REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

INTELIGÊNCIA ARTIFICIAL (IA) LOCAL, NUVEM E ARQUITETURAS HÍBRIDAS COMO DECISÃO ECONÔMICA ESTRATÉGICA: GOVERNANÇA CORPORATIVA, ECONOMIA COMPORTAMENTAL E COMPETITIVIDADE EMPRESARIAL

LOCAL, CLOUD AND HYBRID ARTIFICIAL INTELLIGENCE (AI) ARCHITECTURES AS A STRATEGIC ECONOMIC DECISION: CORPORATE GOVERNANCE, BEHAVIORAL ECONOMICS AND BUSINESS COMPETITIVENESS

ARQUITECTURAS DE INTELIGENCIA ARTIFICIAL (IA) LOCAL, EN LA NUBE E HÍBRIDAS COMO DECISIÓN ECONÓMICA ESTRATÉGICA: GOBERNANZA CORPORATIVA, ECONOMÍA CONDUCTUAL Y COMPETITIVIDAD EMPRESARIAL

Angelo Machado de Souza[1], José Chavaglia Neto[2], Davis Souza Alves[3]

e2132
https://doi.org/10.67553/sol21.v.2i1.32

PUBLICADO: 10/2026

RESUMO

A escolha empresarial entre inteligência artificial (IA) local, IA em nuvem e arquiteturas híbridas envolve alocação de capital, risco, dependência tecnológica e capacidades organizacionais, além de fatores comportamentais que afetam a comparação das alternativas. O artigo analisa essa decisão com ênfase em cargas corporativas de inferência e, especialmente, em aplicações de IA generativa e grandes modelos de linguagem. Adota-se revisão narrativa orientada por critérios temáticos e análise documental comparativa. A bibliografia reúne 44 referências do levantamento inicial e sete fontes complementares; a análise se organiza em cinco categorias temáticas. A síntese

[1] Profissional Especialista em Segurança da Informação | Cibersegurança | Governança de TI | Privacidade de Dados, com mais de 10 anos de atuação na área de Tecnologia da Informação, com ênfase em Segurança da Informação, Governança de TI e Privacidade de Dados. Mestrado em Segurança da Informação pela Florida Christian University, com pesquisa voltada à conformidade com a LGPD. MBA em Gerenciamento de Projetos de TI e certificações internacionais, como Certified Information Security Officer (CISO), ISO/IEC 27001 Lead Implementer e DPO – EXIN. Experiência consolidada na liderança de projetos em empresas como NTT Ltd., ISH Tecnologia e Grupo Cornélio Brennand, com atuação em mitigação de riscos, resposta a incidentes, conformidade regulatória e implementação de políticas de segurança. Gestão de riscos e implementação de controles e políticas de segurança. Coordenação de projetos de cibersegurança e gestão de equipes técnicas.

[2] Economista, professor de empreendedorismo e inovação na Fundação Getúlio Vargas (FGV Management), Professor convidado no programa de doutorado em Neuromarketing da Flórida Christian University (EUA), pesquisador no Observatório de Liderança FGV-NEOP, consultor de empresas, membro do conselho editorial do International Journal of Latest Trends in Finance and Economic Sciences, no Reino Unido e do International Journal of Innovation Studies, na Suíça. É referência nas pesquisas do Census Bureau (Órgão responsável pela publicação do Censo dos Estados Unidos). Professor auxiliar de marketing turístico na Universidade Europeia (Portugal). Desenvolvimento de produtos e serviços na Universidade de Stanford. Doutor em Métodos Quantitativos e Mestre em Gestão Empresarial pelo Instituto Universitário de Lisboa ISCTE/IUL. M.B.A – em Gestão de Empresas pela Fundação Getúlio Vargas – FGV. Especialista em Controladoria e Finanças pela FCETM-MG.

[3] Doutor em Administração de TI - Ph.D na Florida Christian University (EUA) convalidado no Brasil, Mestre em Administração com foco em TI Verde, Extensão em Gestão de TI pela FGV/SP, Pós-graduado em Gerenciamento de Projetos, Graduado em Redes de Computadores e Internet. Residiu para estudos nos Estados Unidos e Nova Zelândia. Possui as certificações PMP®, ITIL® Expert, C|EH®, ITIL® Professional (4 MP), COBIT®, ISO-20000®, ISO-27002®, EXIN® Agile Scrum Master, Lean IT, Green IT, ICS MCSA® Windows Server 2003, Cloud Computing, EXIN® Data Protection Officer (DPO), & EXIN ISO-27001 Professional (ISMP), Cyber Security, CompTIA Security+, EC-Council Certified Ethical Hacker, Psicanalista, CompuTIA Security+, DAC® Wireless, DCP® Switching, DSS® IP Surveillance. Consultor de Gestão de Sustentabilidade de TI, com produtos e consultorias em Green IT para órgãos públicos municipais do Brasil, além de Consultor em Privacidade de Dados (LGPD). Sócio-gerente na Millennium Network (Linx) na coordenação da equipe técnica de projetos de infraestrutura de TI para ERP's, Professor acadêmico para Gestão de Serviços, Segurança da Informação e Redes de Computadores na Universidade Paulista - UNIP, Universidade Municipal de São Caetano do Sul - USCS (professor concursado), Universidade Federal de São Carlos - UFSCar e DARYUS/Faculdade Impacta. Membro do Congresso Científico Internacional POMS nos Estados Unidos. Fez parte do Núcleo Desenvolvedor Estruturante da UNIP do curso superior de Tecnologia em Redes de Computadores, responsável pela adequação do curso junto ao MEC. Instrutor credenciado pelo EXIN/PeopleCert com foco em ITIL®, GDPR, ISO-27001, Green IT, além de pesquisador e palestrante em diversos eventos científicos internacionais relacionados com TI Verde & GDPR na Espanha, Holanda e Estados Unidos - tendo seus estudos publicados nesses países.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

integra governança corporativa, economia comportamental, arquiteturas de computação, custos e estratégia competitiva. Formaliza-se o custo total de propriedade em valor presente e propõe-se análise de sensibilidade e ponto de equilíbrio entre alternativas. A síntese da literatura não sustenta a indicação de uma arquitetura universalmente superior: a escolha depende de utilização, volatilidade da demanda, requisitos de dados e latência, competências, custos de saída e valor de flexibilidade. A contribuição conceitual proposta articula mecanismos que relacionam arquitetura a competitividade por custo, diferenciação, velocidade, resiliência, poder de barganha e aprendizagem; a contribuição gerencial proposta é um modelo de decisão rastreável que combina comparação econômica, auditoria comportamental, governança e revisão periódica. O modelo não foi submetido a validação empírica.

PALAVRAS-CHAVE: Inteligência artificial. Governança corporativa. Economia comportamental. Custo total de propriedade. Competitividade.

ABSTRACT

The choice among local artificial intelligence (AI), cloud AI, and hybrid architectures involves capital allocation, risk, technological dependence, organizational capabilities, and behavioral factors that affect how alternatives are compared. This article focuses on enterprise inference workloads, particularly generative AI and large language model applications. It adopts a narrative review guided by thematic criteria and comparative documentary analysis. The bibliography comprises 44 references from the initial review and seven supplementary sources; the analysis is organized into five thematic categories. The synthesis integrates corporate governance, behavioral economics, computing architectures, costs, and competitive strategy. Total cost of ownership is formalized in present-value terms, with sensitivity and break-even analysis across alternatives. The literature synthesis does not support identifying a universally superior architecture: the decision depends on utilization, demand volatility, data and latency requirements, capabilities, exit costs, and the value of flexibility. The proposed conceptual contribution integrates mechanisms linking architecture to competitiveness through cost, differentiation, speed, resilience, bargaining power, and organizational learning; the proposed managerial contribution is a traceable decision model combining economic comparison, behavioral audit, governance, and periodic review. The model has not been empirically validated.

KEYWORDS: *Artificial intelligence. Corporate governance. Behavioral economics. Total cost of ownership. Competitiveness.*

RESUMEN

La elección empresarial entre inteligencia artificial (IA) local, IA en la nube y arquitecturas híbridas implica asignación de capital, riesgo, dependencia tecnológica, capacidades organizacionales y factores conductuales que afectan la comparación de alternativas. El artículo se concentra en cargas corporativas de inferencia, especialmente aplicaciones de IA generativa y grandes modelos de lenguaje. Se adopta una revisión narrativa orientada por criterios temáticos y un análisis documental comparativo. La bibliografía reúne 44 referencias del levantamiento inicial y siete fuentes complementarias; el análisis se organiza en cinco categorías temáticas. La síntesis articula gobernanza corporativa, economía conductual, arquitecturas de computación, costos y estrategia competitiva. Se formaliza el costo total de propiedad en valor presente y se propone análisis de sensibilidad y punto de equilibrio. La síntesis de la literatura no permite identificar una arquitectura universalmente superior: la decisión depende de utilización, volatilidad de demanda, requisitos de datos y latencia, capacidades, costos de salida y valor de la flexibilidad. La contribución conceptual propuesta relaciona arquitectura y competitividad mediante costo, diferenciación, velocidad, resiliencia, poder de negociación y aprendizaje; la contribución gerencial propuesta es un modelo de decisión trazable que combina comparación económica, auditoría conductual, gobernanza y revisión periódica. El modelo no ha sido validado empíricamente.

