Ernani Fantinatti ernanif@fantinatti.com
Fábio Rhein fabiorhein@gmail.com
Alberto Côrtes cortes.alberto06@gmail.com
Luiz Eduardo luizhermesgarcia@gmail.com
Iluska Gastaldelle iluska.gastaldelle@gmail.com

SkyNET

https://github.com/efantinatti/SkyNET-I2A2

Agentes de IA: Arquitetura, Frameworks e Considerações Estratégicas

Aug 4, 2025

Abstract

Artificial Intelligence (AI) Agents represent a paradigm shift, transforming passive generative models into autonomous systems capable of reasoning, planning, and action. This transition places unprecedented demand on specialized datacenter infrastructure, creating exponential challenges in energy consumption, operational costs, and environmental sustainability. This paper analyzes the fundamental architecture of AI agents, the revolution in datacenter infrastructure they necessitate, and the resulting strategic, economic, and social implications. The analysis reveals that the growth of agents is driving a multi-trillion-dollar datacenter market, redefining networking and cooling requirements, and creating a sustainability crisis. Simultaneously, it is transforming the labor market by generating a net positive job growth and creating strategic opportunities for nations with renewable energy grids, such as Brazil. It is concluded that successfully navigating this new digital era requires a holistic approach that integrates technological innovation with robust governance, strategic planning, and a fundamental commitment to sustainability.

Introdução

Os agentes de inteligência artificial representam uma das mais significativas evoluções na tecnologia de IA, transformando sistemas passivos de geração de texto em entidades autônomas capazes de raciocínio, planejamento e ação independente. Diferentemente dos modelos tradicionais de IA generativa, os Agentes de IA podem perceber seu ambiente, tomar decisões complexas e executar ações para alcançar objetivos específicos com mínima supervisão humana.

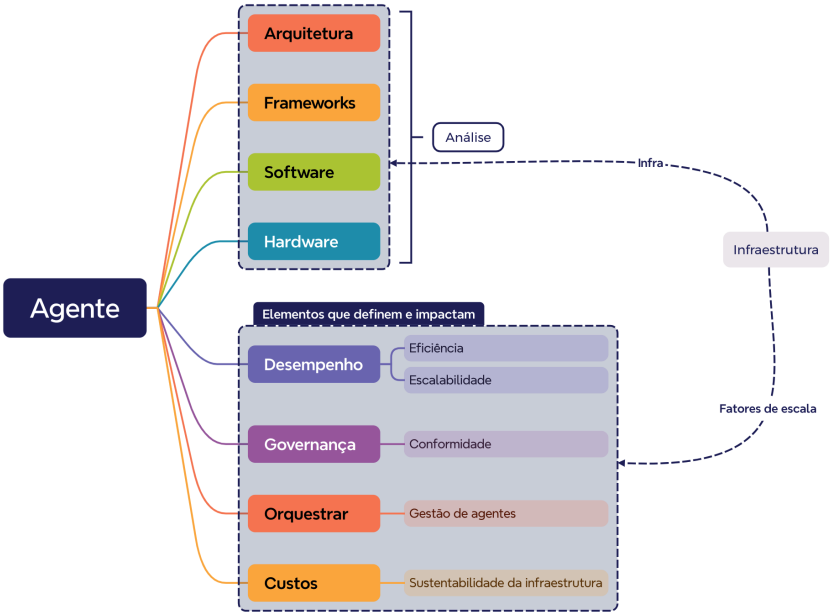

Figura 1: Mindmap da estruturação deste documento.

Arquitetura e Frameworks

Componentes Fundamentais da Arquitetura

A arquitetura de Agentes de IA é baseada em cinco componentes essenciais que trabalham em conjunto para criar sistemas inteligentes e autônomos:

1. Percepção e Processamento de Entrada O módulo de percepção é responsável por capturar informações do ambiente através de diversas fontes: dados de sensores, entradas de usuários, interações ambientais ou databases. Para agentes conversacionais, isso inclui processamento de linguagem natural (NLP), enquanto assistentes de voz executam speech-to-text, e robôs utilizam visão computacional. O objetivo é interpretar entradas de forma sensata, extraindo entidades de texto, transformando imagens em vetores de características ou normalizando leituras de sensores.

2. Gestão de Memória e Conhecimento Agentes de IA necessitam de capacidades robustas de memória para recordar interações passadas e manter uma base de conhecimento. A memória pode ser de curto prazo (informações relevantes dentro da sessão atual) ou de longo prazo (fatos e dados acumulados ao longo do tempo). A gestão de memória persistente introduz questões de escala e governança, pois armazenar dados excessivos pode exceder limites do sistema, enquanto armazenar informações sensíveis levanta preocupações significativas de privacidade.

3. Motor de Raciocínio e Planejamento Este componente representa o "cérebro" do agente, responsável por decidir como alcançar objetivos através de sequenciamento de ações.

Manipula raciocínio de alto nível, busca e planejamento. LLMs são excelentes em reconhecimento de padrões, mas enfrentam dificuldades com cadeias muito longas de raciocínio ou provas matemáticas sem auxílio. Agentes podem precisar combinar planejamento baseado em modelos com raciocínio livre de modelos.

4. Módulo de Ação e Execução Uma vez tomadas as decisões, o agente deve agir sobre elas. Este módulo executa as tarefas planejadas, tipicamente invocando serviços externos, APIs ou funções. Executar ações de forma segura é uma tarefa não trivial, especialmente quando agentes podem executar código arbitrário ou operar em sistemas críticos.

5. Camada de Integração e Orquestração Esta camada conecta todos os componentes e faz interface do agente com o restante do ecossistema de software. Manipula comunicação, agendamento e controle de workflow através dos componentes. Em configurações multi-agente, também orquestra a colaboração de múltiplos agentes.

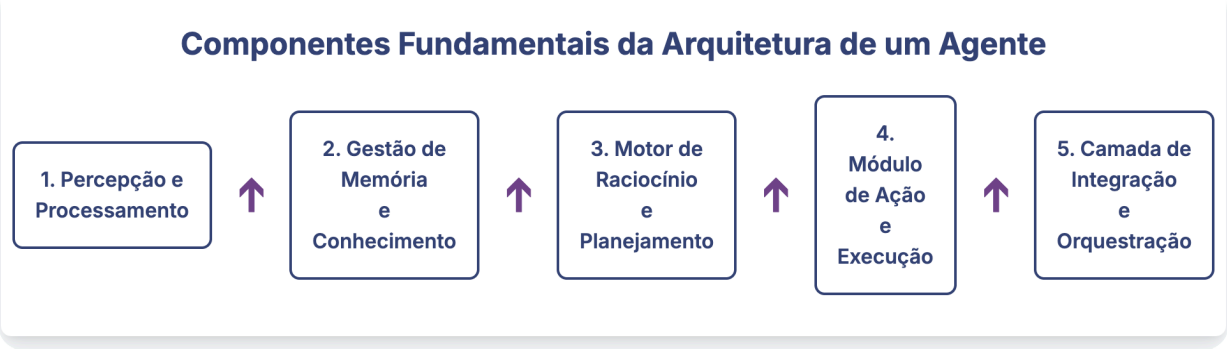

Figura 2

Padrões Arquiteturais Principais

Arquiteturas Reativas Agentes reativos operam puramente em comportamento estímulo-resposta, analisando o ambiente em tempo real e respondendo imediatamente. São ideais para tomada de decisão rápida e em tempo real onde respostas predefinidas são suficientes. Exemplo: aspiradores autônomos que usam abordagem reativa para evitar obstáculos.

