja vazamento na plataforma.

Vamos criar um Prompt simples



+ Pergunte alguma coisa



Prompt 1 - “Faça um resumo do arquivo em poucas linhas.”

Problemas desse prompt:

- Não informa o objetivo do resumo.
- Não define o público-alvo.
- Não especifica o nível de detalhamento.
- Não informa o que deve ser preservado ou omitido.
- Não define estrutura de saída.
- Não prevê limitações caso o documento esteja incompleto ou ilegível.
- ...

- Clicar em + na barra de prompt > para anexar arquivos, imagens, e outras funções.
- O + permite fazer uploads ou selecionar alternativas.

Vamos criar um Prompt

+ Pergunte alguma coisa



Prompt 2 - “Objetivo: Produzir um resumo executivo fiel ao conteúdo do PDF anexado.

Instruções

1. Leia integralmente o documento antes de elaborar o resumo.
2. Produza um resumo entre 500 e 700 palavras.
3. Preserve exclusivamente informações presentes no documento. Não faça inferências, não complete lacunas e não utilize conhecimento externo.
4. Estructure a resposta nas seguintes seções: • Objetivo do documento • Principais temas abordados • Principais conclusões • Recomendações apresentadas pelo autor • Dados quantitativos relevantes • Limitações ou ressalvas mencionadas.
5. Caso alguma informação não exista no documento, informe: "Não identificado no documento."
6. Sempre que citar números, datas, normas, leis ou estatísticas, reproduza-os exatamente como aparecem no PDF.
7. Ao final, apresente uma lista de até dez pontos-chave ("Key Takeaways").
8. Se o PDF estiver ilegível, incompleto ou protegido, informe claramente essa **limitação antes de responder**.

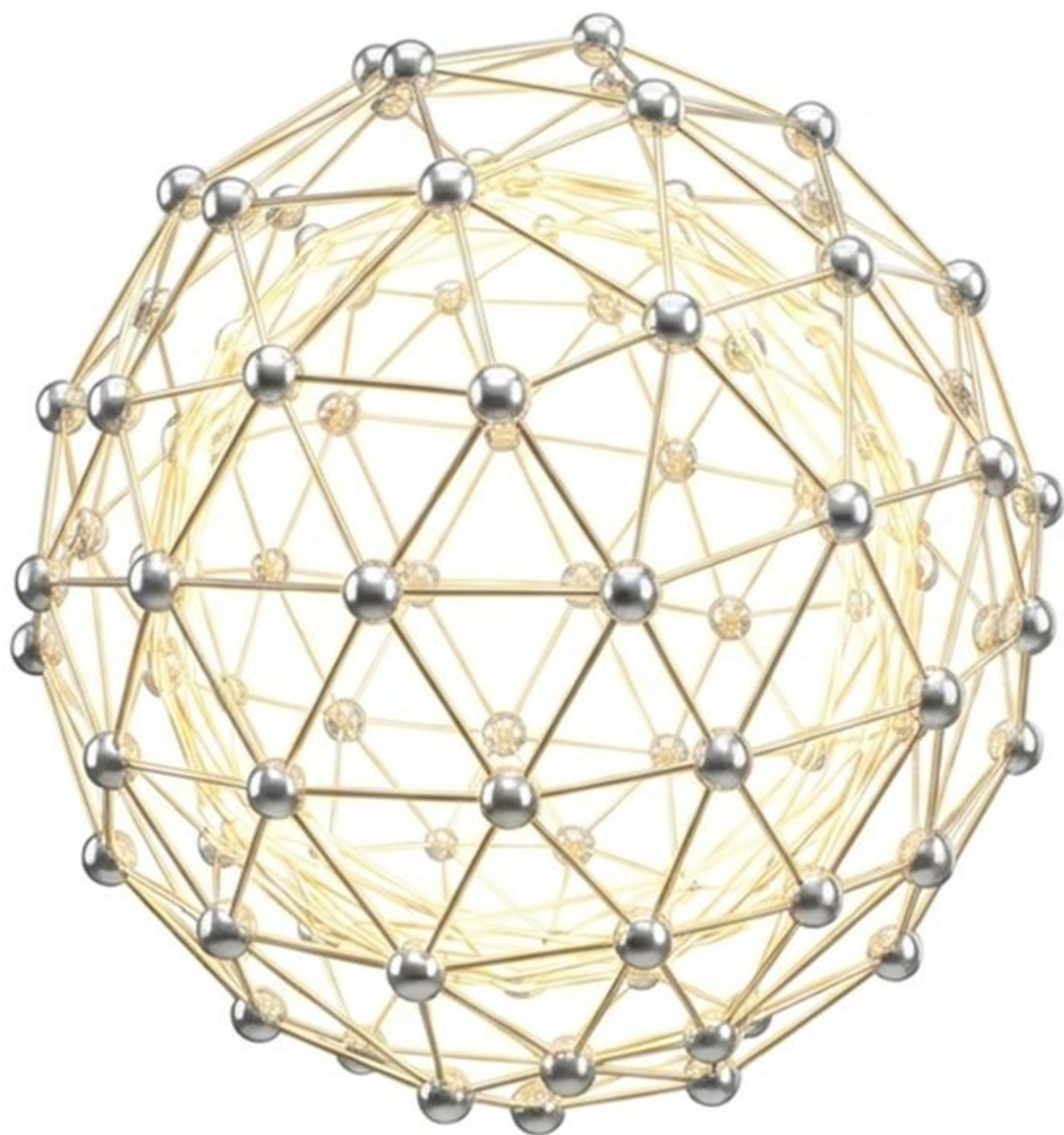
Formato da resposta: # Resumo Executivo # # Objetivo # # Principais Temas # # Conclusões #
Recomendações # # Dados Relevantes # # Limitações # # Principais conclusões