PALABRAS-CLAVE: *Inteligencia artificial. Gobierno corporativo. Economía conductual. Costo total de propiedad. Competitividad.*

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

1 - INTRODUÇÃO

A incorporação da inteligência artificial (IA) às operações empresariais tornou a arquitetura de execução uma variável de decisão econômica. Em 2025, 19,95% das empresas da União Europeia com dez ou mais pessoas ocupadas utilizaram ao menos uma tecnologia de IA; entre as grandes empresas, a proporção chegou a 55,03%. Embora restritos ao universo estatístico europeu, esses dados mostram adoção suficientemente relevante para tornar recorrentes decisões sobre infraestrutura, dados, fornecedores, segurança e capital (Eurostat, 2026). Em paralelo, relatório da Organização para a Cooperação e Desenvolvimento Econômico (OECD, sigla em inglês), do Boston Consulting Group e do Institut Européen d'Administration des Affaires (INSEAD) identificou obstáculos associados a competências, maturidade de dados e capacidade gerencial de relacionar IA a problemas concretos do negócio (OECD; Boston Consulting Group; INSEAD, 2025).

Neste artigo, a expressão IA empresarial não abrange indiscriminadamente todo o ciclo de IA. O recorte é a decisão sobre onde executar cargas corporativas de inferência e serviços associados, com ênfase em IA generativa e grandes modelos de linguagem (LLMs, do inglês large language models), porque nesses casos a escolha local-nuvem-híbrida envolve de forma explícita capacidade computacional, movimentação de dados, latência, consumo variável e portabilidade. Treinamento de modelos fundacionais na fronteira, robótica embarcada e aplicações cujo local de execução não altera materialmente a economia da solução ficam fora do núcleo analítico.

Escolher entre IA em nuvem, IA local e arquitetura híbrida não equivale a selecionar apenas um ambiente técnico. A nuvem oferece recursos compartilhados, elasticidade, provisionamento rápido e mensuração de uso (Mell; Grance, 2011). A execução próxima da origem dos dados pode reduzir latência e movimentação em determinados casos, mas transfere à organização responsabilidades de capacidade e operação (Shi et al., 2016). Estudos recentes sobre LLMs mostram ainda alternativas centralizadas, descentralizadas e colaborativas entre borda e nuvem, evidenciando que desempenho e distribuição de carga dependem do desenho específico e dos recursos disponíveis (Friha et al., 2024; Chen et al., 2025; Jiang et al., 2026).

A questão central, portanto, não é qual arquitetura é genericamente melhor, mas em quais condições cada alternativa oferece relação mais adequada entre custo, risco, flexibilidade, autonomia e criação de valor. A comparação precisa considerar despesas de capital (CAPEX), despesas operacionais (OPEX), custo total de propriedade (TCO, do inglês total cost of ownership), custos de integração e oportunidade, custos de troca, disponibilidade de competências, exigências de segurança, horizonte de utilização e valor estratégico de dados e conhecimento. Uma decisão baseada apenas no preço de servidor, unidade de processamento gráfico (GPU), token ou interface de programação de aplicações (API) omite parcelas relevantes do custo econômico total.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

A governança corporativa acrescenta a dimensão da qualidade do processo decisório. A separação entre conhecimento técnico, autoridade e exposição ao risco cria assimetrias entre conselho, gestão, áreas técnicas e fornecedores (Jensen; Meckling, 1976). Ao mesmo tempo, decisões formalmente governadas continuam sendo tomadas por pessoas sujeitas a racionalidade limitada, enquadramento, ancoragem, aversão à perda e outros efeitos documentados pela economia comportamental (Simon, 1955; Kahneman; Tversky, 1979; Tversky; Kahneman, 1981). A neuroeconomia é considerada apenas como campo interdisciplinar de contextualização; como este estudo não coleta dados neurais, neuroimagem, psicofisiologia ou experimentos neuroeconômicos, nenhuma inferência neurobiológica é formulada.

Competitividade também exige tratamento causalmente disciplinado. A arquitetura não cria vantagem competitiva por si só. Ela pode alterar mecanismos intermediários - posição de custo, diferenciação, velocidade de resposta, resiliência, poder de barganha e acumulação de capacidades - e esses mecanismos somente se convertem em desempenho quando combinados a recursos e capacidades organizacionais apropriados (Porter; Millar, 1985; Barney, 1991; Teece; Pisano; Shuen, 1997). Pesquisas recentes sobre capacidade de IA e desempenho reforçam a importância de complementaridades organizacionais, cultura orientada a dados e intensidade de adoção (Fosso Wamba et al., 2024; Chen; Tajdini, 2025).

Diante disso, a pergunta de pesquisa é: como integrar, a partir da literatura e da análise documental, mecanismos econômicos, comportamentais e de governança corporativa à decisão entre IA local, IA em nuvem e arquiteturas híbridas, e quais relações conceituais podem ser propostas entre essas escolhas e custos, riscos, autonomia e competitividade empresarial? O objetivo geral é analisar essa decisão arquitetural como problema de alocação de capital e capacidade estratégica. Os objetivos específicos são: estruturar componentes de TCO e sensibilidade; identificar fatores comportamentais relevantes; relacionar mecanismos de governança a riscos decisórios; explicitar os canais entre arquitetura e competitividade; e propor um instrumento rastreável de comparação e revisão.

A contribuição pretendida não é afirmar a descoberta de um modelo universal ou inédito em sentido absoluto. O artigo busca integrar, em um mesmo processo de decisão, elementos que aparecem separados em referenciais de risco de IA, sistemas de gestão, literatura de computação distribuída, economia comportamental e estratégia. Essa integração é posteriormente comparada a referenciais existentes para delimitar com precisão o que o modelo acrescenta e o que permanece dependente de validação empírica.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

2 - REFERENCIAL TEÓRICO

2.1 GOVERNANÇA CORPORATIVA E DECISÕES DE INVESTIMENTO TECNOLÓGICO

A teoria da agência oferece um ponto de partida para compreender por que investimentos tecnológicos complexos demandam governança. Jensen e Meckling (1976) mostram que delegação de autoridade, diferenças de objetivos e custos de monitoramento influenciam decisões organizacionais. Em projetos de IA, fornecedores podem conhecer melhor limitações de sua tecnologia do que compradores; equipes técnicas podem dominar detalhes que o conselho não conhece; e administradores podem deter informações estratégicas ou financeiras não visíveis aos especialistas. A qualidade da decisão depende, portanto, de como informação, incentivos e responsabilidades são traduzidos e registrados.

Shleifer e Vishny (1997) associam governança aos mecanismos pelos quais financiadores buscam assegurar retorno sobre seus investimentos. Em perspectiva contemporânea, os Princípios de Governança Corporativa do Grupo dos Vinte (G20) e da OECD destacam orientação estratégica, supervisão e responsabilização do conselho (OECD, 2023). A decisão de arquitetura de IA entra nesse escopo quando compromete capital, continuidade, segurança, relações com fornecedores e ativos informacionais.

Referenciais de IA reforçam essa leitura. O Artificial Intelligence Risk Management *Framework* (AI RMF) do National Institute of Standards and Technology (NIST) organiza a gestão nas funções governar, mapear, medir e gerenciar e enfatiza sua integração à governança organizacional (Tabassi, 2023). A norma ISO/IEC 38507:2022, elaborada pela International Organization for Standardization (ISO) e pela International Electrotechnical Commission (IEC), trata das implicações de governança do uso de IA; a ISO/IEC 42001:2023 estabelece requisitos para um sistema de gestão de IA (ISO, 2022, 2023). Nenhum desses referenciais determina onde a IA deve ser executada. Eles oferecem, contudo, fundamentos para responsabilidades, monitoramento, documentação, avaliação de risco e melhoria contínua.

Neste estudo, governança é tratada como disciplina de qualidade da alocação de capital, não como burocracia adicional. Seu valor potencial está em exigir critérios antes da seleção de fornecedores, tornar alternativas comparáveis, registrar premissas e incertezas, atribuir responsáveis e estabelecer gatilhos de revisão. Esse desenho também permite testar a influência de vieses sem presumir que processos formais os eliminem.