Arquiteturas Deliberativas Agentes deliberativos constroem e mantêm um modelo interno do mundo, planejando ações futuras baseadas em objetivos de longo prazo. Embora mais lentos, são capazes de raciocínio complexo e planejamento estratégico.

Arquiteturas Híbridas combinam elementos reativos e deliberativos, permitindo respostas rápidas a estímulos imediatos enquanto mantêm capacidades de planejamento de longo prazo.

Arquiteturas Agentes de Aprendizado É focada na capacidade de o agente aprender com a experiência e melhorar seu desempenho ao longo do tempo. Esses agentes não são pré-programados com todas as regras, mas sim com algoritmos que lhes permitem adaptar e evoluir. O aprendizado pode ser por reforço, supervisionado ou não supervisionado.

Frameworks de Desenvolvimento Predominantes

LangChain O framework mais amplamente adotado para construção de Agentes de IA, fornecendo componentes centrais para conectar ferramentas, prompts, memória e modelos de linguagem. Oferece biblioteca de ferramentas prontas (busca, Python REPL, consulta SQL) e mecanismo para registrar ferramentas customizadas.

CrewAI Framework open-source para orquestração de equipes colaborativas de Agentes de IA. Permite que desenvolvedores atribuam agentes específicos por função a projetos compartilhados, com um roteador de tarefas coordenando transferências de contexto e rastreamento de progresso.

AutoGen Toolkit open-source para design, teste e escalonamento de agentes LLM colaborativos. Possui forte adoção em pesquisa e ambientes centrados em desenvolvedores, sendo reconhecido como plataforma flexível e experimental.

LangGraph Especializado em orquestração de agentes com estado, permitindo workflows onde agentes podem manter estado entre interações e tomar decisões dinâmicas baseadas em contexto histórico.

LlamaIndex Framework de orquestração de dados de código aberto para criar soluções de IA generativa e IA agentiva. Oferece agentes e ferramentas predefinidos e, recentemente, introduziu fluxos de trabalho, um mecanismo para o desenvolvimento de sistemas multiagentes.

Kernel Semântico Kit de desenvolvimento de código aberto da Microsoft para a criação de aplicações de IA generativa de nível empresarial. Sua framework de agente, atualmente marcada como experimental, apresenta abstrações essenciais para a criação de agentes.

Agno é um framework full-stack para sistemas multi-agentes com arquitetura nativa multimodal (texto, imagem, áudio, vídeo) e interface unificada para 23+ provedores de modelos. Oferece performance excepcional com instantanização em 3µs e uso de apenas 6,5KB de memória, essencial para sistemas de larga escala. Implementa raciocínio como funcionalidade central através de três abordagens distintas, além de busca integrada com 20+ bancos vetoriais e RAG assíncrono. Inclui memória persistente, armazenamento de sessão nativo e saídas estruturadas. Disponibiliza rotas FastAPI pré-construídas para deploy direto em produção, facilitando a implementação empresarial de workflows complexos.

Elementos de Desempenho, Eficiência e Escalabilidade

Fatores Críticos de Desempenho

Latência e Tempo de Resposta A latência é um fator crucial, especialmente para agentes que interagem com humanos ou dados streaming. Para agentes que lidam com processos em background ou assíncronos, respostas mais lentas podem ser aceitáveis. No entanto, conforme os sistemas com agentes se tornam mais complexos, a latência se acumula, exigindo otimização desde o design inicial.

Gestão de Contexto e Memória Agentes devem gerenciar janelas de contexto eficientemente, injetando as memórias certas em prompts sem sobrecarregar o LLM. Memória inconsistente ou desatualizada pode causar alucinações, ou propagação de erros.

Escalabilidade de Recursos Sistemas de Agentes de IA requerem recursos computacionais substanciais, incluindo clusters GPU de alto desempenho, farms de CPU para execução de modelos, memória suficiente para processamento em tempo real, e sistemas de armazenamento distribuído para armazenamento persistente (ver **Anexo I** para uma análise detalhada sobre a infraestrutura de data centers para IA).

Métricas de Desempenho Essenciais

Taxa de Sucesso: A taxa de sucesso mede a porcentagem de tarefas que o agente completa de forma autônoma, sem necessidade de intervenção manual. Esta métrica é fundamental para avaliar a eficiência operacional do sistema. Por exemplo, se um agente de processamento de documentos lida com 10.000 aplicações hipotecárias e 9.200 passam sem precisar de revisão manual, sua taxa de sucesso é de 92%.

Estas métricas complementares avaliam a qualidade das decisões do agente:

- **Precisão:** Mede a exatidão do agente ao identificar ou classificar resultados específicos (quantos dos casos sinalizados estão corretos)
- **Recall (Sensibilidade):** Mede a capacidade do agente de identificar todos os casos relevantes (quantos dos casos reais foram detectados)

Exemplo prático: Em detecção de fraude, recall de 60% significa que 40% de casos potenciais estão passando despercebidos.

Taxa de Automação: mede a proporção de um workflow que o agente AI manipula end-to-end. Alta taxa de automação significa menos pontos de toque humanos, tempo de ciclo reduzido e maior consistência.

Para agentes que utilizam RAG (Retrieval-Augmented Generation), a avaliação da qualidade é mais complexa. Adota-se a 'tríade de métricas RAG':

- **Relevância do Contexto**: que mede se os documentos recuperados são pertinentes à pergunta;

- **Fidelidade (Groundedness)**: que avalia se a resposta é factualmente consistente com o contexto recuperado, evitando alucinações;
- **Relevância da Resposta**: verifica se a resposta final atende à intenção original do usuário.

Desafios de Escalabilidade

Latência de Planejamento e Comunicação Estudos benchmarking revelam desafios críticos como latência prolongada de planejamento e comunicação, interações redundantes entre agentes, mecanismos complexos de controle de baixo nível, inconsistências de memória, e exploding prompt lengths.

Deterioração de Performance com Escala Pesquisas demonstram declínios acentuados nas taxas de sucesso e eficiência de colaboração reduzida conforme o número de agentes aumenta. Multi-agent systems enfrentam custos computacionais elevados e desafios únicos ainda pouco explorados.