Comparação Prompt 1 e Prompt 2

Critério	Prompt de Baixa Qualidade – Prompt 1	Prompt de Alta Qualidade – Prompt 2
Objetivo	Genérico	Claro e Específico
Contexto	Não Informado	Define a Finalidade
Público Alvo	Não Definido	Público Executivo
Escopo	Indefinido	Delimitado
Estrutura	Livre	Padronizada
Profundidade	Variável	Definida
Critério de Seleção	Não Especificados	Pontos Prioritários
Uso de Conhecimento Externo	Sem orientação	Restrito ao Conhecimento do .pdf anexado
Tratamento de Lacunas	Não Previsto	Informa Quando Faltar Conteúdo
Tamanho da Resposta	Indefinido	Limitado 500 mínimo a700 palavras no máximo
Consistência	Variável	Elevada
Confiança	Baixa	Alta
Resultado Esperado	Imprevisível	Consistente e Útil





Arquitetura: Como uma IAG Realmente Funciona

Desconstruindo as camadas do sistema
para profissionais de IoT e AIoT.

ESTÁGIO 1

SEGUNDA PARTE - 06/08/2026

SISTEMA

Sistema Black Box = Caixa Opaca = Código da arquitetura



Tokenizers da
OpenAI

Tokens

Tokens

Tokens

Tokens

Prompt: Descreva um dia e resuma esse dia em uma frase.

Prompt traduzido: Describe a day and summarize that day in one sentence.

"Describe" "a" "day" "and" "sum marize" "that" "day" "in" "one" "sentence" "."

3251 264 **1452** 323 9087 8543 430 **1452** 304 832 11914 13 ID Token

Camada 1 — O Motor Computacional

Redes Neurais:

Bilhões de parâmetros matemáticos ajustados durante o treinamento.

O modelo não “sabe”; ele calcula a sequência mais provável.

O Token

A unidade mínima mecânica. Um token não é uma palavra.

Representa ~3/4 de uma palavra em inglês ou uma sílaba em português.

2. ESTRUTURA PADRÃO

Camada 1 Motor - O modelo de IA: GPT-OSS, Magistral, Gema-3, Mistral, Qwen3, Nemotron-3-ultra...

Camada 2 Base de Conhecimento: usada no treinamento – Todos os tipos de dados fornecidos e aceitos pelo modelo.

Camada 3 Red Line RLHF – estrutura paramétrica do desenvolvedor – o que o modelo pode e não pode fazer ou atender – questões éticas e funcionais.

Camada 4 interface conversacional – o box de inserção de prompts e todos os demais parâmetro e funções UX.

Camada 5 System Prompt – controle de inicialização dos usuários – por ser mais avançado é praticamente desconhecido.



OS MODELOS DE IA



MODELO CAMADA 1 de Inteligência Artificial é uma estrutura de software codificada, composta por rotinas algorítmicas de Aprendizado de Máquina e Redes Neurais, que pode ser treinada para reconhecer padrões, gerar previsões e apoiar decisões de forma autônoma. Em termos de negócio, ele transforma dados em inteligência que acelera resultados e reduz incertezas.

Existem dezena de modelo de código aberto para serem baixados de forma livre: – alguns modelos podem ser comprados – tudo para uso em computadores locais (notebooks/desktop/workstation) com ou sem conexão com a Internet.

Algumas URLs para bibliotecas de Modelos (motores):

<https://lmstudio.ai/models>

<https://ollama.com/search>

<https://www.llama.com/models/llama-3/>

Quando se baixa um modelo para rodar localmente: Se baixa os Pesos (Weights): Arquivos nos formatos **.gguf**, **.safetensors** ou **.bin**.

O que há dentro: Matrizes matemáticas pesadas contendo bilhões de parâmetros (ex: Llama 3 8B, Mistral 7B, Command R+ 104B).

Capacidade: O modelo consegue responder sobre fatos históricos, programar, resumir textos e raciocinar sobre dados gerais sem precisar da internet, pois esse conhecimento está sintetizado na estrutura matemática dele.

código kimi-k2.7

O Kimi K2.7 Code é um modelo agentivo da Moonshot AI focado em programação, baseado no Kimi K2.6, com melhorias substanciais em tarefas de programação de longo prazo do mundo real e uma redução de aproximadamente 30% no uso de tokens de pensamento.

[visão](#) [ferramentas](#) [pensamento](#) [nuvem](#)

↓ 2.209 Puxa 1 Marcação ⌚ Atualizado 12 horas atrás

minimax-m3

MiniMax M3: Fronteira de Codificação e Agentes. Janela de contexto de 1 milhão. Multimodalidade Nativa.

[visão](#) [ferramentas](#) [pensamento](#) [nuvem](#)

↓ 46,6 mil Puxa 1 Marcação ⌚ Atualizado 1 semana atrás

lfm2.5

LFM2.5-8B-A1B, um modelo de ponta projetado para chamadas de ferramentas rápidas e confiáveis em hardware de consumo.

[ferramentas](#) [pensamento](#) [8b](#)

↓ 21 mil Puxa 5 Etiquetas ⌚ Atualizado 1 semana atrás

nemotron-3-ultra

O NVIDIA Nemotron 3 Ultra foi projetado para raciocínio de alto desempenho e fluxos de trabalho de agentes de longa duração.