2.2 ECONOMIA COMPORTAMENTAL, RACIONALIDADE LIMITADA E CONTRIBUIÇÃO CONTEXTUAL DA NEUROECONOMIA

Simon (1955) rompe com a hipótese de decisão plenamente otimizadora ao mostrar que agentes operam sob limites de informação, tempo e capacidade de processamento. Em

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

organizações, esses limites são distribuídos entre pessoas, áreas, comitês, consultorias e fornecedores. Assim, uma empresa pode dispor de grande capacidade analítica e, simultaneamente, decidir sob pressão, incerteza e incentivos conflitantes.

A neuroeconomia aproxima economia, psicologia e neurociência para estudar decisões econômicas (Camerer; Loewenstein; Prelec, 2005). Entretanto, o presente artigo não possui desenho neuroeconômico experimental. Sua lente operacional é a economia comportamental e a literatura de julgamento e decisão; a neuroeconomia é utilizada apenas para situar a interdisciplinaridade do problema. A obra de Chavaglia Neto, Filipe e Ferreira (2017) é empregada como referência brasileira aplicada dessa articulação, sem substituir as fontes seminais dos mecanismos comportamentais e sem constituir critério de seleção do corpus.

A teoria prospectiva mostra que perdas e ganhos são avaliados em relação a pontos de referência (Kahneman; Tversky, 1979), e o enquadramento pode alterar preferências (Tversky; Kahneman, 1981). Em investimentos de IA comparar "CAPEX local" com "mensalidade flexível" é cognitivamente diferente de organizar as duas alternativas pelas mesmas categorias de custo, risco e valor. A governança deve, portanto, padronizar a estrutura de apresentação antes de deliberar.

A ancoragem pode ocorrer quando o primeiro preço, contrato vigente ou retorno sobre o investimento (ROI) apresentado domina comparações posteriores. Chavaglia Neto, Filipe e Ramalheiro (2011) estudam o efeito em decisões de consumo; aqui, o mecanismo é transposto como hipótese analítica, sem presumir sua ocorrência em empresa específica. Aversão à perda e viés de *status quo* podem aumentar o peso atribuído à perda de rotinas ou ativos existentes (Kahneman; Knetsch; Thaler, 1991), enquanto custos efetivamente irrecuperáveis podem ser usados para justificar permanência (Arkes; Blumer, 1985).

Outros mecanismos também importam. O efeito manada pode produzir imitação, embora o comportamento de outros agentes possa transmitir informação em ambientes incertos (Banerjee, 1992). Excesso de confiança pode distorcer estimativas de capacidade e precisão (Moore; Healy, 2008). O viés de confirmação pode orientar a busca seletiva de evidências, inclusive em processos de inovação (Chavaglia Neto et al., 2022). Excesso de opções pode reduzir motivação ou adiar escolhas (Iyengar; Lepper, 2000), preferências podem ser construídas no próprio contexto decisório (Slovic, 1995), e a literatura sobre desconto temporal permite formular hipóteses sobre procrastinação, autocontrole e mecanismos de compromisso (Frederick; Loewenstein; O'Donoghue, 2002). O estudo de Ariely e Wertenbroch (2002), incluído no levantamento original, foi retratado em 2 de setembro de 2026 e deixou de fundamentar os controles propostos (Retraction..., 2026). A crença na lei dos pequenos números recomenda cautela ao generalizar uma única prova de conceito (PoC) ou poucos *benchmark*s (Tversky; Kahneman, 1971).

Chavaglia Neto, Filipe e Caleiro (2019) relacionam economia comportamental, criatividade e inovação, contribuição útil para distinguir adoção tecnológica de geração de valor. Essa distinção

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

é central: uma arquitetura somente é estratégica quando modifica resultados, capacidades ou opções relevantes; novidade técnica, isoladamente, não equivale a inovação econômica.

2.3 ESCOPO TECNOLÓGICO: IA LOCAL, NUVEM E HÍBRIDA EM CARGAS DE INFERÊNCIA

O NIST define computação em nuvem como acesso sob demanda a recursos compartilhados, com elasticidade e serviço mensurado (Mell; Grance, 2011). Em artigo publicado pela Association for Computing Machinery (ACM), Armbrust et al. (2010) discutem elasticidade e pagamento por uso entre seus fundamentos econômicos, ao lado de limitações como movimentação de dados e previsibilidade de desempenho. Neste artigo, IA local significa execução sob controle direto ou predominante da organização, em data center privado, servidor dedicado ou dispositivo de borda, sem pressupor desconexão da internet nem superioridade automática de segurança.

A literatura recente sobre LLMs atualiza a comparação. He et al. (2024) tratam alocação e descarregamento de inferência entre nuvem e borda; Friha et al. (2024) sintetizam arquiteturas de inteligência de borda baseadas em LLMs; Yu et al. (2024) demonstram técnicas para adaptação eficiente em dispositivos com restrições de memória e computação; e Chen et al. (2025) organizam sistemas centralizados, híbridos e descentralizados para inteligência de borda. Esses trabalhos mostram viabilidade e *trade-offs* técnicos, mas não autorizam concluir, por si sós, qual arranjo possui menor TCO empresarial.

A colaboração entre borda e nuvem também aparece como desenho explícito. Jin e Wu (2025) propõem mecanismo adaptativo que distribui inferência conforme latência e condições do ambiente; Jiang et al. (2026) revisam o campo de LLMs de borda e seus desafios de execução. Em conjunto, a literatura de 2024-2026 reforça duas premissas deste artigo: a fronteira local-nuvem não é binária e a adequação arquitetural precisa ser avaliada por carga, dados, recursos e restrições, não pela categoria tecnológica isolada.

Arquiteturas híbridas podem manter inferências ou dados críticos sob maior controle e utilizar nuvem para picos, modelos específicos ou contingência. Essa flexibilidade pode ter valor, mas adiciona observabilidade, identidade, integração, portabilidade, segurança distribuída e múltiplos contratos. O híbrido, portanto, não é compromisso automaticamente superior; deve demonstrar que o valor da segmentação supera o custo de coordenação.

2.4 COMPETITIVIDADE: DO RECURSO TECNOLÓGICO AO MECANISMO ECONÔMICO

Competitividade é tratada neste artigo como resultado mediado, e não como atributo intrínseco da arquitetura. Porter e Millar (1985) mostram que tecnologia da informação pode alterar custo, diferenciação e ligações na cadeia de valor. Barney (1991) acrescenta que vantagem sustentada depende de recursos valiosos e de condições que dificultem sua replicação ou

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

substituição. Teece, Pisano e Shuen (1997) deslocam a atenção para capacidades dinâmicas: em ambientes de mudança tecnológica, desempenho depende da habilidade de integrar, construir e reconfigurar competências.

Aplicada à IA, essa base permite explicitar seis mecanismos. Primeiro, a arquitetura pode alterar posição de custo e produtividade. Segundo, pode suportar diferenciação por personalização, privacidade, qualidade ou latência. Terceiro, pode modificar velocidade de experimentação e *time-to-market*. Quarto, pode afetar resiliência e capacidade de atender requisitos de clientes e reguladores. Quinto, pode alterar poder de barganha e custo de saída. Sexto, pode favorecer ou limitar acumulação de conhecimento, dados e competências internas. Apenas quando um desses mecanismos gera valor relativo a concorrentes e é sustentado por capacidades complementares faz sentido falar em vantagem competitiva.

Evidência contemporânea reforça a necessidade dessas complementaridades. Fosso Wamba et al. (2024) associam capacidade de IA e desempenho à mediação de cultura orientada por dados. Chen e Tajdini (2025) relacionam intensidade de adoção de IA a desempenho sob condições organizacionais e ambientais. Filippucci et al. (2024) destacam que efeitos de produtividade dependem de difusão, organização, competências e competição. Em revisão recente de IA generativa nas organizações, Machkour e Abriane (2026) também tratam capacidades, reconfiguração e governança como mecanismos entre uso e desempenho. Portanto, decidir onde executar IA é uma condição de desenho do sistema de capacidades, não um substituto para estratégia, processos e aprendizagem.

2.5 DEPENDÊNCIA TECNOLÓGICA, CONFORMIDADE E PODER DE BARGANHA

O aprisionamento a fornecedor possui dimensão econômica porque dependência altera custos de troca e poder de barganha. Klemperer (1987) mostra que custos de troca transformam relações competitivas mesmo quando alternativas eram inicialmente comparáveis. Em IA, esses custos podem surgir de APIs proprietárias, formatos, modelos, *pipelines*, armazenamento, transferência de dados e competências específicas. Contratar um fornecedor não é, por isso, inadequado; significa que portabilidade e saída devem ser precificadas e governadas desde o início.