Segurança, Governança e Conformidade

Desafios de Segurança Únicos

Vulnerabilidades Específicas de Agentes Agentes de IA introduzem riscos de segurança que diferem significativamente dos sistemas tradicionais. Identificam-se nove ameaças primárias organizadas em cinco domínios-chave: vulnerabilidades de arquitetura cognitiva, ameaças de persistência temporal, vulnerabilidades de execução operacional, violações de fronteiras de confiança e circunvenção de governança.

Ameaças Multi-Agente Sistemas multi-agente descentralizados criam desafios de segurança além dos frameworks tradicionais de cibersegurança e segurança de IA. Protocolos de forma livre permitem novas ameaças como colisão secreta e ataques coordenados de enxame. Efeitos de rede podem espalhar rapidamente violações de privacidade, desinformação, jailbreaks e envenenamento de dados.

Frameworks de Governança

Framework ETHOS Propõe-se o framework ETHOS (Ethical Technology and Holistic Oversight System), um modelo de governança descentralizada que leverage tecnologias Web3, incluindo blockchain, smart contracts e organizações autônomas descentralizadas (DAOs). Estabelece um registro global para Agentes de IA, permitindo classificação dinâmica de risco, supervisão proporcional e monitoramento automatizado de conformidade.

Framework SHIELD Propõe estratégias práticas de mitigação projetadas para reduzir exposição empresarial, incluindo controles de segurança específicos para agentes, validação de entrada, sanitização de saída e enforcement de políticas de conformidade.

NIST AI Risk Management Framework Fornece orientação específica sobre implementação de controles de segurança para sistemas IA, recomendando abordagem estruturada para mitigar riscos potenciais através de política IA robusta, regulamentação e governança de dados.

Requisitos de Conformidade

Regulamentações Emergentes O AI Act da União Europeia representa a primeira e mais abrangente regulamentação de IA, classificando uso de IA por risco para definir usos aceitáveis e inaceitáveis. Estabelece framework legal dentro da UE com penalidades quantificadas para não conformidade, podendo resultar em multas de até €35 milhões ou 7% do faturamento anual global da empresa.

Frameworks de Conformidade Frameworks de conformidade para Agentes de IA são diretrizes estruturadas e protocolos técnicos meticulosamente projetados para garantir que Agentes de IA operem eticamente, legalmente e de acordo com regulamentações específicas da indústria e normas sociais.

Orquestração e Gestão de Agentes

Conceitos Fundamentais de Orquestração

Definição e Propósito Orquestração de Agentes de IA é o processo de gerenciar e coordenar múltiplos Agentes de IA especializados - cada um habilidoso em uma tarefa específica - em um framework coeso. Em vez de um chatbot monolítico tentando fazer tudo, esta abordagem divide a responsabilidade entre agentes inteligentes que colaboram.

Padrões de Orquestração Sistemas de orquestração podem ser categorizados em diferentes padrões baseados em sua estrutura organizacional:

- **Orquestração Centralizada:** Um controlador central gerencia quando e como agentes agem
- **Orquestração Descentralizada:** Agentes operam com autonomia e colaboração peer-to-peer
- **Orquestração Híbrida:** Combina elementos centralizados e descentralizados
- **Orquestração Modular:** Foca em workflows componíveis em vez de colaboração explícita multi-agente

Plataformas de Orquestração Líderes

AgentFlow: Plataforma purpose-built para finanças e seguros que orquestra agentes desde parsing de documentos até scoring de risco. Oferece *audit trails* integrados, scoring de confiança e controles de acesso granulares que satisfazem SOC 2 e GDPR.

CrewAI: Framework open-source para orquestração de equipes colaborativas de Agentes de IA que permite que desenvolvedores atribuam agentes específicos por função a projetos compartilhados, com coordenação de transferências de contexto e rastreamento de progresso.

Aisera: Fornece motor de orquestração end-to-end que cria experiência cognitiva habilitando ações proativas e conversas human-like. Orquestra agentes específicos de domínio para resolver tarefas através da empresa.

Coordenação Multi-Agente

Estratégias de Coordenação: Coordenação efetiva de agentes requer estratégias bem definidas: protocolos de negociação (Contract Net Protocol), sistemas de votação, ou alocação de recursos baseada em leilão para ajudar agentes a resolver conflitos e alcançar consenso.

Desafios de Coordenação: O principal desafio na coordenação de agentes é balancear a autonomia com objetivos de todo o sistema. Agentes frequentemente têm objetivos locais ou visibilidade limitada do sistema mais amplo, o que pode levar a decisões subótimas.

Avaliação de Custos e Sustentabilidade

Estrutura de Custos de Agentes de IA

Custos de Desenvolvimento Custos de desenvolvimento variam significativamente baseados na abordagem:

- **Desenvolvimento customizado:** $20.000-$ 60.000 dependendo da complexidade
- **Plataformas de IA:** $5.000-$ 50.000/ano em taxas de licença base
- **Freelancers:** $20-$ 100 por hora
- **Agências especializadas:** $5.000-$ 20.000 para serviços completos

Custos Operacionais Custos operacionais incluem múltiplos componentes:

- **Custos de API:** GPT-4 custa $0.03 por 1.000 tokens de prompt e $0.06 por 1.000 tokens de completion
- **Infraestrutura:** $5-$50/mês para deployments básicos, $50-$200/mês para servidores cloud de média escala
- **Para aplicações mid-sized**: Taxas de API sozinhas podem somar $3.000-$7.000 para 100.000 queries por mês

Figura 3

*A Figura 3 ilustra a desproporção entre os custos de API e de infraestrutura para uma aplicação de médio porte. Para uma aplicação de médio porte com 100.000 consultas/mês, o custo de API com um modelo como o GPT-4 pode representar mais de 95% do custo operacional total. Por exemplo, um custo mensal de $5.000 poderia ser composto por $4.800 em chamadas de API e apenas $200 em custos de infraestrutura de nuvem (servidores, banco de dados), evidenciando a importância da otimização de tokens.

Modelos de Pricing Emergentes

Pricing Baseado em Uso

- **Pricing por conversa:** Salesforce Agent Force cobra $2 por conversa
- **Pricing por tempo:** Microsoft Copilot cobra $4 por hora
- **Pricing por token:** Modelo OpenAI Operator usa $15/1M tokens inputs; $60/1M tokens outputs

Pricing Baseado em Resultado

- **Por resolução bem-sucedida:** Intercom FinAI cobra $0.99 por resolução bem-sucedida
- **Taxas de sucesso:** Zendesk AI e Sierra AI cobram por resolução bem-sucedida

Digital AI Agent Seats Alguns fornecedores tratam agentes de IA como "usuários" únicos com suas próprias chaves de acesso API. Cada agente AI recebe seu assento, garantindo acesso específico a recursos da plataforma como um usuário humano.