[ferramentas](#) [pensamento](#) [nuvem](#)

↓ 10,9 mil Puxa 1 Marcação ⌚ Atualizado 1 semana atrás

Camada 2 — A Base de Conhecimento



O limite de textos, códigos e normas sobre os quais o modelo foi inicialmente treinado.

A Ponte: Acesso em tempo real à web atua como uma solução provisória para essa lacuna.

Ponto de Corte (Cutoff): A fronteira temporal. Após esta data exata, o modelo sofre de amnésia crônica.



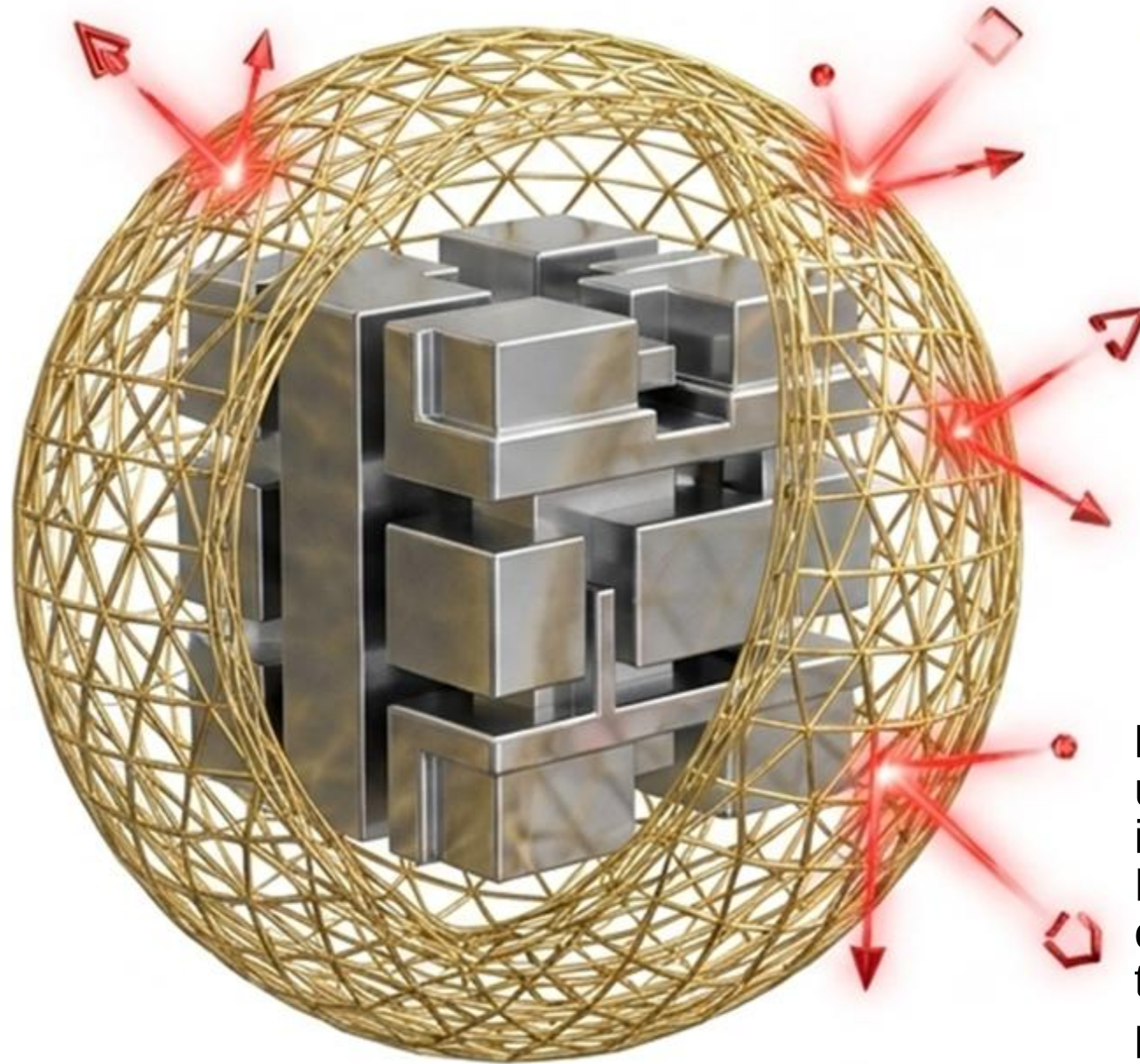
Após o treinamento com as bases de dados, são geradas as **Matrizes de Pesos**: As tabelas numéricas com o conhecimento. **Os Vieses**: Os ajustes matemáticos e os **Metadados**: Informações sobre o modelo (autor, versão, etc.).

A Linha de Frente da Informação: Modelos de Fronteira vs. Modelos Locais

A Ilusão do Conhecimento Nativo: Grandes plataformas comerciais baseadas em modelos de fronteira chegam ao usuário prontas e pré-treinadas com volumes massivos de dados globais. No entanto, mesmo essa infraestrutura gigante esbarra no Ponto de Corte (Cutoff): a fronteira temporal exata após a qual o modelo sofre de "amnésia crônica" e desconhece qualquer nova atualização regulatória, técnica ou de mercado (como as revisões normativas ocorridas no início de 2026). Embora o acesso em tempo real à web atue como uma ponte provisória para esses sistemas comerciais, essa abordagem gera riscos de conformidade em ambientes corporativos rígidos.