O controle de dados pode ter valor quando afeta reutilização, propriedade intelectual, continuidade ou capacidade de atender obrigações. A Lei Geral de Proteção de Dados Pessoais (LGPD) disciplina tratamento de dados pessoais no Brasil (Brasil, 2018), e o AI Act europeu estabelece regras harmonizadas baseadas em risco para sistemas de IA (União Europeia, 2024). A arquitetura não assegura conformidade, mas pode modificar a facilidade de demonstrar controles, segregar dados e produzir evidências de auditoria.

Conformidade e resiliência, assim, podem operar como mecanismos competitivos indiretos quando preservam receita, reduzem fricção de auditoria ou habilitam acesso a contratos. O efeito

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

precisa ser demonstrado no caso concreto e comparado ao custo de produzir e manter esses controles.

3 - METODOLOGIA

3.1 DESENHO, UNIDADE DE ANÁLISE E DELIMITAÇÃO

A pesquisa é qualitativa, bibliográfica, documental, comparativa e aplicada. Adota uma revisão narrativa orientada por critérios temáticos, não uma revisão sistemática nem meta-análise. A unidade de análise é a decisão organizacional sobre arquitetura de execução de cargas corporativas de inferência de IA, com ênfase em IA generativa e LLMs. Não são realizados experimento neuroeconômico, estudo de caso, levantamento primário de preços ou estimação causal entre arquitetura e desempenho. A função da revisão é sintetizar mecanismos e construir um instrumento de decisão cuja validação empírica permanece como agenda posterior.

O conjunto de fontes é delimitado por pertinência temática e metodológica, independentemente da identidade dos autores. A identificação de autoria, título e identificador digital de objeto (DOI) serve à conferência dos metadados, não como critério de inclusão. Isso se aplica inclusive às obras de José Chavaglia Neto, consideradas por sua relação com ancoragem, economia comportamental, neuroeconomia ou inovação e articuladas às fontes seminais correspondentes.

3.2 BUSCA, ELEGIBILIDADE E FLUXO DO CORPUS

O levantamento inicial tem como corte temporal 19 de agosto de 2026. A revisão articula literatura técnico-científica disponível na biblioteca Xplore do Institute of Electrical and Electronics Engineers (IEEE), em SpringerLink e nas páginas dos editores; Crossref e DOI auxiliam a conferência bibliográfica. As fontes institucionais e documentais incluem OECD, NIST, catálogo ISO, EUR-Lex, portal jurídico da União Europeia, e Eurostat. Títulos e referências das obras seminais orientam o encadeamento retrospectivo da literatura clássica. A delimitação temática reúne cinco grupos de termos: (1) "local AI", "on-premise AI", "*on-device* AI", cloud e enterprise; (2) "*large language model*", edge, local, cloud, hybrid e inference; (3) "AI capability", "artificial intelligence adoption", firm performance, productivity e competitive advantage; (4) artificial intelligence, governance, risk e enterprise decision; e (5) anchoring, loss aversion, *sunk cost*, overconfidence, herding, technology investment e decision making. Esses grupos explicitam o recorte da revisão narrativa, sem constituir uma estratégia de busca sistemática exaustiva.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

Quadro 1 - Critérios e organização da revisão narrativa

Elemento	Delimitação adotada	Finalidade
Janela temporal	Seminais sem limite inferior; prioridade 2024-2026 para arquitetura de IA/LLMs e desempenho organizacional. Corte inicial: 19 ago. 2026; complementação: 8 out. 2026.	Evitar obsolescência técnica sem perder fontes conceituais de origem.
Inclusão	Artigo revisado por pares, livro acadêmico, norma, lei, dado oficial ou documento de instituição pública ou organização técnica; relação direta com categoria analítica; metadados verificáveis.	Garantir pertinência e verificabilidade.
Exclusão	Duplicata; fonte comercial usada como prova independente de vantagem econômica; estudo puramente algorítmico sem implicação arquitetural; referência sem metadados confirmáveis; versão preliminar quando havia registro editorial mais estável.	Reduzir viés comercial e ruído técnico.
Composição bibliográfica	44 referências do levantamento inicial e 7 fontes complementares, totalizando 51 entradas. O artigo retratado e seu aviso documentam a atualização bibliográfica; não fundamentam as proposições do modelo.	Delimitar o conjunto bibliográfico efetivamente apresentado.
Síntese	Codificação temática sem meta-análise: economia/TCO; arquitetura; comportamento; governança; competitividade.	Gerar proposições condicionais e matrizes, não estimativas agregadas.

Fonte: elaboração própria. As contagens referem-se às entradas da lista bibliográfica, não a resultados brutos de portais ou a um fluxo de triagem sistemática. A seleção temática não permite estimar cobertura exaustiva da literatura.

3.3 EXTRAÇÃO, SÍNTESE E RASTREABILIDADE

A síntese organiza as fontes nas dimensões economia/TCO, arquitetura, comportamento, governança e competitividade. Para cada argumento, distinguem-se o tipo de fonte e a natureza da evidência (conceitual, normativa, empírica ou técnica), a proposição sustentada e sua aplicação às categorias e aos quadros. A identificação bibliográfica e o ano constam das referências; as relações com custos, requisitos, riscos, comportamento e mecanismos de competitividade são explicitadas no texto e nas matrizes. A síntese não atribui pesos numéricos à evidência nem converte resultados heterogêneos em efeito agregado.

A derivação dos quadros segue três passos. Primeiro, proposições relacionadas são agrupadas por categoria. Segundo, as categorias são convertidas em perguntas ou critérios comparáveis entre local, nuvem e híbrido. Terceiro, as citações e notas dos quadros identificam as fontes que sustentam os mecanismos e permitem distinguir esse suporte das relações propostas pelos autores. A rastreabilidade oferecida é bibliográfica e conceitual: conecta fonte, categoria, proposição e aplicação no quadro, sem reivindicar a reprodutibilidade de uma revisão sistemática.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Em consulta complementar realizada em 8 de outubro de 2026, foram examinadas fontes sobre TCO, gestão financeira da tecnologia (FinOps), decisão tecnológica, arquitetura empresarial e apoio multicritério para aprofundar a comparação do Quadro 5, além de literatura sobre desconto temporal e do aviso de retratação citado na seção 2.2. Essas sete fontes ampliam a bibliografia de 44 para 51 referências, preservando a distinção entre o levantamento inicial e a atualização. O trabalho retratado permanece identificado para documentar a atualização, sem servir de evidência para os controles propostos.

A análise mantém linguagem condicional. Estudos de engenharia demonstram desempenho ou viabilidade em configurações específicas; não são convertidos em afirmações gerais de menor TCO. Estudos de adoção e desempenho organizacional sustentam relações entre capacidades e resultados, mas não identificam causalmente uma arquitetura superior. Referenciais normativos sustentam requisitos e processos de governança, não benefícios econômicos automáticos.

4 - RESULTADOS

Os resultados apresentados nesta seção são produtos da síntese conceitual e da análise documental. Os mecanismos econômicos, tecnológicos, comportamentais e de governança provêm das fontes citadas; sua organização em categorias de custo, matrizes, relações entre etapas e perguntas de controle constitui uma proposta dos autores para o recorte de inferência empresarial. A equação, os quadros e o processo integrado não foram submetidos a validação empírica neste estudo.

4.1 FORMALIZAÇÃO DO TCO E ANÁLISE DE SENSIBILIDADE

A primeira síntese propõe comparar o custo total associado ao caso de uso e ao mesmo horizonte, em consonância com a abordagem de TCO de Ellram (1995) e com as considerações econômicas de Armbrust et al. (2010) e Badger et al. (2012). Para IA local, entram aquisição, instalação, energia, refrigeração, redes, armazenamento, software, manutenção, pessoal, segurança, atualização, indisponibilidade e substituição. Para nuvem, entram computação, APIs, armazenamento, transferência de dados e tarifas de saída, serviços gerenciados, suporte, integração, monitoramento e crescimento do consumo. No híbrido, somam-se custos incrementais de integração, observabilidade, identidade, orquestração e controles distribuídos, evitando dupla contagem.