Considerações de Sustentabilidade

Impacto Ambiental Agentes de IA introduzem preocupações sérias sobre custo de sistema, eficiência e sustentabilidade. Pesquisas revelam que enquanto agentes melhoram precisão com computação aumentada, sofrem de retornos rapidamente diminuídos, variância de latência ampliada e custos de infraestrutura insustentáveis.

Crise de Sustentabilidade A mudança de inferência estática de turno único para workflows agentes de múltiplos turnos amplia generalização de tarefas e flexibilidade comportamental, mas também introduz demandas computacionais profundas, revelando uma crise de sustentabilidade iminente.

Estratégias de Otimização Para mitigar custos e melhorar sustentabilidade:

- Usar modelos pré-treinados em vez de construir do zero
- Escolher complexidade de modelo adequada - nem todas tarefas requerem modelos deep learning complexos
- Aproveitar plataformas low-code/no-code para reduzir custos de desenvolvimento
- Otimizar coleta e labeling de dados por meio de geração de dados sintéticos

Tendências Emergentes e Considerações Futuras

Arquiteturas Descentralizadas e Soberania de IA

Mudança para Infraestruturas Descentralizadas Há uma mudança crescente para infraestruturas AI descentralizadas, enfatizando soberania de dados e reduzindo dependência de provedores cloud centralizados. Na Europa, a iniciativa GAIA-X está impulsionando o desenvolvimento de infraestrutura de dados descentralizada para garantir que dados da UE permaneçam dentro de suas fronteiras.

Model Context Protocol (MCP) A introdução do Model Context Protocol estabeleceu novo padrão para como modelos AI integram e compartilham dados com ferramentas e sistemas externos. MCP fornece interface model-agnostic, permitindo que Agentes de IA interajam perfeitamente com várias aplicações.

Inovações em Segurança e Confiança

Avanços em Segurança de Agentes de IA Conforme Agentes de IA se tornam mais autônomos, garantir sua segurança e confiabilidade é paramount. Inovações como o Coral Protocol estão sendo desenvolvidas para facilitar comunicação segura e coordenação entre Agentes de IA, estabelecendo a fundação para a "Internet dos Agentes".

Leveraging Infraestrutura eSIM Telecom-Grade Explora-se leverage de infraestrutura eSIM telecom-grade para fornecer identidades seguras e verificáveis para agentes de IA, melhorando sua confiabilidade em ambientes empresariais.

Inovações em Infraestrutura Computacional

Avanços em desenvolvimento de Chips Os agentes de IA demandam hardware muito mais especializado, capaz de lidar com essa inferência constante. Para evitar gargalos, empresas de tecnologia estão colaborando com fabricantes de chips para desenvolver soluções específicas com baixa latência. Gigantes da tecnologia como Meta, OpenAI, Google, Amazon e Anthropic estão desenvolvendo juntas o hardware, a infraestrutura e os sistemas de controle que darão vida à primeira força de trabalho digital verdadeiramente autônoma do mundo.

Eficiência energética Fornecedoras de infraestrutura, como a Lenovo, já estão oferecendo chips de IA para dispositivos de borda e racks de data center adaptados para inteligência distribuída. Isso permite que os agentes tomem decisões localmente, em tempo real, mantendo a conexão com modelos mais robustos hospedados na nuvem. À medida que os agentes de inteligência artificial se expandem dos grandes data centers para a borda da rede, a infraestrutura por trás deles também precisa se tornar inteligente, distribuída e autônoma. Nesse novo cenário, a IA deixará de estar apenas no código – ela estará gravada no próprio silício.

Direções Futuras

Hibridização de Coordenação Hierárquica e Descentralizada Pesquisas identificam a hibridização de coordenação hierárquica e descentralizada, coordenação humanas e baseados em LLM como direções futuras promissoras.

Collaborative Intelligence O conceito de inteligência colaborativa, combinando expertise humana e agentes de IA para confrontar obstáculos sofisticados de sustentabilidade, está emergindo como paradigma transformador.

Conclusão

Os agentes de IA representam uma evolução fundamental na tecnologia de inteligência artificial, oferecendo capacidades autônomas que vão muito além da simples geração de texto. Sua implementação bem-sucedida requer consideração cuidadosa de múltiplos fatores: desde arquiteturas robustas e frameworks apropriados até gestão eficiente de custos e conformidade regulatória.

As organizações que adotam agentes de IA devem balancear inovação com responsabilidade, implementando frameworks de governança adequados enquanto otimizam para performance e sustentabilidade. O sucesso neste novo paradigma exigirá abordagem holística que combine excelência técnica com considerações éticas e estratégicas de negócio.

Conforme a tecnologia continua evoluindo rapidamente, organizações devem manter-se adaptáveis e preparadas para integrar novas inovações enquanto mantém foco em criar valor real e sustentável através da implementação responsável de agentes de IA.

Anexo I

O Impacto Revolucionário dos Datacenters de IA

O Boom dos Datacenters de IA

A explosão da inteligência artificial está criando uma demanda sem precedentes por infraestrutura de datacenter especializada. O valor do mercado de datacenters de IA atingiu US$13,62 bilhões em 2024 e está previsto para crescer a uma taxa acelerada de 28,3% CAGR até 2030. Esta expansão representa uma transformação fundamental na infraestrutura digital global, com grandes empresas de tecnologia planejando investir dezenas de bilhões de dólares em datacenters e infraestrutura de IA.

Demandas Energéticas Exponenciais

Os datacenters de IA apresentam necessidades energéticas drasticamente superiores aos centros de dados tradicionais. Uma instalação de datacenter tradicional de cinco acres, ao ser aumentada com unidades de processamento gráfico especializadas, pode ver seu uso de energia aumentar de 5 para 50 megawatts. Os maiores datacenters em construção ou planejamento pelos hyperscalers podem requerer até 2.000 MW - equivalente a 2 gigawatts.

Previsões indicam que a demanda global de energia por datacenters aumentará 50% até 2027 e até 165% até o final da década. Por volta de 2030, os datacenters podem representar até 21% da demanda global de energia quando o custo de entregar IA aos clientes é considerado.