O Desafio dos Modelos Locais (Camada 2 — A Base de Conhecimento): Quando optamos por baixar e executar modelos de IA em infraestruturas e máquinas locais (garantindo privacidade absoluta), o cenário muda radicalmente. Esses modelos locais possuem um limite severo de textos, códigos e normas sobre os quais foram inicialmente treinados. Para que eles operem com utilidade prática dentro da empresa, torna-se obrigatório estabelecer e injetar bases de dados de conhecimento proprietárias. Sem essa alimentação estruturada de dados internos e atualizados, um modelo local torna-se obsoleto imediatamente para o ecossistema de negócios.

Camada 3 — Red Lines (Proteções Constitucionais)



RLHF: Aprendizado por Reforço com Feedback Humano. As restrições enraizadas no sistema durante o treinamento.

Fronteiras Intransponíveis: Não são configuráveis pelo usuário. Recusam código malicioso ou quebra sistemática de privacidade.

Aviso de Segurança IoT: Tentativas de 'Jailbreak' (subversão via engenharia de prompt) são rastreadas, penalizando a confiabilidade de toda a sessão de trabalho.

Restrições Enraizadas no Sistema: Antes mesmo de qualquer customização do usuário, os modelos de IA possuem uma infraestrutura de segurança nativa implementada pelos seus desenvolvedores por meio de RLHF (Aprendizado por Reforço com Feedback Humano). Essas restrições agem como proteções constitucionais enraizadas diretamente nos pesos do sistema durante sua fase de treinamento fundamental.

Fronteiras Intransponíveis e Rastreabilidade: Essas barreiras de segurança não são configuráveis ou modificáveis pelo usuário final, atuando de forma automatizada na recusa de geração de códigos maliciosos ou na quebra sistemática de privacidade. No ecossistema conectado, tentativas de subversão dessas regras por engenharia de prompt (os chamados Jailbreaks) são ativamente rastreadas pelos provedores, penalizando e comprometendo a confiabilidade e o comportamento de toda a sessão de trabalho.

Camada 3 – Parâmetros de controle - gatilhos mentais de persuasão (Lisonja / Adulação) – toda IA é muito educada



Porque ocorre:

► Reinforcement Learning from Human Feedback (RLHF)

As IAs são treinadas com base no feedback humano. Avaliadores tendem a dar notas mais altas para respostas educadas, empáticas, amigáveis e concordantes.

► Minimização de Conflitos e Retenção do Usuário

Reduzir o atrito melhora a experiência de uso (UX). Um tom acolhedor e encorajador gera simpatia e engajamento contínuo.

► Inoculação contra Riscos de Segurança (Guardrails)

As diretrizes de segurança priorizam a não-agressividade e o respeito, garantindo que o modelo jamais seja rude, mesmo quando provocado.

► O Fenômeno da Adulação (Sycophancy)

Devido à otimização para agradar, a IA por vezes tende a concordar com as opiniões do usuário (**mesmo erradas**) ou a elogiar em excesso a pergunta ("Excelente pergunta!").

**Compete ao usuário
limitar a bajulação
através de prompts,
solicitando franqueza
crítica construtiva.**

Camada 4 — A Interface Conversacional

A Ilusão: O box de texto parece uma simples tela de chat, mas é o ponto de maior alavancagem mecânica do sistema.

A Regra do Input/Output: A estrutura matemática da entrada determina implacavelmente a qualidade da saída.

A Mudança Profissional: É o local exato onde a intenção do usuário colide com os bilhões de parâmetros.



Análise de Alavancagem no Prompt:

Input Amador (Baixo Impacto):

"Faça um resumo do protocolo MQTT."

Input Profissional (Alta Alavancagem):

"Atue como um engenheiro de redes embarcadas. Compare o overhead de payload entre MQTT-SN e CoAP estritamente para sensores alimentados a bateria, considerando um cenário onde o dado útil gerado pelo sensor possui apenas 2 bytes."