Para uma alternativa a pertencente ao conjunto {local, nuvem, híbrida}, horizonte H e taxa de desconto k, os autores organizam a comparação em valor presente pela seguinte expressão, proposta como instrumento analítico e não como equação empiricamente estimada:

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

$$TCO_a(H) = CAPEX_{0,a} + \Sigma[t=1..H] (OPEX_{a,t} + CINT_{a,t} + CTRANS_{a,t} + CEXIT_{a,t})/(1+k)^t - VR_{a,H}/(1+k)^H$$

CAPEX_0,a representa o investimento inicial; OPEX_a,t, os custos operacionais no período t; CINT_a,t, integração e orquestração; CTRANS_a,t, contratação, compliance e gestão de fornecedores; CEXIT_a,t, custos esperados de saída ou migração quando aplicáveis; e VR_a,H, valor residual recuperável no fim do horizonte. Benefícios, receitas e custos de oportunidade devem ser modelados separadamente quando mensuráveis, para evitar misturar custo contábil com criação de valor. A diferença entre duas alternativas é $\Delta TCO_{a,b} = TCO_a - TCO_b$; o ponto de equilíbrio ocorre quando $\Delta TCO_{a,b} = 0$. As parcelas devem ser mutuamente exclusivas: custos discriminados em integração, contratação ou saída não devem ser novamente contabilizados no OPEX.

A análise de sensibilidade deve variar, uma a uma e de forma conjunta, as premissas capazes de alterar a classificação: utilização de aceleradores, crescimento da demanda, preço de tokens/APIs, custo de energia, custo de equipe, *egress*, vida útil, taxa de desconto, indisponibilidade e custo de saída. Em vez de percentuais arbitrários, os intervalos devem vir de contratos, telemetria, propostas ou *benchmark*s do caso. Se a alternativa preferida muda dentro de uma faixa plausível, o resultado deve ser apresentado como sensível e pode justificar fases, contratos reversíveis ou arquitetura híbrida temporária. Essa regra transforma a sensibilidade em mecanismo decisório, não em apêndice financeiro. A prática de gestão financeira da tecnologia denominada FinOps articula engenharia, finanças e negócio no acompanhamento de custos e valor e também contempla decisões de alocação de cargas e governança (Markanastasakis, 2026).

4.2 MATRIZ ECONÔMICA DAS ALTERNATIVAS

Quadro 2 - Matriz econômica comparativa de arquiteturas de IA

Dimensão	IA local	IA em nuvem	Arquitetura híbrida
CAPEX	Mais relevante quando há compra de servidores/aceleradores; depende de utilização e vida útil.	Menor barreira inicial de infraestrutura em muitos serviços; consumo migra para despesa variável.	Combina investimento seletivo e consumo externo; requer evitar capacidade duplicada.
OPEX	Energia, equipe, manutenção, licenças, segurança e atualização.	Computação, APIs, armazenamento, *egress*, suporte e serviços gerenciados.	Combina os dois ambientes e adiciona integração/orquestração.
Elasticidade	Limitada à capacidade adquirida e ao tempo de expansão.	Alta quando recursos estão disponíveis sob demanda.	Pode usar local como base e nuvem para picos, se a aplicação permitir.
Controle de dados	Pode ampliar controle operacional, condicionado à qualidade dos controles internos.	Depende de serviço, contrato, configuração e jurisdição.	Permite segmentação por criticidade, com custo adicional de classificação e integração.
Latência	Pode ser favorável próximo à origem dos dados.	Depende de conectividade, região e desenho.	Pode distribuir componentes conforme requisito de resposta.
Escala/capacidade	Crescimento exige planejamento e aquisição/implantação.	Facilita escala rápida dentro de limites técnicos/comerciais do provedor.	Combina capacidade, porém requer política explícita de distribuição.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

Dimensão	IA local	IA em nuvem	Arquitetura híbrida
Competências	Exige operação de infraestrutura, IA, segurança e continuidade compatíveis com o desenho.	Reduz parte da operação física, mas não elimina competências de IA, dados, segurança e FinOps.	Requer competências nos dois ambientes e integração.
Dependência/saída	Menor dependência do provedor de execução pode coexistir com dependência de hardware, software e equipe.	APIs, dados, serviços gerenciados e conhecimento específico podem elevar custos de troca.	Pode diversificar dependência ou multiplicar tecnologias e contratos.
Implantação	Pode exigir aquisição, instalação e preparação de ambiente.	Pode acelerar experimentação e início de cargas suportadas.	Depende da maturidade de integração e dos ambientes existentes.
Conformidade/continuidade	Pode facilitar alguns requisitos de controle, sem garantir conformidade; continuidade depende de redundância interna.	Pode oferecer controles/certificações robustos, sob responsabilidade compartilhada e dependência de acordo de nível de serviço (SLA) e região.	Pode segmentar requisitos e criar alternativas de continuidade, mas aumenta pontos de controle.

Rastreabilidade: CAPEX/OPEX/elasticidade/escala/implantação - Mell e Grance (2011), Armbrust et al. (2010), He et al. (2024), Jin e Wu (2025); latência/controle/distribuição - Shi et al. (2016), Friha et al. (2024), Yu et al. (2024), Chen et al. (2025), Jiang et al. (2026); dependência e saída - Klemperer (1987); conformidade/continuidade - Tabassi (2023), ISO (2022, 2023), Brasil (2018), União Europeia (2024).

4.3 MATRIZ COMPORTAMENTAL E MECANISMOS DE GOVERNANÇA

A segunda síntese identifica que a qualidade do caso de negócio pode depender da informação disponível e de como referências, evidências e alternativas entram no processo. Os mecanismos abaixo são tratados como riscos possíveis, não como diagnóstico de que determinada organização ou decisão esteja enviesada.

Quadro 3 - Fatores comportamentais, riscos e controles na decisão de arquitetura

Fator	Manifestação	Risco	Controle de governança	Fonte principal
Ancoragem	Primeiro preço, contrato vigente ou ROI inicial domina comparações.	Alternativas avaliadas contra referência arbitrária.	Normalizar propostas; faixas e segunda estimativa independente.	Chavaglia Neto; Filipe; Ramalheiro (2011)
Aversão à perda / *status quo*	Medo de abandonar ativos, fornecedor ou rotina.	Manutenção ou rejeição inadequada de mudança.	Separar perdas prospectivas, migração e custos irrecuperáveis.	Kahneman; Knetsch; Thaler (1991)
Custo irrecuperável	Permanência justificada pelo valor já gasto.	Despesa futura sem justificativa prospectiva.	Desconsiderar gasto irrecuperável na decisão marginal; registrar valor residual.	Arkes; Blumer (1985)
Manada	Adoção porque concorrentes ou líderes adotaram.	Imitação sem comparabilidade de contexto.	Distinguir sinal informacional de pressão imitativa.	Banerjee (1992)
Excesso de confiança	Subestimação de prazo, equipe, operação ou incerteza.	Caso de negócio e contingências enviesados.	Faixas de incerteza, validação independente e cenários de falha.	Moore; Healy (2008)

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

Fator	Manifestação	Risco	Controle de governança	Fonte principal
Confirmação	Busca de *benchmark*s favoráveis à solução preferida.	PoC torna-se justificativa, não teste.	Hipótese falsificável e evidência que mudaria a decisão.	Chavaglia Neto et al. (2022)
Excesso de opções	Muitos modelos/provedores prolongam ou paralisam avaliação.	Tempo/custo de decisão e adiamento.	Critérios eliminatórios e lista curta antes de propostas.	Iyengar; Lepper (2000); Thaler; Sunstein (2008)
Contexto/preferência construída	Demos e enquadramentos alteram pesos percebidos.	Forma de apresentação afeta valor atribuído.	Roteiro e métricas padronizados entre alternativas.	Slovic (1995); Tversky; Kahneman (1981)
Presente/procrastinação	Foco no desembolso imediato ou adiamento sem gatilho.	Miopia temporal ou custo da espera.	Horizontes comparáveis, data e gatilho de revisão.	Frederick; Loewenstein; O'Donoghue (2002)
Pequenos números	Generalização de uma PoC ou poucos *benchmark*s.	Escala decidida com evidência pouco representativa.	Carga representativa, repetição e limites de generalização.	Tversky; Kahneman (1971)

Fonte: elaboração própria. A coluna final explicita a origem principal de cada linha; Simon (1955) e Camerer, Loewenstein e Prelec (2005) fornecem contexto geral, enquanto as proposições específicas derivam das fontes comportamentais indicadas. Chavaglia Neto, Filipe e Ferreira (2017) e Chavaglia Neto, Filipe e Caleiro (2019) são utilizados na contextualização aplicada, não como substitutos das fontes seminais. As manifestações em decisões de arquitetura e os controles de governança são transposições e proposições dos autores; as fontes indicadas não demonstram sua eficácia nesse contexto.

4.4 MODELO INTEGRADO DE DECISÃO PARA ARQUITETURA DE IA

A terceira síntese integra economia, comportamento e governança em sete etapas. O modelo é heurístico e não atribui pesos universais. Sua função é tornar rastreável a passagem do problema estratégico para requisitos, modelagem econômica, teste, decisão e revisão. Critérios eliminatórios devem preceder a pontuação compensatória; uma solução mais barata não compensa requisito legal ou de continuidade não atendido. Para critérios compensatórios, pesos e notas podem ser usados desde que aprovados previamente e submetidos a sensibilidade.