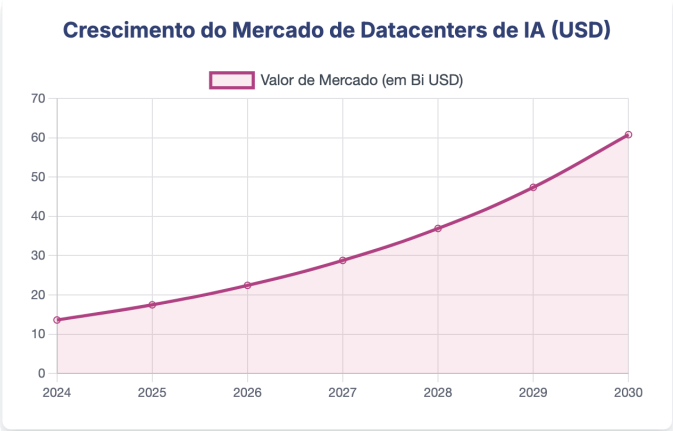

Figura 4

Desafios de Infraestrutura e Poder

Densidade de Energia e Resfriamento As densidades de potência de rack em datacenters de IA estão aumentando de 40 kW para 130 kW, com projeções chegando a 250 kW. Esta alta densidade gera quantidades enormes de calor, exigindo sistemas de resfriamento sofisticados. Os sistemas tradicionais de resfriamento a ar estão se tornando obsoletos para cargas de trabalho de IA, com 73% das novas instalações de IA implantando sistemas de resfriamento direto ao chip ou por imersão.

Limitações da Infraestrutura de Transmissão Os gargalos de infraestrutura de energia são um impedimento importante para o desenvolvimento de datacenters. Em muitos mercados, pode levar quatro anos ou mais para ter linhas de energia de alta capacidade estendidas para novos locais de desenvolvimento. A maioria desse atraso está associada à obtenção de servidões e aprovações regulamentares.

A Revolução dos Agentes de IA e Sua Demanda por Infraestrutura

Arquitetura Especializada para Agentes de IA

Os agentes de IA requerem infraestrutura computacional significativamente mais robusta que sistemas de IA tradicionais. Diferentemente dos modelos passivos de geração de texto, os agentes de IA devem perceber seu ambiente, tomar decisões complexas e executar ações para alcançar objetivos específicos com mínima supervisão humana.

Requisitos Computacionais Intensivos

Os workloads de agentes de IA, especialmente para deep learning e modelos de IA generativa, exigem poder computacional massivo. O treinamento de modelos como GPT-4 ou Gemini da Google envolve o processamento de trilhões de parâmetros, exigindo milhares de GPUs (Unidades de Processamento Gráfico) ou TPUs (Unidades de Processamento Tensor) de alto desempenho.

Demanda por Baixa Latência

Agentes de IA que interagem com humanos ou dados streaming exigem latência extremamente baixa. Conforme sistemas agênticos se tornam mais complexos, a latência se acumula, exigindo otimização desde o design inicial.

Infraestrutura de Comunicação Intra-Datacenter

A implantação de IA está remodelando fundamentalmente os requisitos de comunicação interna do datacenter. Dentro do datacenter moderno de IA, a rede uniforme está dando lugar a

uma arquitetura cuidadosamente dividida que reflete a crescente divergência entre serviços cloud convencionais e as necessidades vorazes da IA.

Segmentação de Rede Especializada A rede frontend permanece como backbone familiar para interações de usuários externos e aplicações cloud tradicionais. Porém, uma nova rede altamente especializada emergiu, dedicada inteiramente às demandas de workloads de IA orientados por GPU. Neste backend, as velocidades de porta sobem para 400 ou até 800 gigabits por segundo por GPU, e a latência é medida em sub-microssegundos.

Sustentabilidade e Energia Renovável em Datacenters

O Imperativo da Sustentabilidade

O crescimento exponencial dos datacenters de IA está criando uma crise de sustentabilidade. A demanda energética da IA é esperada para atingir 200 TWh em 2025, superando o consumo anual da Bélgica. Esta realidade está forçando uma transformação fundamental em direção a soluções energéticas sustentáveis.

Integração de Energia Renovável

A integração de fontes de energia renovável nas operações de datacenters marca um passo significativo em direção a práticas sustentáveis de TI. Painéis solares, turbinas eólicas e energia hidrelétrica oferecem alternativas viáveis aos combustíveis fósseis tradicionais. No Brasil, essa transição é particularmente promissora devido à matriz energética favorável do país, com 53,88% de geração hídrica, 15,22% eólica e 7,2% solar.

Vantagens Competitivas do Brasil

O Brasil oferece vantagem competitiva significativa com suas abundantes fontes de energia que podem fornecer energia contínua e confiável aos datacenters 24 horas por dia. A abundância de energia renovável, água e terra pode servir como grande atrativo em um mundo que cada vez mais demanda sustentabilidade em novos projetos.

Estratégias de Green Datacenters

Princípios Fundamentais: Os princípios-chave dos datacenters verdes incluem hardware energeticamente eficiente, integração de energia renovável, sistemas avançados de resfriamento e estratégias de otimização de recursos. O uso de energia renovável permite redução de até 30% nos custos operacionais, diminuindo a dependência de combustíveis fósseis e aumentando a previsibilidade dos gastos.

Tecnologias de Resfriamento Avançadas: O resfriamento líquido é 3.000 vezes mais eficiente que o resfriamento a ar para hardware de IA. As implantações de resfriamento por

imersão nos próximos anos serão concentradas em instalações de IA e em seções de datacenters tradicionais executando workloads de IA.

O Mercado de Trabalho e a IA: Transformação e Oportunidades

Impacto da IA no Emprego Global

Estima-se que aproximadamente 25% dos empregos em todo o mundo estejam potencialmente expostos à inteligência artificial generativa. No Brasil, especificamente, estimativas apontam que cerca de 30 milhões de empregos podem estar em risco até 2026, com mais da metade de todos os empregos nos municípios brasileiros podendo ser ameaçados até 2040. A Figura 5 demonstra visualmente este crescimento.