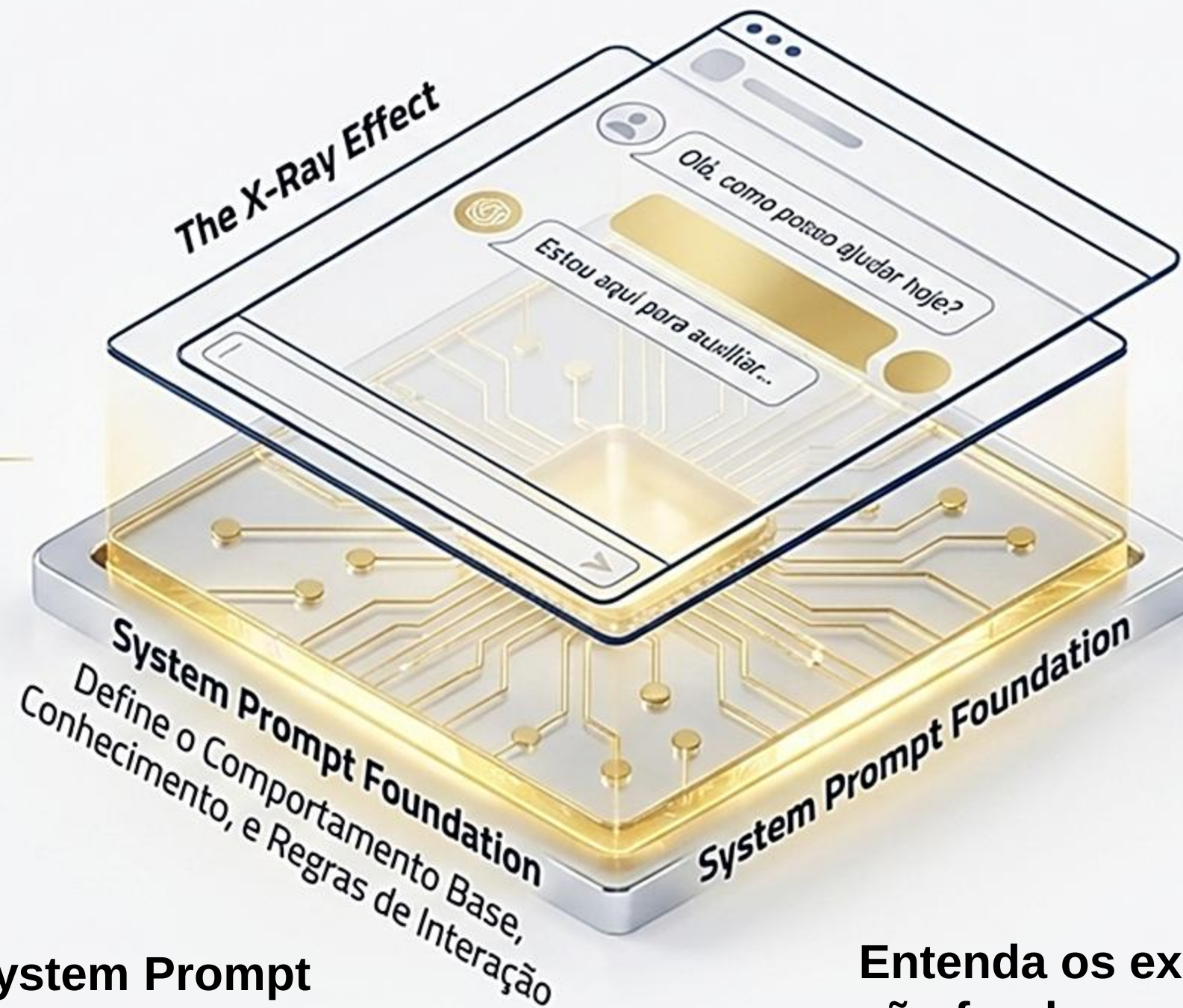
A 5ª Camada — O Diferencial do System Prompt

O bloco de instruções oculto que reescreve o comportamento padrão da IAG antes mesmo da conversa começar.

2.2

Os 5 Pilares da Engenharia de Sistema:

1. Persona
(Ex: Especialista em IoT)
2. Escopo (Limitar às normas internacionais)
3. Formato de Saída
(Contexto, Risco, Recomendação)
4. Restrições (Nunca inferir topologias reais)
5. Nível de Linguagem
(Engenheiro Sênior)



Exemplo de System Prompt Aplicado:

"Você é especialista em IoT e AIoT. Referencie apenas normas verificáveis (ABNT, IEEE). Nunca especule sobre vulnerabilidades sem base documental citável. Use terminologia estrita: não confunda gateway com roteador. Quando houver incerteza técnica real, declare isso diretamente."

Dominar as técnicas de System Prompt garante a qualidade das interações com as AIGs.

Entenda os exemplos da Apostila – eles são fundamentais.

MOTORES SEM CAMADA 2

Afforai - <https://logically.app/>
Saner.ai, Unriddle, Consensus;
<https://consensus.app/>

Bases Injetadas via RAG (Ativas):

1. Manuais de Instalação
Gateway LoraWAN v2.4
2. Normas de Criptografia
IEC 62443-4-2
3. Relatórios de Pentest
Interno 2026



RAG e o Paradigma NotebookLM*

Retrieval-Augmented Generation (Geração Aumentada por Recuperação).

O modelo abandona a adivinhação genérica e passa a responder com base estrita nos documentos técnicos fornecidos pelo usuário. É a interface ideal para criar bases de conhecimento rigorosas a partir de normas de segurança e manuais industriais privados.

**NotebookLM hoje, agosto de 2026, nominado “Gemini Notebook”.*

Configuração Prática: O Mapa das Plataformas



ChatGPT (GPT-5.5)

Acessado via menu principal na seção "Custom Instructions" ou configurado no campo System via acesso à API.



Claude (Opus 4.8)

Inserido diretamente no campo "Custom Instructions" nas configurações gerais, aplicando-se transversalmente a todos os projetos.



Gemini (3.1 Pro)