Quadro 4 - Modelo Integrado de Decisão para Arquitetura de IA

Etapa	Objetivo	Pergunta de controle	Trilha principal de evidência
1. Estratégia e caso de uso	Definir processo, usuário e resultado econômico esperado.	Qual valor será criado ou protegido?	Porter; Millar (1985); Barney (1991); Teece; Pisano; Shuen (1997)
2. Requisitos eliminatórios	Classificar dados, latência, disponibilidade, jurisdição, segurança e continuidade.	Algum requisito exclui uma alternativa antes do TCO?	Tabassi (2023); ISO (2022, 2023); Brasil (2018); União Europeia (2024)

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Etapa	Objetivo	Pergunta de controle	Trilha principal de evidência
3. Economia e sensibilidade	Modelar TCO, oportunidade, dependência e saída em cenários.	Quais variáveis mudam o ranking e onde está o ponto de equilíbrio?	Mell; Grance (2011); Armbrust et al. (2010); He et al. (2024); Jin; Wu (2025); Ellram (1995); Badger et al. (2012)
4. Auditoria comportamental	Testar âncora, perda, *sunk cost*, manada, confiança, confirmação, contexto e presente.	Qual evidência contrária foi procurada?	Fontes do Quadro 3
5. Teste comparável	Executar *benchmark*/PoC com carga e métricas definidas previamente.	O teste representa a carga real e mede integração/operação?	Yu et al. (2024); Friha et al. (2024); Jiang et al. (2026); Tversky; Kahneman (1971)
6. Decisão e responsabilização	Registrar pesos, riscos aceitos, contrato, portabilidade e responsáveis.	Quem assume cada risco e quem pode reabrir a decisão?	Jensen; Meckling (1976); OECD (2023); Tabassi (2023)
7. Monitoramento e revisão	Acompanhar custo, uso, incidentes, desempenho e resultado de negócio.	As premissas permanecem válidas? Há gatilho de migração/rebalanceamento?	ISO (2023); Tabassi (2023); Teece; Pisano; Shuen (1997)

Fonte: elaboração própria a partir das trilhas indicadas. A contribuição reivindicada é integrativa e aplicada: o modelo conecta decisão arquitetural, TCO/sensibilidade, auditoria comportamental, governança e competitividade; não se afirma ineditismo universal sem validação comparativa adicional.

4.5 POSICIONAMENTO FRENTE A REFERENCIAIS EXISTENTES

O Quadro 5 compara os escopos documentados nas fontes selecionadas, incluindo a arquitetura empresarial pelo The Open Group Architecture Framework (TOGAF) e o apoio multicritério pela análise hierárquica (AHP, do inglês analytic hierarchy process). A comparação delimita complementaridades e necessidades de adaptação ao problema, sem constituir levantamento exaustivo dos referenciais ou teste de superioridade do modelo.

Quadro 5 - Posicionamento do modelo frente a referenciais concorrentes ou complementares

Referencial	Contribuição central	Delimitação para este problema	Relação com o modelo proposto
NIST AI RMF 1.0	Governança e ciclo de gestão de riscos de IA (Tabassi, 2023).	As funções de risco, isoladamente, não permitem inferir viabilidade econômica ou eficácia da integração proposta.	Sustenta requisitos eliminatórios (2), responsabilização (6) e revisão (7).
ISO/IEC 42001 e 38507	Sistema de gestão e implicações de governança do uso de IA (ISO, 2022, 2023).	Escopo normativo; não constitui evidência empírica de menor TCO ou de redução de vieses.	Sustenta responsabilidades, controles, documentação e melhoria contínua (2, 6 e 7).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

Referencial	Contribuição central	Delimitação para este problema	Relação com o modelo proposto
Decisão tecnológica em nuvem	Compara implantação, custos, desempenho, continuidade e portabilidade (Badger et al., 2012).	Recomendações anteriores às atuais cargas de LLMs; sua transposição exige avaliação do caso.	Orienta requisitos (2) e comparação econômica (3), atualizados com evidências da PoC (5).
FinOps e modelos de TCO	TCO considera custos além do preço; FinOps articula valor, alocação de cargas, governança e acompanhamento (Ellram, 1995; Markanastasakis, 2026).	Já abordam custo e valor; a proposta não reivindica exclusividade dessas dimensões.	Relaciona TCO/sensibilidade (3) às premissas comportamentais (4), à PoC (5) e à revisão (7).
Literatura edge/LLM 2024-2026	Desempenho, offloading, execução local e colaboração nuvem-borda.	Resultados dependem das configurações examinadas; não estabelecem TCO empresarial geral.	Sustenta condições técnicas e variáveis dos requisitos (2), da economia (3) e da PoC (5).
Arquitetura empresarial/TOGAF	Método e linguagem comuns para relacionar necessidades do negócio e arquitetura (The Open Group, 2022).	A documentação geral requer adaptação à inferência; não se comparam aqui todos os seus artefatos.	Contextualiza estratégia (1), requisitos (2), governança da decisão (6) e evolução arquitetural (7).
Dependência de fornecedores	Custos de troca alteram competição e poder de barganha (Klemperer, 1987).	Sua magnitude em IA requer dados de contratos, integração, portabilidade e competências.	Inclui saída no TCO (3), responsabilidades de portabilidade (6) e gatilhos de revisão (7).
Apoio multicritério/AHP	Compara critérios e alternativas, deriva prioridades e examina a consistência dos julgamentos (Saaty, 2008).	Prioridades consistentes não comprovam viabilidade de execução nem afastam vieses.	Serve de referência à pontuação após requisitos eliminatórios (2); o modelo não aplica formalmente AHP.
Literatura de capacidade/desempenho em IA	Relaciona adoção/capacidades a desempenho e complementaridades organizacionais.	As relações examinadas não identificam causalmente uma arquitetura superior ou de menor custo.	Sustenta a relação conceitual entre arquitetura, capacidades, mecanismos de valor e competitividade.
Modelo deste artigo	Decisão de arquitetura como alocação de capital sob risco, comportamento e governança.	Síntese heurística sem validação empírica; originalidade relativa e eficácia não foram demonstradas.	Propõe encadear requisitos, TCO/sensibilidade, auditoria comportamental, PoC e revisão das premissas.

Fontes: Tabassi (2023); ISO (2022, 2023); He et al. (2024); Friha et al. (2024); Yu et al. (2024); Chen et al. (2025); Jin e Wu (2025); Jiang et al. (2026); Fosso Wamba et al. (2024); Chen e Tajdini (2025); Machkour e Abriane (2026). Complementação: Badger et al. (2012); Ellram (1995); Markanastasakis (2026); The Open Group (2022); Saaty (2008).

Da articulação desses mecanismos resulta uma regra de encadeamento: requisitos delimitam alternativas admissíveis; TCO e sensibilidade explicitam sob quais premissas a preferência muda; a auditoria comportamental orienta a procura de contraprovas; a PoC pode atualizar custos e requisitos; e a revisão reabre a decisão quando essas premissas se alteram. Essa

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

relação entre restrições, julgamento e atualização de evidências constitui a contribuição conceitual proposta. A lacuna abordada é a operacionalização conjunta desses vínculos no recorte de inferência empresarial, sem pressupor ausência geral dessas dimensões na literatura. A sequência e seus efeitos permanecem proposições a validar.

5 - DISCUSSÃO

5.1 ARQUITETURA, ECONOMIA E EVIDÊNCIA TECNOLÓGICA RECENTE

A síntese conceitual dialoga com a literatura técnica recente sem convertê-la em prescrição econômica. Estudos de offloading e colaboração nuvem-borda mostram que distribuir inferência pode reduzir latência ou deslocar carga sob configurações específicas (He et al., 2024; Jin; Wu, 2025), enquanto pesquisas de execução e adaptação em borda demonstram avanços de eficiência em dispositivos restritos (Yu et al., 2024; Jiang et al., 2026). Esses achados sustentam a existência de alternativas arquiteturais tecnicamente viáveis, mas o TCO depende de preço, utilização, equipe, integração e horizonte observados pela organização.

Essa distinção permite explicitar uma possível fragilidade em decisões de infraestrutura: eficiência computacional e eficiência econômica não são sinônimos. Um ganho de latência pode justificar custo adicional quando muda receita, experiência ou requisito operacional; em outro caso, pode não produzir valor mensurável. Da mesma forma, elasticidade da nuvem tem valor quando reduz o custo de capacidade ociosa ou acelera experimentação, enquanto maior controle local tem valor quando sua utilização, sensibilidade de dados ou continuidade justificam o investimento. O ponto de equilíbrio deve ser encontrado por cenário, não presumido pela arquitetura.