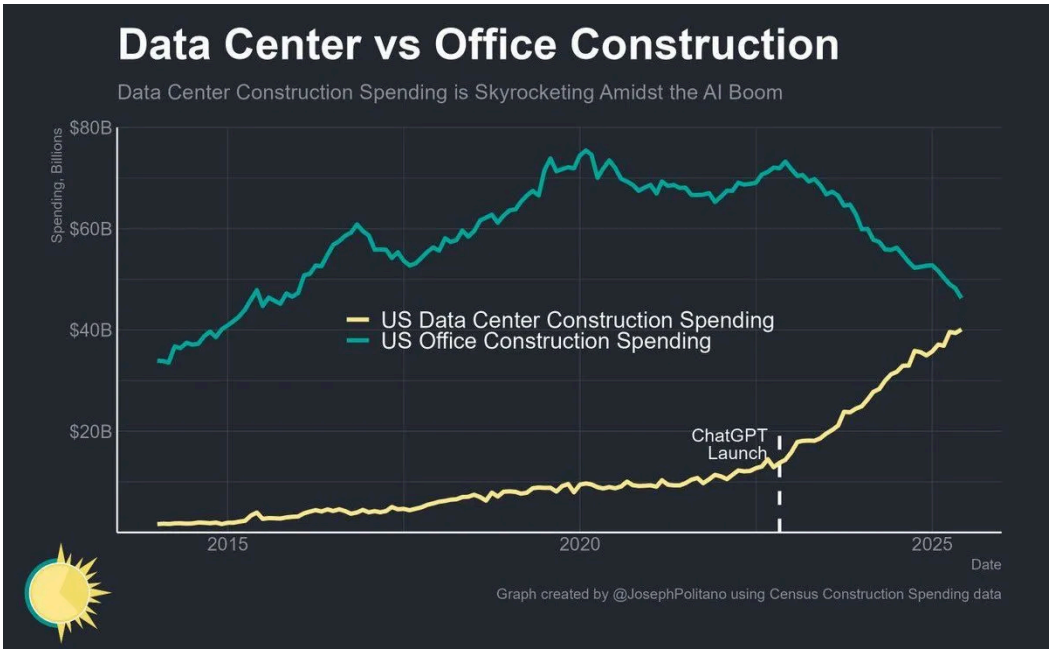

Figura 5

Crescimento Explosivo de Vagas de IA no Brasil

Contrariando as preocupações sobre desemprego, dados mostram um crescimento explosivo em postagens de trabalho relacionadas à IA no Brasil: de 19.000 em 2021 para 73.000 em 2024, representando um aumento de quase quatro vezes em apenas três anos. A participação de postagens de trabalho exigindo habilidades de IA cresceu de forma constante, atingindo 1,1% em 2024.

Criação vs. Eliminação de Empregos

Geração Líquida Positiva Pesquisas indicam que a inteligência artificial tem potencial para gerar mais empregos do que substituir a mão de obra humana. Um estudo aponta que a IA pode criar cerca de 2,7 milhões de empregos líquidos apenas no Reino Unido até 2037. Previsões indicam que enquanto 85 milhões de empregos podem ser deslocados até 2025, 97 milhões de novos papéis podem emergir (Figura 4).

Novas Profissões Emergentes A IA está criando diversos novos postos de trabalho e estimulando outros já existentes, especialmente nas áreas de tecnologia da informação, marketing digital, ciência de dados e assistência virtual. Profissões como especialistas em big data, engenheiros de fintech, peritos em inteligência artificial, analistas de dados surgem entre as que mais vão crescer.

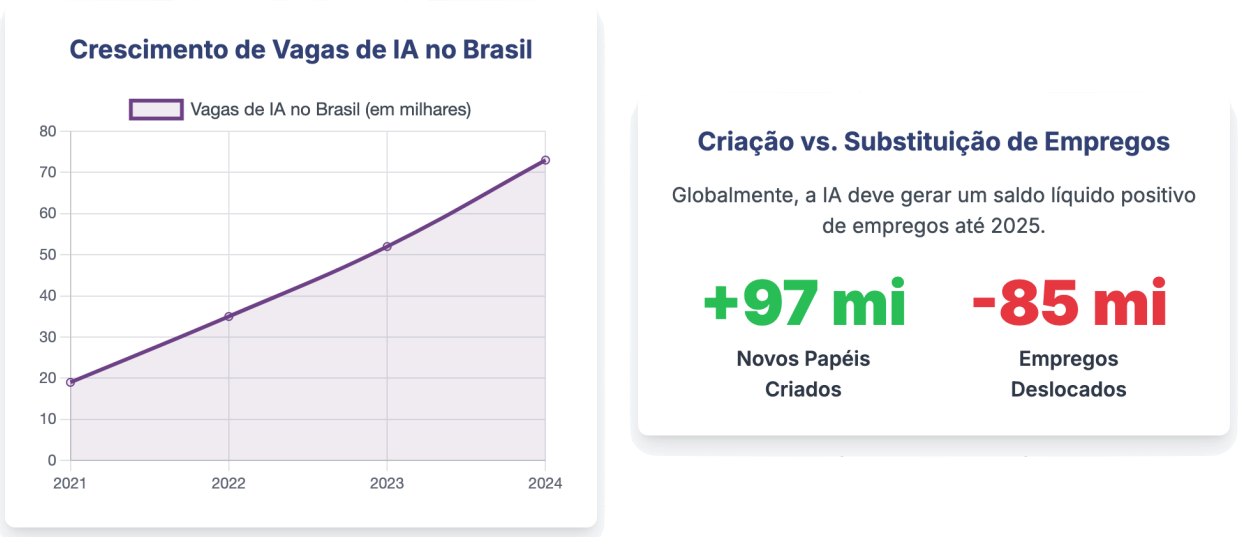

Figura 6

Desafios e Oportunidades no Brasil

Uso Disseminado mas Formação Limitada Uma pesquisa revela que 68% dos profissionais brasileiros já usam ferramentas de IA pelo menos uma vez por dia, mas apenas 33% têm acesso formal e treinamento no trabalho. Impressionantes 90% dos trabalhadores brasileiros acreditam que a IA melhorará sua eficácia no trabalho e 84% dizem estar animados com o futuro do trabalho com IA.

Impacto Diferenciado por Setor Mulheres e trabalhadores altamente educados enfrentam maior exposição ocupacional à IA, tanto em alta quanto em baixa complementaridade. Trabalhadores na cauda superior da distribuição de ganhos são mais propensos a estar em ocupações com alta exposição, mas também alto potencial de complementaridade.

Políticas e Governança para o Futuro Digital

Framework Regulatório Brasileiro

O governo brasileiro está no meio de discussões sobre uma nova política nacional para datacenters, com o objetivo de aproveitar o que o governo vê como uma "janela de oportunidade" para atrair investimentos ao setor. Isenções de impostos federais e redução de tarifas de importação de equipamentos estão entre os incentivos que o governo oferecerá aos investidores.

Ausência do Ministério do Meio Ambiente

Uma preocupação significativa é que, em discussões governamentais sobre datacenters envolvendo centenas de funcionários federais, o Ministério do Meio Ambiente não tem participado. Esta ausência é preocupante dado o potencial impacto ambiental dos datacenters, particularmente em relação ao uso de água e energia.

Declaração BRICS sobre IA e Trabalho

Os países BRICS aprovaram uma declaração delineando um caminho para abordar os impactos da inteligência artificial e mudanças climáticas no mercado de trabalho, com ênfase na proteção social e requalificação profissional. O documento reconhece que a inteligência artificial está remodelando radicalmente as relações de trabalho, criando novas oportunidades mas também trazendo riscos como deslocamento de empregos e crescimento das desigualdades.

Investimentos e Projeções Futuras

Corrida de US$ 7 Trilhões

Pesquisas mostram que até 2030, os datacenters estão projetados para exigir US$6,7 trilhões mundialmente para acompanhar a demanda por poder computacional. Em um cenário de demanda acelerada, o crescimento pode exigir investimentos de até US$7,9 trilhões, com 205 GW incrementais de capacidade de datacenter relacionada à IA sendo adicionados entre 2025 e 2030.