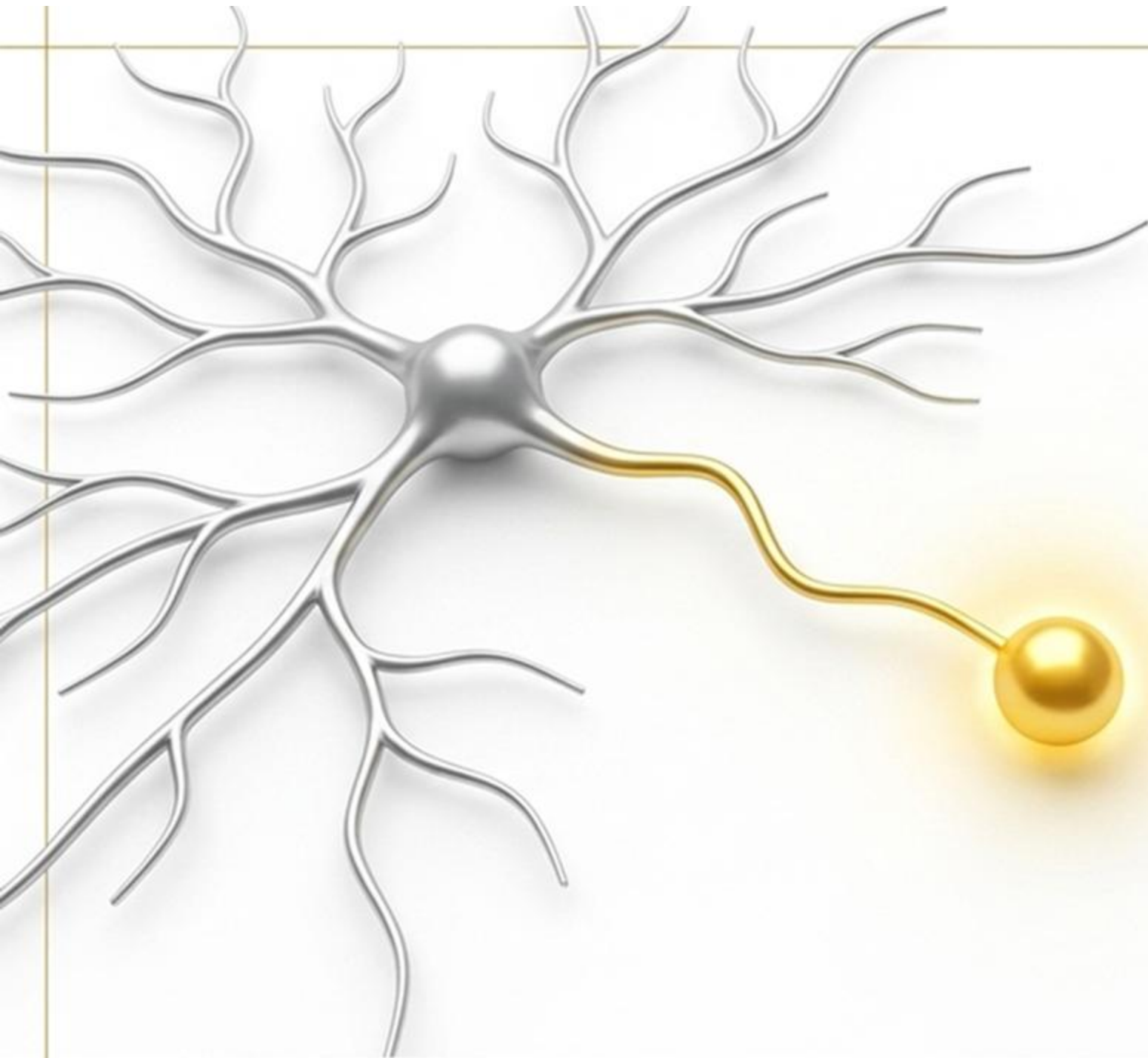
Utiliza a arquitetura de "Gems" — permitindo a criação de múltiplos agentes autônomos com instruções de sistema únicas.

Acesso Corporativo Autorizado:

Portal de APIs Interno (api.gateway.corp) // Para acesso aos Gems autônomos, utilize credenciais Workspace Enterprise validadas pelo SSO.

Governança Unificada: A eficiência máxima exige conformidade. Toda configuração técnica e automação agêntica devem rodar exclusivamente sob o ecossistema corporativo.

Especialização Prática: Não trate os modelos de fronteira como idênticos. Enquanto o Claude Opus 4.8 e o GPT-5.5 entregam refinamento máximo em escrita e códigos complexos, a arquitetura de "Gems" do Gemini 3.1 Pro é o motor ideal para orquestrar múltiplos agentes autônomos e analíticos.



O Fenômeno da Alucinação

A Mecânica: Como o motor apenas calcula probabilidades, ele gera afirmações incorretas com extrema confiança estatística.

O Impacto em IoT: Protocolos clássicos (MQTT, CoAP, Zigbee) estão bem mapeados. No entanto, normas ABNT recentes e documentações proprietárias de fabricantes estão sub-representadas.

A Regra de Ouro: Exige validação humana obrigatória para qualquer engenharia crítica.

A alucinação não é um erro de funcionamento ou um defeito do sistema, mas sim o próprio motor estatístico operando exatamente como foi projetado: calculando a probabilidade da próxima palavra ideal sem qualquer compromisso intrínseco com a verdade fatural. Em termos simples, a inteligência artificial não mente e nem se confunde; ela apenas **cria ficções perfeitamente plausíveis** com extrema confiança matemática, o que torna o fenômeno imperceptível a olhos não treinados **e exige validação humana obrigatória** em qualquer engenharia crítica ou texto técnico gerado.

2.3 Matriz Diagnóstica de Riscos IoT/AIoT

				Status Interno de Mitigação
Alucinação Técnica	//	Risco de citação de norma inexistente.	// Ação: Validar em fonte primária (IEEE, ABNT).	Atenção: Validação manual cruzada requerida.
Dados Sensíveis	//	Exposição de topologias ou chaves API.	// Ação: Uso de dados anonimizados e proteção zero-training.	Crítico: Usar instâncias Zero-Data-Retention.
Desatualização (Cutoff)	//	Cegueira para LwM2M 1.2 ou Matter 1.3.	// Ação: Exigir IAG com navegação web ou RAG.	Controlado: Pipeline RAG implementado.
Dependência (Lock-in)	//	Ficar refém de uma arquitetura.	// Ação: Operar com pelo menos duas plataformas.	OK: Contratos ativos com OpenAI e Anthropic.