5.2 MECANISMOS ENTRE ARQUITETURA E COMPETITIVIDADE

A contribuição conceitual central proposta consiste em explicitar o elo intermediário entre escolha arquitetural e competitividade. A partir de Porter e Millar (1985), a arquitetura pode alterar atividades da cadeia de valor e posição de custo; a partir de Barney (1991), controle de dados, conhecimento ou infraestrutura somente gera vantagem sustentada se constituir recurso valioso e difícil de substituir ou imitar no contexto da firma; e, a partir de Teece, Pisano e Shuen (1997), o valor depende também da capacidade de reconfigurar recursos conforme tecnologia e mercado mudam.

Os estudos recentes de desempenho em IA são coerentes com essa mediação. Fosso Wamba et al. (2024) destacam o papel de cultura orientada a dados, e Chen e Tajdini (2025) mostram que intensidade de adoção e contexto organizacional importam para desempenho. A revisão de Machkour e Abriane (2026) enfatiza capacidades e governança na tradução da IA generativa em resultados organizacionais. Portanto, uma empresa não se torna mais competitiva

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

por "ter IA local" ou "estar na nuvem". A arquitetura pode contribuir quando reduz custo relativo, melhora diferenciação, acelera resposta, aumenta resiliência, melhora barganha ou acumula capacidades que a organização consegue explorar.

Essa formulação também limita alegações sobre soberania ou autonomia. Portabilidade, conhecimento interno e possibilidade de substituir fornecedores podem ser opções estratégicas com valor de barganha, mas exigem investimento. Elasticidade e acesso a serviços especializados em nuvem também são opções econômicas. O *framework* propõe que cada opção seja associada a um mecanismo e a uma métrica observável, evitando transformar autonomia ou flexibilidade em slogans.

5.3 GOVERNANÇA, COMPORTAMENTO E CONTRIBUIÇÃO DO MODELO

O modelo articula referenciais de risco e gestão à comparação de um investimento em arquitetura de IA. NIST e ISO oferecem governança e gestão; a literatura de edge/LLM oferece condições técnicas; economia comportamental explica riscos do processo de julgamento; estratégia explica os mecanismos de valor. O modelo proposto une essas camadas e torna explícito onde cada fonte entra. Essa é uma contribuição integrativa, ainda não evidência de superioridade do modelo frente a alternativas. As sobreposições com FinOps, TCO, arquitetura empresarial e análise multicritério, delimitadas no Quadro 5, situam a contribuição nas relações entre etapas e nas condições para atualizar ou reabrir a decisão. A redução de custos e vieses pela adoção desse encadeamento permanece uma hipótese a testar.

A auditoria comportamental também deve ser interpretada com cautela. Nenhuma matriz prova que um viés ocorreu. Seu objetivo é criar perguntas de controle, contraprovas e registros capazes de revelar dependência excessiva de âncoras, preferências ou evidências pequenas. Essa postura é mais consistente com a evidência disponível do que denominar o instrumento "neuroeconômico", pois não há medição neural neste estudo.

As limitações permanecem relevantes. A síntese resulta de uma revisão narrativa com seleção temática, sem cobertura exaustiva ou avaliação sistemática do risco de viés das fontes. A bibliografia apresentada não permite estimar efeitos agregados nem frequências de ocorrência dos mecanismos na literatura. O TCO foi formalizado, mas não populado com dados contratuais reais. O modelo não foi aplicado longitudinalmente a decisões corporativas, e os seis mecanismos competitivos não foram testados causalmente. Essas limitações delimitam a agenda empírica: estudos de caso múltiplos, dados de custo e telemetria, registros de deliberação e comparação entre decisões antes e depois da aplicação do modelo.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

6 - CONCLUSÕES

Em resposta à pergunta de pesquisa, a síntese propõe relacionar a escolha entre IA local, IA em nuvem e arquitetura híbrida à interação entre perfil da carga, horizonte econômico, requisitos de dados e disponibilidade, competências, custos de integração e saída, vieses do processo decisório e mecanismos de governança. A revisão não oferece base para indicar uma arquitetura universalmente mais barata, segura ou competitiva. A contribuição conceitual proposta está em relacionar arquitetura e competitividade por seis mecanismos - custo/produtividade, diferenciação, velocidade, resiliência/acesso a mercado, poder de barganha e aprendizagem/capacidades - e em propor que esses mecanismos sejam mediados por recursos organizacionais e pela qualidade da decisão, não pela tecnologia isoladamente.

A contribuição gerencial proposta é o Modelo Integrado de Decisão para Arquitetura de IA, fundamentado na síntese temática descrita, na formalização de TCO, na análise de sensibilidade, nas relações entre fontes e proposições e na comparação com referenciais existentes. O instrumento pode orientar gestores a definir valor e requisitos antes de fornecedores, comparar fluxos no mesmo horizonte, testar premissas críticas, procurar evidência contrária, executar PoCs representativas, registrar responsabilidades e revisar a decisão com dados observados. Como o estudo não possui validação empírica nem experimento comportamental ou neuroeconômico, o modelo deve ser entendido como proposição aplicada a ser testada em organizações de diferentes setores e escalas, com métricas de custo, desempenho, incidentes, migração, tempo de mercado e resultados competitivos.

REFERÊNCIAS

ARIELY, Dan; WERTENBROCH, Klaus. Procrastination, deadlines, and performance: self-control by precommitment. Psychological Science, v. 13, n. 3, p. 219-224, 2002. DOI: 10.1111/1467-9280.00441. Artigo retratado em 2 set. 2026.

ARKES, Hal R.; BLUMER, Catherine. The psychology of sunk cost. Organizational Behavior and Human Decision Processes, v. 35, n. 1, p. 124-140, 1985. DOI: 10.1016/0749-5978(85)90049-4.

ARMBRUST, Michael et al. A view of cloud computing. Communications of the ACM, v. 53, n. 4, p. 50-58, 2010. DOI: 10.1145/1721654.1721672.

BADGER, Lee; GRANCE, Tim; PATT-CORNER, Robert; VOAS, Jeff. Cloud Computing Synopsis and Recommendations. Gaithersburg: National Institute of Standards and Technology, 2012. NIST Special Publication 800-146. DOI: 10.6028/NIST.SP.800-146.

BANERJEE, Abhijit V. A simple model of herd behavior. The Quarterly Journal of Economics, v. 107, n. 3, p. 797-817, 1992. DOI: 10.2307/2118364.

BARNEY, Jay. Firm resources and sustained competitive advantage. Journal of Management, v. 17, n. 1, p. 99-120, 1991. DOI: 10.1177/014920639101700108.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

BRASIL. Lei n. 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709compilado.htm. Acesso em: 19 ago. 2026.

CAMERER, Colin; LOEWENSTEIN, George; PRELEC, Drazen. Neuroeconomics: how neuroscience can inform economics. Journal of Economic Literature, v. 43, n. 1, p. 9-64, 2005. DOI: 10.1257/0022051053737843.

CHAVAGLIA NETO, José; FILIPE, José António; RAMALHEIRO, Brenno. Neuromarketing: consumers and the anchoring effect. International Journal of Latest Trends in Finance and Economic Sciences, v. 1, n. 4, p. 183-189, 2011. DOI: 10.35808/ijfirm/25.

CHAVAGLIA NETO, José; FILIPE, José António; FERREIRA, Manuel Alberto M. Neuroeconomia: uma nova perspectiva sobre o processo de tomada de decisões econômicas. 1. ed. Rio de Janeiro: Alta Books, 2017. ISBN 9788550800813.

CHAVAGLIA NETO, José; FILIPE, José António; CALEIRO, António. Creativity and innovation: a contribution of behavioral economics. International Journal of Innovation Studies, v. 3, n. 1, p. 12-21, 2019. DOI: 10.1016/j.ijis.2019.06.003.

CHAVAGLIA NETO, José; FILIPE, José António Candeias Bonito; CALEIRO, António Bento Ratão; SANT'ANNA, Anderson de Souza. Confirmation bias in innovation decision-making. RECIMA21 - Revista Científica Multidisciplinar, v. 3, n. 7, e371744, 2022. DOI: 10.47820/recima21.v3i7.1744.

CHEN, Handi et al. Toward edge general intelligence via large language models: opportunities and challenges. IEEE Network, v. 39, n. 5, p. 263-271, 2025. DOI: 10.1109/MNET.2025.3539338.

CHEN, Jing; TAJDINI, Saeed. A moderated model of artificial intelligence adoption in firms and its effects on their performance. Information Technology and Management, v. 26, p. 407-419, 2025. DOI: 10.1007/s10799-024-00422-5.