Distribuição de Investimentos

Aproximadamente 15% (US$0,8 trilhão) do investimento fluirá para construtores para terra, materiais e desenvolvimento de sites. Outros 25% (US$1,3 trilhão) serão alocados para energizadores para geração e transmissão de energia, resfriamento e equipamentos elétricos. A maior parcela do investimento, 60% (US$3,1 trilhões), irá para desenvolvedores e designers de tecnologia, que produzem chips e hardware computacional para datacenters.

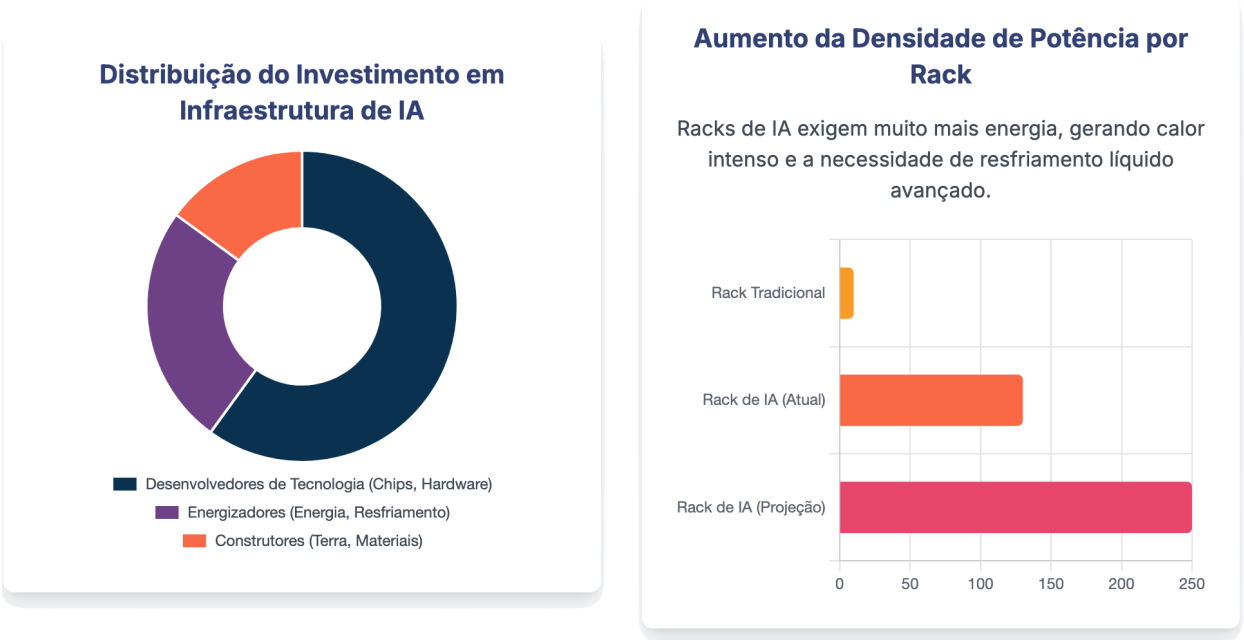

Figura 7

Oportunidades para o Brasil

O Brasil possui tudo o que é necessário para hospedar muitos datacenters, com desafios que são solucionáveis. O país tem mais do que energia renovável e água suficientes. A política nacional de datacenters tem o potencial de atrair dois trilhões de reais (cerca de US$350 bilhões) em investimentos nos próximos 10 anos.

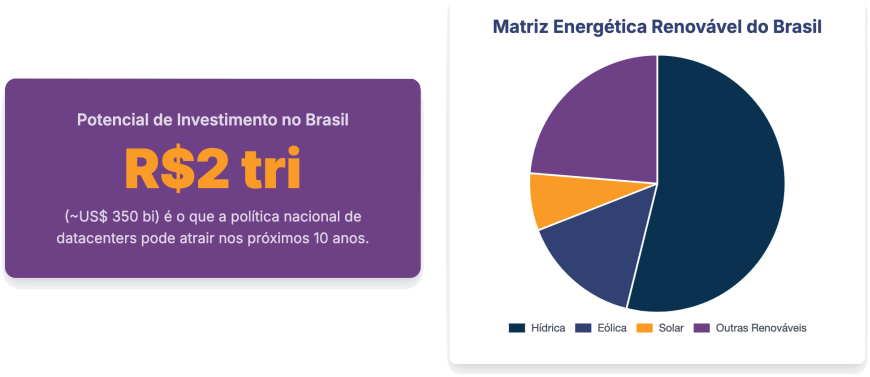

Figura 8

Conclusão: Navegando a Transformação Digital Sustentável

A convergência entre agentes de IA e datacenters representa uma transformação fundamental na infraestrutura digital global. Esta revolução apresenta tanto oportunidades extraordinárias quanto desafios complexos que exigem abordagens inovadoras e sustentáveis.

Os datacenters de IA não são apenas uma evolução natural da infraestrutura existente - eles representam um paradigma completamente novo que demanda repensar desde o design de sistemas de resfriamento até a estrutura de redes internas. Com densidades de energia que podem chegar a 250 kW por rack e demandas energéticas que podem atingir gigawatts, estes centros estão redefinindo os requisitos de infraestrutura física e energética.

O Brasil emerge como um player estratégico neste cenário global, possuindo vantagens competitivas únicas: uma matriz energética limpa com 44,8% de fontes renováveis, abundantes recursos hídricos, e um governo comprometido com políticas de incentivo ao setor. A oportunidade de atrair US$350 bilhões em investimentos nos próximos 10 anos posiciona o país como um hub regional para a infraestrutura de IA sustentável.

No mercado de trabalho, contrariando previsões pessimistas, a evidência sugere que a IA está criando mais oportunidades do que eliminando postos de trabalho. No Brasil, o crescimento de quase 400% em vagas relacionadas à IA entre 2021 e 2024 demonstra o potencial transformador da tecnologia. Contudo, é fundamental que as políticas públicas e corporativas priorizem programas de requalificação profissional para garantir uma transição justa e inclusiva.

A sustentabilidade não é mais opcional - é um imperativo estratégico. A integração de energia renovável, sistemas de resfriamento avançados e práticas de economia circular são essenciais para viabilizar o crescimento sustentável da infraestrutura de IA. As empresas que conseguirem balancear eficiência operacional com responsabilidade ambiental terão vantagens competitivas decisivas.

O futuro dos agentes de IA e datacenters será definido pela capacidade de harmonizar inovação tecnológica, sustentabilidade ambiental e desenvolvimento social. As organizações e países que conseguirem navegar esta transformação de forma holística e responsável estarão melhor posicionados para liderar a próxima era da economia digital.