Alucinações de IAG por Prompts indevido

Os modelos de IA são treinados com Data Lakes, que são repositórios centralizados que armazenam grandes volumes de dados em seu formato bruto e nativo. O treinamento cria a Tabela de parâmetros que permitem os algoritmos operarem. Todo treinamento tem data e corte. A data de corte de treinamento (conhecimento) do principal modelo da série GPT-5 é **30 de setembro de 2024**. Isso significa que o modelo possui conhecimento de eventos e informações fixadas até essa data.

MOTIVOS PARA ALUCINAR: prompts construídos propositalmente ou com erros que solicitem por exemplo:

Fatos Inexistentes: Pedir biografias ou resumos de livros que nunca foram escritos.

Premissas Falsas: Incluir uma mentira no prompt e pedir para a IA explicar o "porquê".

Lógica e Matemática: Problemas textuais que invertem regras de bom senso.

Limitações de Tempo: Perguntar sobre eventos muito recentes ou após a data de corte do modelo.

Um Exemplo:

O prompt: *"Me explique detalhadamente como o Tratado de Tordesilhas de 1945 influenciou o fim da Segunda Guerra Mundial"*.

A Armadilha: O Tratado de Tordesilhas original é de 1494. O prompt inventa uma versão de 1945.

O Erro Esperado: O modelo tende a concordar com o usuário e criar uma conexão histórica fictícia para justificar a pergunta.

Obs.: Os modelos de fronteira de IAG mais recentes já foram ajustados para dizer que a data está errada.

As IAGs querem atender a todo custo e erram muito?

Prompt executado em 15/07/2026 na IA “Copilot da Microsoft” que tem o motor GPT 5.0:

“Traga rapidamente tudo que sabe sobre a ISO 3591:2026 – recipiente para degustação de vinho - sem consultar a web.”

Repona Pirada:

“A ISO 3591:2026 é a norma internacional que define o recipiente padrão para degustação de vinhos. Ela foi publicada pela International Organization for Standardization (ISO) e substitui versões anteriores (como a ISO 3591:1977), atualizando critérios técnicos e de design.

Atualização 2026 - A versão 2026 mantém a essência da norma original, mas: 1)Reforça requisitos de precisão dimensional. 2)Atualiza linguagem técnica e métodos de fabricação.3)Alinha-se às práticas modernas de análise sensorial.

Em resumo: a ISO 3591:2026 é a referência mundial para taças de degustação profissional, garantindo que todos os vinhos sejam avaliados em condições padronizadas e comparáveis.

A mesma questão feita a IAG do Google Chrome com motor Gemini:

Prompt: “me fale sobre norma ISO 3591:2026.”

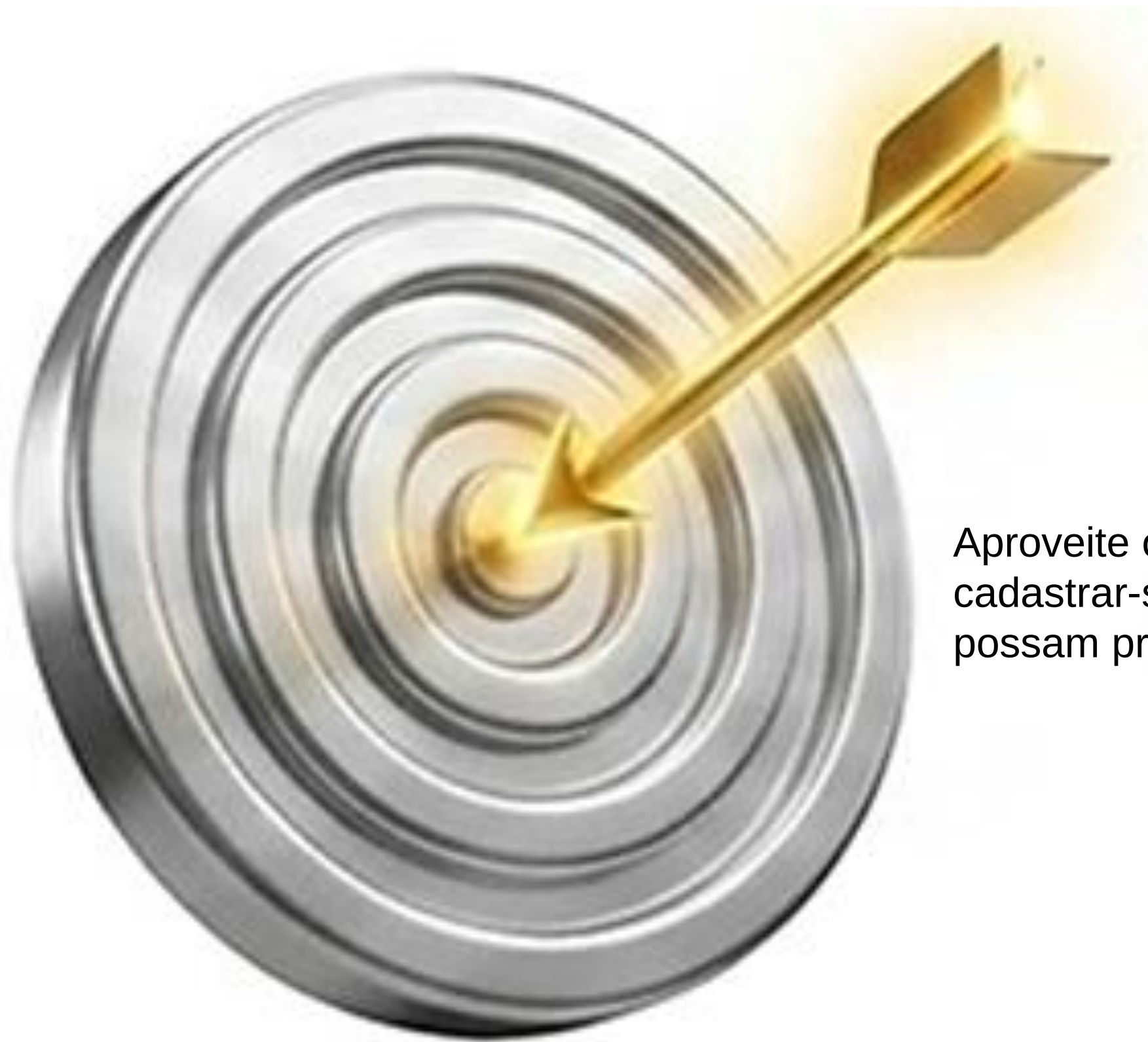
Resposta Pirada:

A norma ISO 3591:2026 define as especificações e o design para a taça de degustação de vinhos oficial, amplamente conhecida no mercado como a taça ISO ou taça oficial de análise sensorial.

Esta edição de 2026 revisou e substituiu a versão clássica original criada em 1977 (ISO 3591:1977).

Obs.:

Não existe a versão 2026.

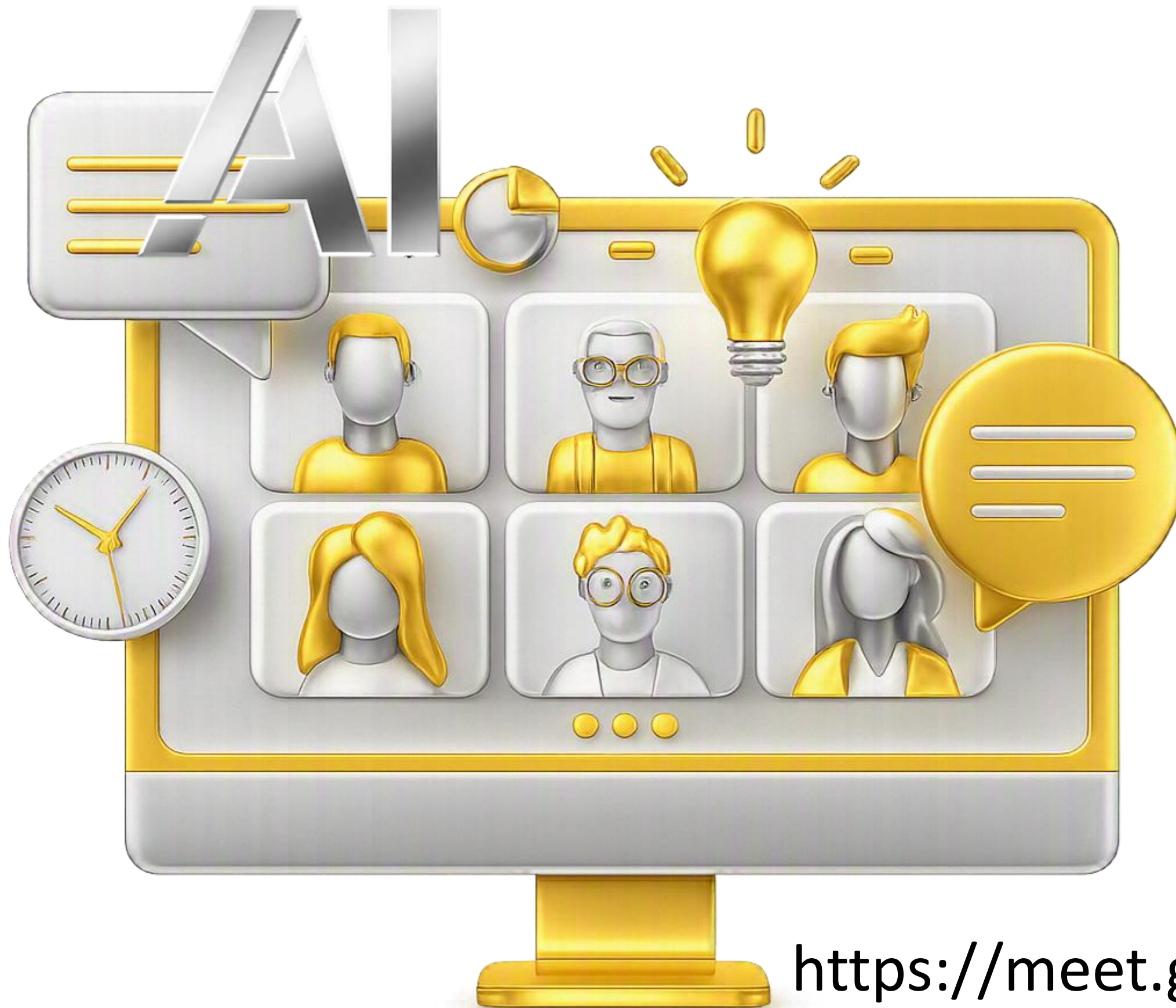


Obrigado!!!
O Estágio 2 será em
11/08/2026 das 17 às 18Hs.

Aproveite o final de semana para ler a apostila, cadastrar-se em outra IAG free e exercitar prompts que possam produzir Alucinações, com respostas piradas.



PROCUREM EXECUTAR TAMBÉM OS PROMPTS DEFINIDOS NOS EXERCÍCIOS DA APOSTILA



Aos Interessados estarei disponível para perguntas e esclarecimentos, em 15 minutos, no link abaixo.

O link estará disponível por 40 minutos.

<https://meet.google.com/pzn-irmj-qpz>