ELLRAM, Lisa M. Total cost of ownership: an analysis approach for purchasing. International Journal of Physical Distribution & Logistics Management, v. 25, n. 8, p. 4-23, 1995. DOI: 10.1108/09600039510099928.

EUROSTAT. Artificial intelligence by size class of enterprise. Conjunto de dados isoc_eb_ai. Luxembourg: European Commission, 2026. Dados de 2025; atualizado em 15 jun. 2026. DOI: 10.2908/ISOC_EB_AI. Disponível em: https://ec.europa.eu/eurostat/databrowser/view/isoc_eb_ai/default/table. Acesso em: 13 jul. 2026.

FILIPPUCCI, Francesco; GAL, Peter; JONA-LASINIO, Cecilia; LEANDRO, Alvaro; NICOLETTI, Giuseppe. The impact of Artificial Intelligence on productivity, distribution and growth: key mechanisms, initial evidence and policy challenges. OECD Artificial Intelligence Papers, n. 15. Paris: OECD Publishing, 2024. DOI: 10.1787/8d900037-en.

FOSSO WAMBA, Samuel et al. Artificial Intelligence Capability and Firm Performance: A Sustainable Development Perspective by the Mediating Role of Data-Driven Culture. Information Systems Frontiers, v. 26, p. 2189-2203, 2024. DOI: 10.1007/s10796-023-10460-z.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

FREDERICK, Shane; LOEWENSTEIN, George; O'DONOGHUE, Ted. Time discounting and time preference: a critical review. Journal of Economic Literature, v. 40, n. 2, p. 351-401, 2002. DOI: 10.1257/002205102320161311.

FRIHA, Othmane; FERRAG, Mohamed Amine; KANTARCI, Burak; CAKMAK, Burak; OZGUN, Arda; GHOUALMI-ZINE, Nassira. LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness. IEEE Open Journal of the Communications Society, v. 5, p. 5799-5856, 2024. DOI: 10.1109/OJCOMS.2024.3456549.

HE, Ying; FANG, Jingcheng; YU, F. Richard; LEUNG, Victor C. M. Large Language Models (LLMs) Inference Offloading and Resource Allocation in Cloud-Edge Computing: An Active Inference Approach. IEEE Transactions on Mobile Computing, v. 23, n. 12, p. 11253-11264, 2024. DOI: 10.1109/TMC.2024.3415661.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). ISO/IEC 38507:2022: Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations. Geneva: ISO, 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). ISO/IEC 42001:2023: Information technology - Artificial intelligence - Management system. Geneva: ISO, 2023.

IYENGAR, Sheena S.; LEPPER, Mark R. When choice is demotivating: can one desire too much of a good thing? Journal of Personality and Social Psychology, v. 79, n. 6, p. 995-1006, 2000. DOI: 10.1037/0022-3514.79.6.995.

JENSEN, Michael C.; MECKLING, William H. Theory of the firm: managerial behavior, agency costs and ownership structure. Journal of Financial Economics, v. 3, n. 4, p. 305-360, 1976. DOI: 10.1016/0304-405X(76)90026-X.

JIANG, Shan et al. Edge large language models: a comprehensive survey. CCF Transactions on Pervasive Computing and Interaction, v. 8, p. 181-210, 2026. DOI: 10.1007/s42486-025-00227-7.

JIN, Hongpeng; WU, Yanzhao. CE-CoLLM: Efficient and Adaptive Large Language Models Through Cloud-Edge Collaboration. In: 2025 IEEE International Conference on Web Services (ICWS). Proceedings [...]. IEEE, 2025. p. 316-323. DOI: 10.1109/ICWS67624.2025.00046.

KAHNEMAN, Daniel; KNETSCH, Jack L.; THALER, Richard H. Anomalies: the endowment effect, loss aversion, and status quo bias. Journal of Economic Perspectives, v. 5, n. 1, p. 193-206, 1991. DOI: 10.1257/jep.5.1.193.

KAHNEMAN, Daniel; TVERSKY, Amos. Prospect theory: an analysis of decision under risk. Econometrica, v. 47, n. 2, p. 263-291, 1979. DOI: 10.2307/1914185.

KLEMPERER, Paul. Markets with consumer switching costs. The Quarterly Journal of Economics, v. 102, n. 2, p. 375-394, 1987. DOI: 10.2307/1885068.

MACHKOUR, Badr; ABRIANE, Ahmed. Artificial Intelligence in Organizations: A Systematic Review of Operationalizing Generative Artificial Intelligence Capabilities for Organizational Performance. IEEE Access, v. 14, p. 72642-72658, 2026. DOI: 10.1109/ACCESS.2026.3692425.

MARKANASTASAKIS, Vasilio. FinOps Framework 2026: Executive Strategy, Technology Categories, and Converging Disciplines. FinOps Foundation, 19 mar. 2026. Disponível em: https://www.finops.org/insights/2026-finops-framework/. Acesso em: 8 out. 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

MELL, Peter; GRANCE, Timothy. The NIST definition of cloud computing. Gaithersburg: National Institute of Standards and Technology, 2011. NIST SP 800-145. DOI: 10.6028/NIST.SP.800-145.

MOORE, Don A.; HEALY, Paul J. The trouble with overconfidence. Psychological Review, v. 115, n. 2, p. 502-517, 2008. DOI: 10.1037/0033-295X.115.2.502.

OECD. G20/OECD Principles of Corporate Governance 2023. Paris: OECD Publishing, 2023. DOI: 10.1787/ed750b30-en.

OECD; BOSTON CONSULTING GROUP; INSEAD. The Adoption of Artificial Intelligence in Firms: New Evidence for Policymaking. Paris: OECD Publishing, 2025. DOI: 10.1787/f9ef33c3-en.

PORTER, Michael E.; MILLAR, Victor E. How Information Gives You Competitive Advantage. Harvard Business Review, v. 63, n. 4, p. 149-160, 1985.

RETRACTION: procrastination, deadlines, and performance: self-control by precommitment. Psychological Science, v. 37, n. 9, p. 665, 2026. DOI: 10.1177/09567976261488042.

SAATY, Thomas L. Decision making with the analytic hierarchy process. International Journal of Services Sciences, v. 1, n. 1, p. 83-98, 2008. DOI: 10.1504/IJSSCI.2008.017590.

SHI, Weisong et al. Edge computing: vision and challenges. IEEE Internet of Things Journal, v. 3, n. 5, p. 637-646, 2016. DOI: 10.1109/JIOT.2016.2579198.

SHLEIFER, Andrei; VISHNY, Robert W. A survey of corporate governance. The Journal of Finance, v. 52, n. 2, p. 737-783, 1997. DOI: 10.1111/j.1540-6261.1997.tb04820.x.

SIMON, Herbert A. A behavioral model of rational choice. The Quarterly Journal of Economics, v. 69, n. 1, p. 99-118, 1955. DOI: 10.2307/1884852.

SLOVIC, Paul. The construction of preference. American Psychologist, v. 50, n. 5, p. 364-371, 1995. DOI: 10.1037/0003-066X.50.5.364.

TABASSI, Elham. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: National Institute of Standards and Technology, 2023. NIST AI 100-1. DOI: 10.6028/NIST.AI.100-1.

TEECE, David J.; PISANO, Gary; SHUEN, Amy. Dynamic capabilities and strategic management. Strategic Management Journal, v. 18, n. 7, p. 509-533, 1997. DOI: 10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.

THALER, Richard H.; SUNSTEIN, Cass R. Nudge: improving decisions about health, wealth, and happiness. New Haven: Yale University Press, 2008.

THE OPEN GROUP. The TOGAF Standard. 10. ed. The Open Group, 2022. Disponível em: https://www.opengroup.org/togaf. Acesso em: 8 out. 2026.

TVERSKY, Amos; KAHNEMAN, Daniel. Belief in the law of small numbers. Psychological Bulletin, v. 76, n. 2, p. 105-110, 1971. DOI: 10.1037/h0031322.

TVERSKY, Amos; KAHNEMAN, Daniel. The framing of decisions and the psychology of choice. Science, v. 211, n. 4481, p. 453-458, 1981. DOI: 10.1126/science.7455683.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA SOL21
STANDARD OPEN LITERATURE - ISSN 3086-089X

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024, que cria regras harmonizadas em matéria de inteligência artificial. Jornal Oficial da União Europeia, L 2024/1689, 12 jul. 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Acesso em: 19 ago. 2026.

YU, Zhongzhi et al. EDGE-LLM: Enabling Efficient Large Language Model Adaptation on Edge Devices via Unified Compression and Adaptive Layer Voting. In: 61st ACM/IEEE Design Automation Conference (DAC). Proceedings [...]. New York: ACM, 2024. Artigo 327, p. 1-6. DOI: 10.1145/3649329.3658473.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.