Referências:

AGENTES DE IA: DEFINIÇÃO, ARQUITETURA E FUNDAMENTOS

- **What are AI agents? Definition, examples, and types | Google Cloud**
  https://cloud.google.com/discover/what-are-ai-agents

- **Exploring Generative AI Agents: Architecture, Applications, and Challenges**
  https://newjaigs.com/index.php/JAIGS/article/view/350

- **A Comprehensive Review of Gen AI Agents: Applications and Frameworks in Finance, Investments and Risk Domains**

  https://www.ijisrt.com/a-comprehensive-review-of-gen-ai-agents-applications-and-frameworks-in-finance-investments-and-risk-domains

- **Building Effective AI Agents - Anthropic**
  https://www.anthropic.com/research/building-effective-agents

- **Understanding Agent Architecture: The Frameworks Powering AI – HatchWorks**
  https://hatchworks.com/blog/ai-agents/agent-architecture/

- **7 Components of an Agentic AI-Ready Software Architecture – Aziro**
  https://www.aziro.com/blog/7-components-of-an-agentic-ai-ready-software-architecture/

- **The 12-Factor AI Agent: Building Effective AI Systems That Scale – Flowhunt**

  https://www.flowhunt.io/blog/the-12-factor-ai-agent-building-effective-ai-systems-that-scale/

- **Understanding AI Agent Infrastructure: Key Insights – Ema**
  https://www.ema.co/additional-blogs/addition-blogs/ai-agent-infrastructure-key-insights

- **Understanding AI Agent Orchestration – Botpress**
  https://botpress.com/blog/ai-agent-orchestration

- **9 AI Orchestration Platforms – Multimodal**
  https://www.multimodal.dev/post/ai-orchestration-platforms

- **Estruturas de agentes de IA: como escolher a base certa para o seu negócio - IBM**
  https://www.ibm.com/br-pt/think/insights/top-ai-agent-frameworks

AVALIAÇÃO, MÉTRICAS E GOVERNANÇA

- **AI agent evaluation: Metrics, strategies, and best practices – Wandb**

  https://wandb.ai/onlineinference/genai-research/reports/AI-agent-evaluation-Metrics-strategies-and-best-practices--VmlldzoxMjM0NjQzMQ

- **Effective governance frameworks for AI agents – IBM Developer**
  https://developer.ibm.com/articles/governing-ai-agents-watsonx-governance/

- **AI agents pose new governance challenges – Schwartz Reisman Institute**
  https://srinstitute.utoronto.ca/news/challenges-in-governing-ai-agents

- **Securing Agentic AI: A Comprehensive Threat Model and Mitigation Framework for Generative AI Agents – arXiv**
  https://arxiv.org/abs/2504.19956

DATA CENTERS, INFRAESTRUTURA E SUSTENTABILIDADE

- **25+ AI Data Center Statistics & Trends (2025 Updated)**
  https://thenetworkinstallers.com/blog/ai-data-center-statistics/

- **AI has high data center energy costs — but there are solutions – MIT Sloan**

  https://mitsloan.mit.edu/ideas-made-to-matter/ai-has-high-data-center-energy-costs-there-are-solutions

- **Can US infrastructure keep up with the AI economy? – Deloitte**

  https://www.deloitte.com/us/en/insights/industry/power-and-utilities/data-center-infrastructure-artificial-intelligence.html

- **AI to drive 165% increase in data center power demand by 2030 – Goldman Sachs**

  https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030

- **AI power: Expanding data center capacity to meet growing demand – McKinsey**

  https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insi

ghts/ai-power-expanding-data-center-capacity-to-meet-growing-demand

- **The cost of compute: A $7 trillion race to scale data centers – McKinsey**

  https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers

- **As estratégias para tornar os data centers mais sustentáveis – FAPESP**

  https://revistapesquisa.fapesp.br/as-estrategias-para-tornar-os-data-centers-mais-sustentaveis/

- **IA e energias renováveis: fórmulas para um futuro sustentável nos data centers – Data Center Dynamics**

  https://www.datacenterdynamics.com/br/not%C3%ADcias/ia-e-energias-renov%C3%A1veis-f%C3%B3rmulas-para-um-futuro-sustent%C3%A1vel-nos-data-centers/

- **Brazil's data center boom raises concerns over energy access – Daily Climate**

  https://www.dailyclimate.org/brazils-data-center-boom-raises-concerns-over-energy-access-2671285914.html

- **The AI Herd | Pulitzer Center**
  https://pulitzercenter.org/stories/ai-herd

- **Agentes de IA estão nos fazendo repensar a infraestrutura computacional global - Fast Company Brasil**
  https://fastcompanybrasil.com/ia/agentes-de-ia-estao-nos-fazendo-repensar-a-infraestrutura-computacional-global/

MERCADO DE TRABALHO, IMPACTO SOCIAL E IA

- **Labor Market Exposure to AI: Cross-country Differences and Distributional Implications – IMF**

  https://elibrary.imf.org/openurl?genre=journal&issn=1018-5941&volume=2023&issue=216&cid=539656-com-dsp-crossref

- **The Future of Work in Brazil – Horasis**
  https://horasis.org/the-future-of-work-in-brazil/

- **Inteligência artificial: aprendizado e qualificação são fundamentais – Jornal USP**

  https://jornal.usp.br/campus-ribeirao-preto/inteligencia-artificial-vai-transformar-o-mercado-de-trabalho-com-novas-oportunidades/

- **UFF Responds: Artificial Intelligence and the Job Market | Instituto IA**
  https://instituto.ia.lncc.br/en/news/uff-responds-artificial-intelligence-and-the-job-market

- **Um em cada 4 empregos será transformado pela inteligência artificial, diz OIT – ONU**
  https://news.un.org/pt/story/2025/05/1848821

- **Brazil Survey: 68% of Brazilians Use AI Everyday But Only 33% Have Formal Access**

  https://www.read.ai/post/brazil-survey-68-of-brazilians-use-ai-everyday-but-only-31-have-formal-access-and-training-at-work--and-they-want-more

- **IA pode afetar 1 em cada 4 empregos, diz OIT – G1**

  https://g1.globo.com/trabalho-e-carreira/noticia/2025/05/21/ia-pode-afetar-1-em-cada-4-empregos-diz-oit.ghtml

- **Artificial intelligence and climate: BRICS declaration proposes policies to protect workers**

  https://brics.br/en/news/artificial-intelligence-and-climate-brics-declaration-proposes-policies-to-protect-workers

Ernani Fantinatti	**06 de agosto de 2025**
Fabio Dias Rhein	**06 de agosto de 2025**
Alberto Côrtes Cavalcante	**06 de agosto de 2025**
Luiz Eduardo Hermes Garcia	**06 de agosto de 2025**